Title: Mortality Rates Correlate with Drinking Age in the United States

Authors: Mustafa Barez, Jie Huang

Date: April 14th, 2020

Abstract

The present paper studied the effects of legal drinking age on death rates using regression discontinuity design. The study found that the legal drinking age does play a role in significant increases in death rates due to age policies in the United States. This rise in death rates presents evidence of hazardous drinking habits for individuals over the age of 21 and policies noting age restrictions can have a significant impact on these mortality rates.

Introduction

Alcohol consumption brings a number of questions as to the public health policies examined in the United States. According to a study by SafetyLit, alcohol is the third leading preventable cause of death in the United States and is associated with a number of health consequences. These include liver cirrhosis, violence, and death (Centers for Disease Control and Prevention, 2004). The current legal drinking age is 19 and studies show that alcohol is a critical factor in mortality rates for a number of individuals in this age group (Minino et al., 2000). However, other studies present the age of 21 as being the spike in death rates for alcohol consumption (McCartt et al., 2010). In order to provide some evidence for this, the following study examines the relationship between age and mortality rates using regression discontinuity design. Regression discontinuity design is an experimental approach in gaining causality by applying a treatment to gain a cut-off on a continuous interval (Hahn et al., 2001). Here causality is defined as one variable being affected by another. With this, the following study seeks to determine whether or not age has a casual relationship with mortality rates. This is in terms of alcohol in the United States. The study will also determine whether a significant difference exists between age groups above and below 21 years of age for mortality rates. By determining this, policymakers may look to put greater restrictions on public drinking for persons who are around 21 years old. The present study presents a Regression Discontinuity Design experiment comparing age and mortality rates from a United States dataset. The study presents a cut-off point through scatterplot analysis and provides evidence of a causal relationship between age and mortality rates through a t-test analysis. These statistical results will be a strong determinant in presenting a significant difference in death rates between age groups.

Dataset

The following dataset is extracted from the mastering metrics website with the link found in the appendix. The dataset was also found in the journal titled: "the Effect of Alcohol Consumption on Mortality: Regression Discontinuity Evidence from the Minimum Drinking Age" written by Christopher Carpenter and Carlos Dobkin. Here the same dataset in this study is used to determine alcohol consumption consequences based on quantitative information provided.

The attributes in the dataset are as follows: agecell, all, allfitted, internal, internalfitted, external, externalfitted, alcohol, alcoholfitted, homicide, homicidefitted, suicide, suicidefitted, mva, mvafitted, drugs, drugsfitted, externalother, and externalotherfitted. Certain attributes are further described and filtered in the later sections of this paper. It should be noted that due to the attributes provided, the dataset of death rates in the United States focuses on a wide range of root causes that include homicides, suicides, and alcohol consumption. The dataset also presents data that relates to the respective cause of death such as age.

Ethics

When examining the study, taking into account ethical dilemmas is critical. One of the first areas of focus is awareness. The present dataset contains details of individuals and also ensures their anonymity. By ensuring the privacy of the individuals in the study, the datasets collected follow correct ethical procedures in this area. Another area concerns honesty. The present data depicted appears not to misuse data and does not show any bias. This is observed from the descriptions of the attributes of the dataset and the precision of the data. The attributes are also open to the public and research design processes are mentioned in order to avoid conflicts. In terms of the actual data presented, there were no signs of discrimination as groups of visible minorities were not isolated, nor mentioned. According to Gregory (2003), an important ethical issue is the source of funding for research as the relationship between the head researcher and person providing the funding should be known. This way, the public can understand which individuals should be acknowledged for starting the research project and assisting in its completion. The present dataset did not provide details regarding this issue and can be seen as an area for improvement.

Weaknesses

In order to gain a strong understanding of the study, it is important to examine the shortcomings. One of the critical areas that can be improved would be information regarding the source of the data. By providing information on how the data came about and was collected would allow researchers to better understand its authenticity. Another area of improvement revolves around attribute descriptions. Some of the attributes in the dataset, such as agecell, are not as descriptive

and require further research to understand its purpose. By creating separate sections to present details of the dataset, researchers would not be required to spend time locating details in order to describe this to readers. In terms of the scatterplots, Plot 2. presents disadvantages at the cut-off point. Here some of the points are located in the middle are difficult to explain as they do not fit in the two categories and create problems when explaining the regression discontinuity design. Although this is presented in Plot 2. The number of data points is minimal and therefore can be regarded as outliers.

