# Using a Regression Discontinuity Design to Evaluate if the MLDA Affects Underage Drinking and Arrests

Eugene Okada Lee

December 2021

**Abstract**

The purpose of this research is to figure out if the Minimum Legal Drinking Age (MLDA) reduces the proportion of the population that drinks and if the MLDA helps reduce crime. We are using a regression discontinuity design with the United States MLDA of 21 years old as our cut-off and observe whether turning 21 years of age changes the rate of drinking and arrests accordingly. Our results show that the MLDA reduced drinking for individuals that are underage by 9.08 percentage points. It also found that the MLDA reduced arrests for individuals that are underage by 136.9 arrests.

## 1 Introduction

The motivation of this paper is to evaluate the effects of the minimum legal drinking age, which in the United States is set at 21 years of age. In the United States, underage drinking is a major cause of deaths and crimes amongst young adults. Because of this, we want to estimate and observe what the change in rates of crime and drinking are once an individual is legally allowed to drink. For decades, the United States has debated whether or not to decrease the MLDA from 21 to as young as 18 years old, to match the standard of how it is internationally in countries in Europe and Asia. This debate was most recently reintroduced in the summer of 2008 by a group of US college presidents and chancellors who formed an organization called the Amethyst Initiative. The group wanted the country as a whole to reexamine the legal drinking age of 21 and have the government consider lowering it. The argument of the Amethyst Initiative is that the current MLDA promotes more illegal/dangerous binge drinking in young adults, which leads to harmful consequences. This paper is designed to use the MLDA in the US to determine if this policy reduces crime and drinking for individuals who are under the minimum legal drinking age.

To examine the effects of the MLDA we use two different data sets. The first data set comes from the National Health Interview Survey (NHIS), which we use to measure the effect of the MLDA on the rate of drinking. The NHIS data set is survey data that includes information about an individual's age, whether they drink alcohol and other demographic characteristics. We understand that because the NHIS data is an open survey and that there is an existence of bias within the data which includes measurement error, recall bias, and desirability bias. We receive our arrest data from the California Monthly Arrest and Citation Register, which we use to measure the effect of the MLDA on the rate of arrests. This data set includes information of different types of arrests, both related to and unrelated to drinking, as well as the individuals age when they got arrested.

The empirical design we are using for our research is a regression discontinuity design where we use the MLDA of 21 as our discontinuity threshold. Right at our cut off, our observations immediately to the left and right include individuals who are similar to each other. Also at the cut off, there is a discontinuity right at the age of 21 for both drinking and arrests. We created two regression tables: one for drinking with all our demographic variables, and the other for each type of arrest. We visualize and examine the effect at our MLDA cut off with multiple regression discontinuity figures with specific configurations for the observed bin widths, bandwidths, y-range and we use a quadratic polynomial line of best fit.

Finally we implement an IV estimation method to estimate our true effect of drinking on arrests at our cut off. First we calculate and examine how drinking changes at our cutoff which will produce our first stage estimate. Then we examine how the rate of arrests change at our MLDA cutoff which will produce our reduced form. Then we take our first form and divide it by our reduced form to get our local average treatment effect IV estimate, which will tell us our effect of drinking on the arrest rate. Through our methods, we found that there was an effect on the rate of drinking and decreased it by 9.08 percentage points for underages individuals. It also found that at the cut off, our MLDA reduced arrests for individuals that are underage by an 136.9 arrests.

# 2 Data

In order to run our research, we used two different data sets. The first data set came from the well known National Health Interview Survey (NHIS). The data comes from their annual National Health Interview Sample files for the years of 1997 through 2007. This data from our NHIS data set comes from cross-sectional survey results that were conducted and collected through personal household interviews. The main purpose of the NHIS is to provide estimates of a wide range of health status and utilization that nationally represents non-institutionalized individuals of the United States. In our specific research, we obtained data on drinking behavior and demographic characteristics from our NHIS data set of our population of interest. The main data we collected for each individual was whether they drank within the last 30 days and their actual age, which included how far away they were from turning 21. We also acquired demographic data which included if they had a high school degree, their gender, their race (Black, White, or Hispanic), their employment status, their current work wage, and whether they were attending school or not. With this data we are able to estimate the effect of the MLDA on drinking.