Discussion

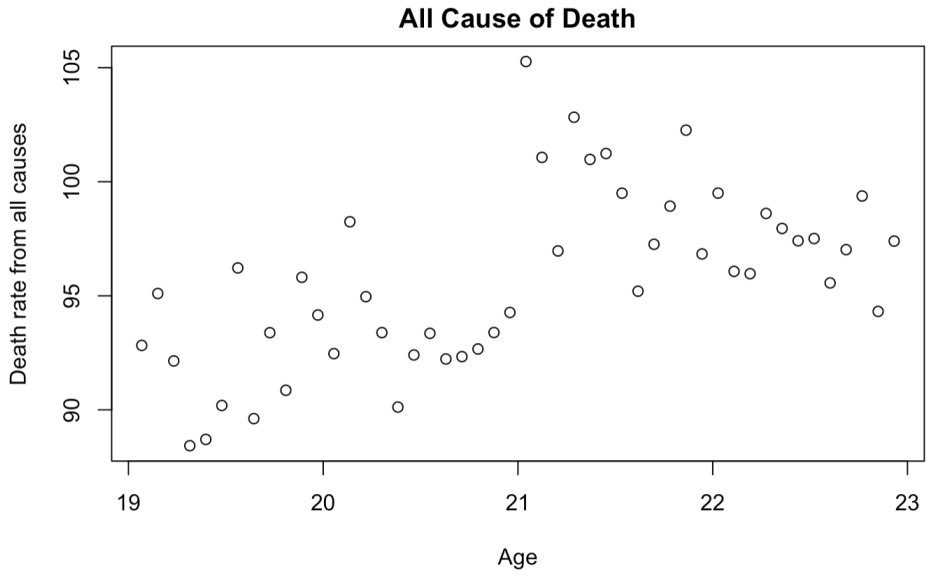
In order to answer the research question, we decided to use the Regression Discontinuity Design method to approach an answer. The motivation for this method is that Regression Discontinuity Design visually demonstrates the outcomes for both the control group and treatment group with a clear cutoff point. Thus we can tell how different the treatment group is compared to the control group. However, before starting the method, there are several assumptions that need to be made. First, whether an individual observation has a treatment is assumed to be random. Second, we need to assume that the observations can not manipulate their independent variable values. Third, we assume that whether an observation's independent variable value falls immediately under or above the cutoff point is random as well. Now we can proceed to conduct Regression Discontinuity Design for the research question.

We were interested in understanding the basics of our data, thus, we summarized the characteristics of variable Age and Death Rates as shown in Table 1. The variable Age ranges from 19.70 to 22.93, and the variable Death Rate ranges from 88.43 to 105.27. Since our research is interested in how the legal drinking age in the US affects the death rate, such ranges of the variables' value are suitable for the study.

Age	**Minimum Value**	**Maximum Value**	**Mean**	**Standard Deviation**
	19.70	22.93	21.00	1.13
Death Rate	**Minimum Value**	**Maximum Value**	**Mean**	**Standard Deviation**
	88.43	105.27	95.67	N/A

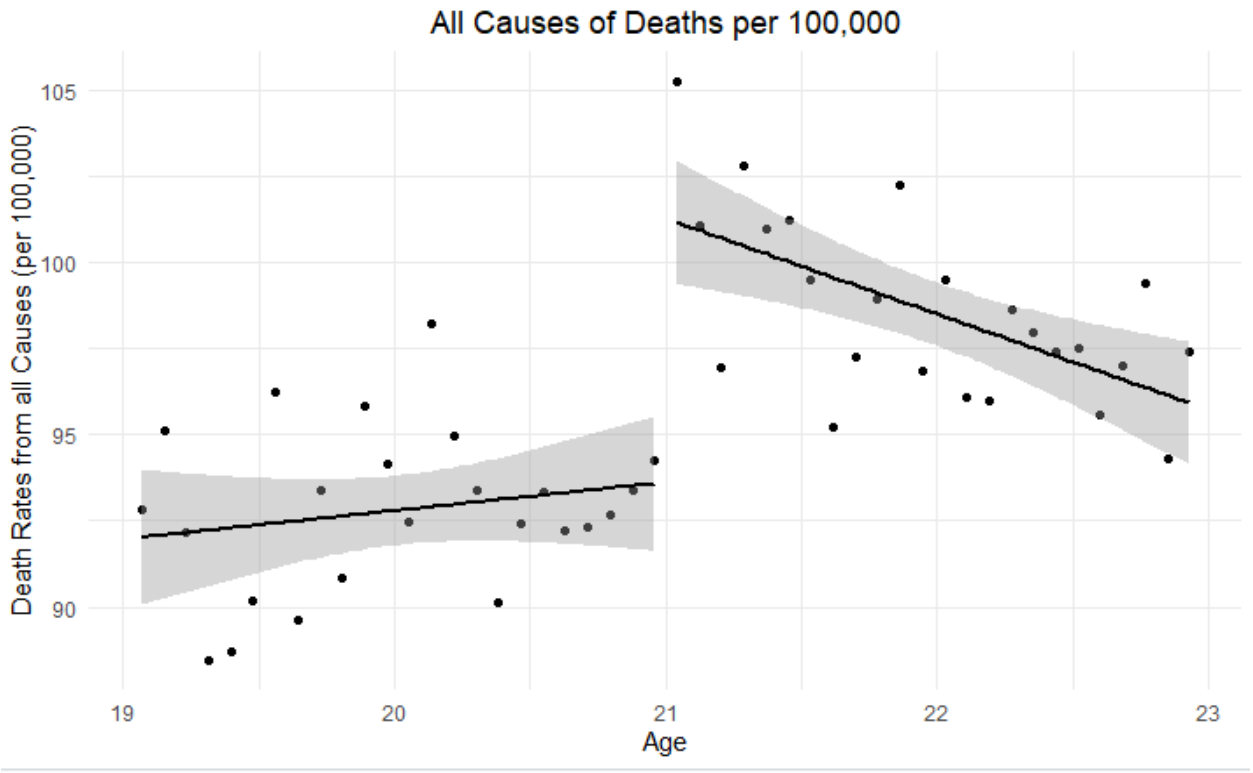
Table 1. Summary of Independent and Dependent Variables

Next, we were interested in checking the distribution of corresponding variable data points. The distribution of data points will help us to briefly understand the trend of variable Death Rate in relation to variable Age. As shown in Plot 1, the maximum death rate is located at around the age of 21, and the minimum death rate is spotted at between the ages of 19 and 20. The plot briefly helps us to realize that age of 21 is a special point that is followed by the highest death rate.



Plot 1. Scatterplot Comparing Age and Death Rates

From here we proceeded with our Regression Discontinuity Design analysis using the rdd package. The RDestimate function allowed us to output a model that shows the dependent variable difference between observations that fall under and over the cutoff point. In addition, the model contains two linear regression lines at the threshold. Since we are interested in the legal drinking age, we put 21 as a cut point in the RDestimate function. As shown in Plot 2, we can see a clear gap between these two linear models around the cut off point and since the boundaries are not sharp, the scatterplot displays fuzzy Regression Discontinuity Design properties. This result took us closer to the success of our initial attempt, which is being able to explain the causal relationship between legal drinking age and death rate.



Plot 2. Scatterplot Displaying RDD Model containing Cut-Off

We further our analysis by implementing a summary function on the Regression Discontinuity model. As Table 2 indicated, the LATE, which stands for the local average treatment effect, has an estimate of 9.001 with a p-value of 1.199e-09. These results indicate that the effect of the treatment, which is being 21 years old or older, is very significant in terms of affecting the death rate. We can be more confident to conclude that the legal drinking age presents a causal relationship with the death rate. More specifically, the policy of the legal drinking age causes the death rate to increase at the age when people can start drinking alcohol.