Since our data from the NHIS is a cross-sectional survey, we understand that there is a possibility we might encounter the following three types of biases: measurement error, recall bias, and desirability bias. We conducted this research knowing that there is an existence of classical measurement error in our surveyed data set because the NHIS tampers with an individual's recorded age by adding or subtracting a predetermined error of up to 3 weeks in order to keep the surveyed individual private and maintain a level of anonymity. With this survey data we also run into systematic errors, one being recall bias where individuals may lack accuracy or completeness of the event or experience from their past. We may run into this because individuals might not be able to recall accurately if they were to have drunk in the last 30 days or not. Lastly, we may run into the second of the systemic biases known as desirability bias which may lead individuals to manipulate their answers in order to be more favorable. In our case, surveyed individuals might under report how much they drank because alcohol consumption for underage people is illegal and a socially taboo behavior which could lead to other external trouble.

Our second data set came from the California Monthly Arrest and Citation Register from the time period of 1979 to the year of 2006. The Monthly Arrest and Citation Register is a database which provides information on felony and misdemeanor level arrests for adults and juveniles in the state of California. From our data set, we obtain the arrest rates per 10,000 person-years, which is broken down by the different arrest types. The different crimes we include are both drinking related and non-drinking related. The list of crimes in our research include: DUI, drunk, possibly risk to self, violation of liquor laws, disorderly conduct, robbery, simple assault, and aggravated assault. With this data we are able to estimate the effect of the MLDA on the rate of arrests.

# 3 Empirical Methods

In our research we want to know if the MLDA helps reduce the amount of alcohol consumed and crimes committed by individuals who are below the age of 21. We do this through a regression discontinuity design, which leverages variation that results from a sharp cutoff that determines who gets a treatment, which in our context is the MLDA which allows 21 year olds to drink alcohol legally. We observe a discontinuity between our control and treatment groups because the treatment is based on a cutoff in a continuous variable. Then by observing this discontinuity, we are able to compare the individuals just right below and above our treatment cutoff. Since the only difference between our two groups is whether or not they are over or under the age of 21, we can assume that these two groups carry similar observable and unobservable characteristics, and only differ in their treatment. Then with these conditions met, we can examine the rate of the treatment and outcomes of people just above and below the cutoff.

In order to create our regression discontinuity figures, there was a necessity to choose a specific bin-width, bandwidth, y-range, and functional form in order to organize our raw represented data and visually analyze what was going on both sides of our cut off. A bin-width is the width of x values to group together. A bin-width of 30 days was selected because the resulting figures were not too noisy, had enough of the raw data represented, and there was a clear discontinuity at our cut off. The bandwidth is the overall range of x values to present across our continuous age variable. The bandwidth selected for the age range was 19 through 23 due to this range capturing the whole pattern that we are trying to observe. Another thing to consider with our bandwidth is what we do not want to capture. With our selected range we do not see any other discontinuities that could exist prior or after our cut off purely based on individuals behaving differently than those at our threshold. With this bandwidth we are confidently able to center our data around our cut off of interest with enough data points on either side of our threshold. A y-range is used to adjust the vertical axis range which will zoom in our graph to better show what is going on at our cut off. The selected y-range for each figure in our research differs based on the subject of our vertical axis. The ranges selected for each figure were selected in an effort to clearly see our discontinuity with no clutter and extra noise. The final specification for our figures and regression tables is the functional form which we selected to be our quadratic polynomial. We selected the quadratic fit because it goes through most of the points on each side of the threshold without making the discontinuity smaller and is statically significant at the 99.9% level unlike our cubic estimate. It also has the smallest standard error compared to the linear function.