	Estimate	**p-value**
LATE	9.001	1.100e-09

Table 2. Summary of Regression Discontinuity Model

To test the validity of our result, we decided to calculate the difference in the average death rate between groups that are under and over the age of 21. We removed all the missing values from the dataset and used the mean function to calculate the average death rate from the two groups, then subtracted the average death rates. The difference between the control group and the

treatment is 5.740044, which is significant considering the death rate is measured with the unit of 100,000 persons.

In addition, we fitted two linear regression models on the scatterplot for age under 21 and over 21 to see the difference in the death rate of these two groups. The models for the study are:

Death_Rate_Under21 = 0.827 * Age + 76.2515
Death_Rate_Over21 = -2.7764 * Age + 159.5847

As the models suggest, if we put 21 as the Age value in the first model, the death rate will be 93.6185. However, if we put 21 as the Age value in the second model, the death rate will become 101.2803, which suggests the higher death rate for people after they pass the legal drinking age.

Furthermore, we decided to test the result by conducting a two-sample t-test. This test can tell us whether a significant difference exists in the average death rate between groups that are under and over the age of 21. As shown in Table 3, the t-test yielded a p-value of 5.185e-10, from which we can conclude that the difference is significant and cannot be overlooked.

t	df	p-value
-7.8593	45.432	5.185e-10

Table 3. Two-Sample T-Test

In terms of the statistical uncertainty, which is also the weakness of Regression Discontinuity Design, the assignment of observations to the control and the treatment groups is not random. In this case, a person who is 21 or older will automatically be assigned to the treatment group. This way of assigning groups could bring statistical uncertainty and biases to the analysis. For example, a person who does not drink alcohol, and turning 21, will be assigned to the treatment group and display behaviors from the control group. In conclusion, Regression Discontinuity Design is dependent on the assumptions made on randomness.

Conclusion

To conclude, the following study examined Regression Discontinuity Design through age groups and death rates and presented causality between one another through scatterplot analysis. This was confirmed through statistical analysis using R packages. The analysis displayed a fuzzy Regression Discontinuity Design pattern with a cut-off at the age of 21. The two-sample t-test displayed a significant difference between average death rates of groups younger and older than

21. Moving forward, researchers could use the developments of this study to provide further information as to why death rates for these age groups differ significantly.

Appendix

The dataset can be found at http://www.masteringmetrics.com/resources/ on chapter 4.

The Github link to the code can be found at https://github.com/Mustafa-barez/Drinking-Age-Mortality-Rates-through-R/blob/master/PS5/PS5.R

References

broom. David Robinson and Alex Hayes (2019). broom: Convert Statistical Analysis Objects into Tidy Tibbles. R package version 0.5.3. https://CRAN.R-project.org/package=broom

Centers for Disease Control and Prevention. (2004). Alcohol-attributable deaths and years of potential life lost--United States, 2001. *MMWR: Morbidity and mortality weekly report*, *53*(37), 866-870.

dplyr. Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 0.8.5. https://CRAN.R-project.org/package=dplyr

foreign. R Core Team (2019). foreign: Read Data Stored by 'Minitab', 'S', 'SAS', 'SPSS', 'Stata', 'Systat', 'Weka', 'dBase', ... R package version 0.8-72. https://CRAN.R-project.org/package=foreign

Gregory, I. (2003). *Ethics in research*. A&C Black.

Hahn, J., Todd, P., & Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression‐discontinuity design. *Econometrica*, *69*(1), 201-209.

knitr. Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.28. https://CRAN.R-project.org/package=knitr

Minino, A. M., & Smith, B. L. (2001). Deaths: preliminary data for 2000.

McCartt, A. T., Hellinga, L. A., & Kirley, B. B. (2010). The effects of minimum legal drinking age 21 laws on alcohol-related driving in the United States. *Journal of Safety Research*, *41*(2), 173-181.

Rdd. Drew Dimmery (2016). rdd: Regression Discontinuity Estimation. R package version 0.57 https://CRAN.R-project.org/package=rdd

Tidyverse. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686