We also create regression analysis tables as a way to check for balance between our two groups on either side of our cutoff. Our first stage (FS) equation calculates the effect of drinking at our cut off age of 21. Our reduced form (RF) equation calculates the effect of arrests at our cut off age of 21. The regression equations we use for our both observations are:

$$FS : Alcohol_i = \rho_0 + \rho_1 Over21_i + g((Age - 21)_A) + \epsilon_A \tag{1}$$

$$RF : Arrest_A = \phi_0 + \phi_1 Over21_A + h((Age - 21)_A) + v_A \tag{2}$$

Our first stage regression equation (1) calculates our estimates for our outcome $Alcohol_i$, where $Over21_i$ is our variable for being above 21. This equation has $g((Age-21)_A)$ as our functional form for the quadratic fit. Our $\rho_0$ is the value of the outcome variable when all of the independent variables are equal to 0 and our $\rho_1$ is the first stage estimate which is the effect of being able to drink legally on alcohol consumption. The $\epsilon_i$ is the error term associated with our first stage regression equation. Our reduced form regression equation (2) calculates our estimates for our outcome $Arrests_A$, where $Over21_A$ is the variable for being above 21. $h((Age - 21)_A)$ is the functional form for the quadratic fit that we've chosen. $\phi_0$ is the value of the outcome variable when all of the independent variables are equal to 0. $\phi_1$ is the reduced form estimate which is the effect of being able to drink legally on arrest rates and $v_A$ is the error term. Our regression tables that we create from these equations will also be using robust standard errors to reflect the possibility of heteroskadacity.

# 4 Results

Table 1: Linear, Quadratic and Cubic Regression Coefficients for Age based on Drinking

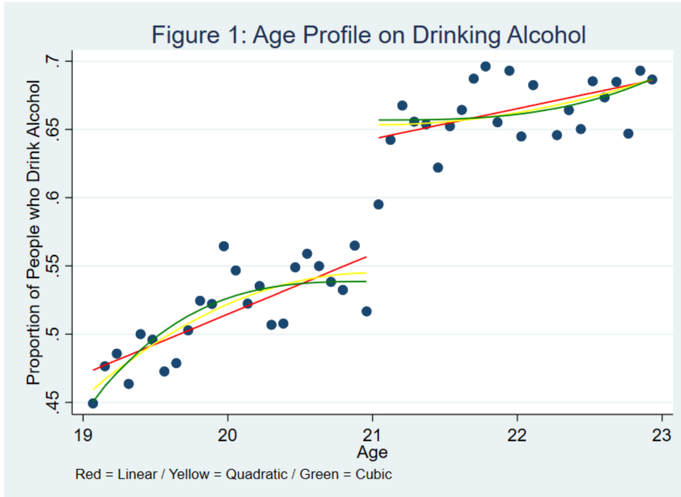
| | (1) | (2) | (3) |
|---|---|---|---|
| | drinks_alcohol | drinks_alcohol | drinks_alcohol |
| Over 21 | 0.0866*** | 0.0908*** | 0.0815** |
| | (6.13) | (4.26) | (2.84) |
| Age | 0.0439*** | -0.0236 | -0.0509 |
| | (4.89) | (-0.66) | (-0.57) |
| Age * Z | -0.0243* | 0.0969 | 0.207 |
| | (-1.97) | (1.96) | (1.67) |
| $\text{Age}^2$ | | -0.0340 | -0.0681 |
| | | (-1.95) | (-0.65) |
| $\text{Age}^2 * Z$ | | 0.00727 | -0.0618 |
| | | (0.30) | (-0.43) |
| $\text{Age}^3$ | | | -0.0114 |
| | | | (-0.33) |
| $\text{Age}^3 * Z$ | | | 0.0457 |
| | | | (0.97) |
| Constant | 0.559*** | 0.536*** | 0.532*** |
| | (54.95) | (35.00) | (25.65) |
| Observations | 18801 | 18801 | 18801 |
| $R^2$ | 0.025 | 0.025 | 0.025 |

$t$ statistics in parentheses

$^{*}$ $p < 0.05$, $^{**}$ $p < 0.01$, $^{***}$ $p < 0.001$

The RD figure for Age profile based on Drinking Alcohol shows us a clear discontinuity at our cut off of 21. Based on our quadratic regression calculations which we chose for our interpretation, it showed there is an approximately a 9.08 percentage point spike increase in drinking as soon as someone turns 21. This result means that the MLDA does in fact reduce underage drinking. This discontinuity intuitively makes sense regardless of the particular specification choices personally made for this research because at the age of 21, it becomes legal for anyone of the MLDA to purchase and consume alcohol.

By looking at our demographic regression table, we see that all of the 10 demo-




graphic characteristics are very uncorrelated with whether you are over 21 or not. When it comes to our regression balance table, we have to consider the occurrence of multiple inferences, which happens when we do more than one significance test for a given table. In our case we do 10 significance tests, one for each variable in our quadratic regression tables (Table 2 and 3). The probability that one or more of our estimates is significant can be calculated using the following formula: 1-(1-0.5)$^n$, where n is the amount of significance tests that are being run in your analysis. In our case we have 10 different variables we are running significance tests for so our probablity for at least one of them being statistically significant is 1-(1-0.5)$^1$0 = 0.401, which is a 40.1 percent of the time. However, all our results are not close to being statistically significant (not capturing any of the multiple inference) which lets us assume that at our cut off,our two groups on each side of our discontinuity is well balanced and that the discontinuity we see is the true effect of just the MLDA alone on drinking.

Table 2: Quadratic Regression Coefficients for Age based on Demographic Variables

| | (1) Black | (2) White | (3) Male | (4) Working Wage | (5) Attending School |
|---|---|---|---|---|---|
| Over 21 | -0.0236 | 0.0133 | 0.0225 | 0.0317 | 0.000886 |
| | (-1.50) | (0.61) | (1.04) | (1.50) | (0.05) |
| Age | 0.0172 | -0.00780 | -0.0731* | -0.0149 | -0.0611* |
| | (0.65) | (-0.21) | (-2.00) | (-0.41) | (-2.07) |
| Age * Z | 0.00653 | 0.0144 | 0.106* | 0.0546 | 0.0430 |
| | (0.18) | (0.28) | (2.09) | (1.12) | (1.13) |
| Age$^2$ | 0.00429 | -0.000207 | -0.0251 | -0.0365* | -0.00151 |
| | (0.34) | (-0.01) | (-1.41) | (-2.06) | (-0.10) |
| Age$^2 * Z$ | -0.0137 | -0.00395 | 0.0113 | 0.0338 | -0.00943 |
| | (-0.78) | (-0.16) | (0.46) | (1.43) | (-0.51) |
| Constant | 0.160*** | 0.554*** | 0.412*** | 0.618*** | 0.165*** |
| | (13.92) | (35.31) | (26.40) | (40.15) | (13.69) |
| Observations | 18801 | 18801 | 18801 | 18801 | 18801 |
| $R^2$ | 0.000 | 0.000 | 0.001 | 0.014 | 0.020 |

$t$ statistics in parentheses

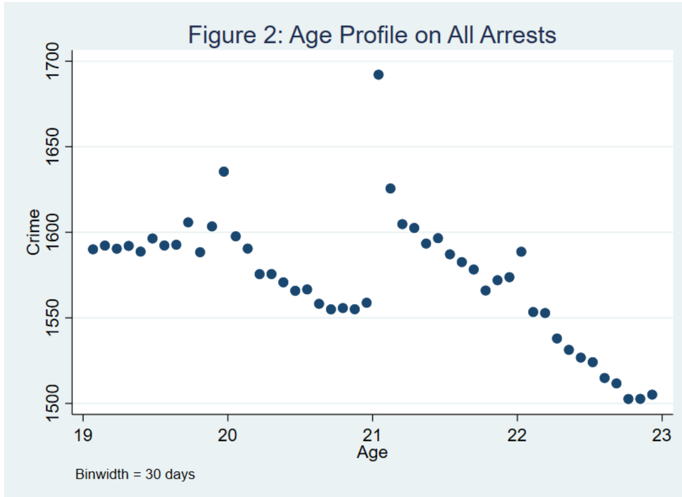
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3: Quadratic Regression Coeffecients for Age based on Demographic Variables

| | (1) High School Diploma | (2) Hispanic | (3) Employed | (4) Married | (5) Uninsured |
|---|---|---|---|---|---|
| Over 21 | 0.0328 | -0.00414 | 0.0286 | -0.0244 | -0.000362 |
| | (1.93) | (-0.22) | (1.35) | (-1.60) | (-0.02) |
| Age | -0.0636* | 0.00224 | -0.0134 | 0.0671** | -0.00133 |
| | (-2.15) | (0.07) | (-0.37) | (2.84) | (-0.04) |
| Age * Z | 0.0856* | -0.0212 | 0.0553 | -0.0469 | -0.0142 |
| | (2.17) | (-0.49) | (1.13) | (-1.34) | (-0.30) |
| Age$^2$ | -0.0436** | 0.00110 | -0.0365* | 0.00806 | -0.0139 |
| | (-2.98) | (0.07) | (-2.06) | (0.74) | (-0.85) |
| Age$^2 * Z$ | 0.0341 | 0.00792 | 0.0326 | 0.00826 | 0.0262 |
| | (1.77) | (0.38) | (1.37) | (0.49) | (1.16) |
| Constant | 0.793*** | 0.242*** | 0.618*** | 0.158*** | 0.309*** |
| | (63.12) | (18.02) | (40.11) | (14.43) | (21.20) |
| Observations | 18801 | 18801 | 18801 | 18801 | 18801 |
| $R^2$ | 0.003 | 0.000 | 0.014 | 0.019 | 0.001 |

$t$ statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$




The RD figure for age profile based on the rate of all arrests shows that there is a discontinuity at our MLDA cut off. Visually based on our graph we see an outlier at the 20, 21, and 22 age mark, which is what we call the birthday effect, where on an individual's birthday they tend to go out and drink more which causes them to get into behavior that gets them arrested more often. We control for this in our regression table, by creating a birthday variable Also, based on our quadratic regression calculations which we chose for our interpretation, it showed there are approximately 10,000 persons-years, there are 136.9 more arrests at our cutoff of 21. This result means that the MLDA does in fact reduce the rate of overall arrests for under aged individuals. For our subcauses we group our drinking related arrest types in figure 3 and all other arrest types in figure 4. We see from our regression table, that 2 (DUI and Drunk Risk to Self) out of our 3 drinking related arrest types (the other being violation of liquor laws) are highly statistically significant with a t-critical value of way over 1.96. Based on figure 3 you can see huge discontinuities between DUI (which increases by 54.34 arrests) and Drunk risk to self (which increases by 36.07 arrests). Violation of liquor laws however goes down by almost an estimate of 65 arrests which makes sense cause once you are 21 you are legally able to have possession of alcohol.

By using our regression table and our figure 4, we can see that there are discon-

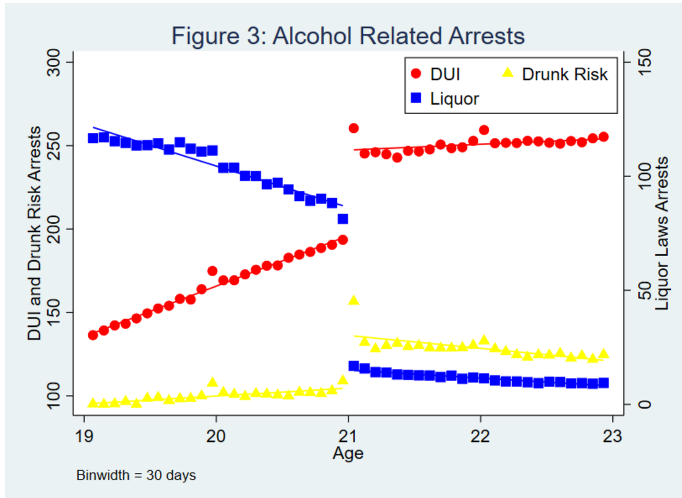


Figure 3: Alcohol Related Arrests

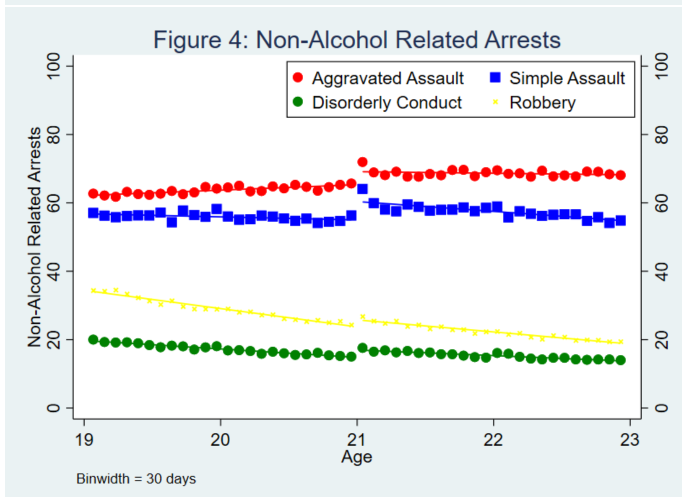


Figure 4: Non-Alcohol Related Arrests

tinuities in all of our non drinking related arrest types as well, however only ranging from 1-5 more arrests at the cut off per 10,000 persons. All of the non drinking related arrest types which include robbery, simple assault, disorderly conduct, and aggravated assault all have spikes right at the 21 cut off and have outliers at the 20th, 21st and 22nd birthday marks. But after the cut off the rate of these arrests all have downward slopes. Based on the regression table all of our arrest types have statistically significant estimated values for being over 21 which means that this additional evidence that drinking is driving the discontinuity for arrest rates.

Table 4: Quadratic Regression Coefficients for Age based on Arrest Type

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
|---|---|---|---|---|---|---|---|---|
| | All Arrests | DUI | Violation of Liquor Laws | Disordely Conduct | Robbery | Simple Assault | Drunk Risk to Self | Aggravated Assault |
| Over 21 | 136.9*** | 52.76*** | -64.97*** | 2.012*** | 1.732*** | 5.247*** | 33.69*** | 4.045*** |
| | (18.80) | (18.36) | (-78.23) | (5.14) | (4.33) | (6.35) | (9.54) | (5.45) |
| Age | -210.5*** | 24.99*** | -33.61*** | -1.505* | -2.901*** | -1.247 | 4.224 | 1.446 |
| | (-31.19) | (10.20) | (-17.39) | (-2.50) | (-4.12) | (-1.15) | (1.94) | (1.26) |
| Age * Z | 139.2*** | -21.93*** | 26.54*** | -0.830 | -2.205* | -1.091 | -19.92** | -2.823 |
| | (12.80) | (-3.51) | (13.30) | (-0.93) | (-2.30) | (-0.59) | (-2.72) | (-1.65) |
| Age$^2$ | -88.39*** | -2.507* | -7.736*** | 0.509 | 1.232*** | -0.115 | -0.0782 | -0.0602 |
| | (-37.95) | (-2.00) | (-8.16) | (1.76) | (3.42) | (-0.21) | (-0.07) | (-0.11) |
| Age$^2 * Z$ | 88.30*** | 2.977 | 9.548*** | -0.109 | -0.422 | -0.145 | 4.538 | 0.599 |
| | (26.04) | (1.07) | (9.78) | (-0.26) | (-0.89) | (-0.17) | (1.44) | (0.72) |
| Birthday | 634.5*** | 129.2*** | 5.134*** | 14.91*** | 5.198*** | 38.01*** | 193.9*** | 23.70*** |
| | (103.72) | (47.34) | (22.62) | (49.41) | (17.57) | (54.10) | (56.51) | (41.39) |
| Constant | 1496.6*** | 194.2*** | 81.16*** | 15.12*** | 24.55*** | 54.86*** | 104.4*** | 65.27*** |
| | (378.39) | (217.06) | (101.59) | (60.55) | (91.30) | (126.26) | (125.38) | (138.45) |
| Observations | 2190 | 1460 | 1460 | 1460 | 1460 | 1460 | 1460 | 1460 |
| $R^2$ | 0.756 | 0.943 | 0.990 | 0.340 | 0.719 | 0.148 | 0.688 | 0.249 |

$t$ statistics in parentheses

$^* p < 0.05$, $^{**} p < 0.01$, $^{***} p < 0.001$

# 5 Conclusion

From the analysis of our resulting regression tables and our regression discontinuity figures, we found that the minimum legal drinking age does in fact reduce the amount of underage drinking and arrests for many different types of crimes (except for the violation of liquor laws). We have brought up the concern of biases existing in our first stage estimates that may screw up our results. As we discussed, the measurement error imposed by the NHIS spreads out our age estimates which leads our estimate to be closer to zero which would lead to downward bias. If there is an existence of recall bias, and individuals are unable to remember accurately how many drinks they had in the last 30 days, that will introduce additional random noise in our data which will mess with our estimates as a whole. Lastly if there is an existence of desirability bias within our data set for those underage that end up under reporting, it will cause upward bias in the over 21 group.

With our RD design, we are able to view the true effects of drinking on arrests, through an instrumental variables (IV) approach where we use the estimates of our first stage which involved drinking based on the MLDA, and divide those by our reduced form which was our estimates for our arrest types based on the MLDA. This is useful for our analysis because by calculating the IV we can find estimates of drinking on arrests that do not incorporate an individual's decision to drink(strictly following the MLDA). By dividing our first stage estimates by our reduced form, we are able to find our local average treatment effect IV point estimate for overall arrests comes out to 893.74. This IV estimate for overall arrests is calculated for only our compliers to the MLDA, so the population of interest only includes those who didn't drink before the cut off and drank after they are of the MLDA of 21. This means that if everyone complied to not drink before turning 21 and then started drinking when they turned 21, the IV point estimate would be the effect of drinking on arrests for everyone that complied.

| Arrest Type | All Arrests | Aggravated Assault | Simple Assault | Robbery |
|---|---|---|---|---|
| IV Estimate (LATE) | 893.74328 | 43.623048 | 64.112949 | 23.648796 |

| Arrest Type | DUI | Drunk Risk | Violation of Liquor Laws | Disorderly Conduct |
|---|---|---|---|---|
| IV Estimate (LATE) | 549.53809 | 366.56794 | -823.83599 | 25.055707 |

When it comes to the practicality, it is hard to accept this IV estimate purely because based on our figure 1 which represents age profile on if an individual drinks or not there is a huge population of non-compliers of the MLDA which results in our IV estimate not really representing what actually happens. There are three assumptions that have to be met in order for an IV estimate to be valid. The first assumption is that the first stage estimate can not be equal to 0. We know this is true because we see a discontinuity in drinking at our cut off age. The second IV assumption is that both our reduced form and first stage estimates are consistent and unbiased. We also know this to be true because there is no element of selection bias into the treatment group because we are observing the two groups at a cut off. The last assumption is that drinking is the only thing that affects an individual to get arrested at the MLDA cut off. We know this is not true because our figure 4 showed that at our cut off non-drinking related arrests also went up. This could also be due to new environments that are made accessible for 21 year old's, such as clubs, casinos, or bars. Even though we cannot use the IV estimation as an accurate indicator of whether drinking alcohol causes arrests, our research still shows signs that the MLDA is effective and does decrease the rate of drinking and arrests for individuals who are underage.