Problemset5

Sharon Allman, Diego Mamanche Castellanos & Ke-Li Chiu

11/04/2020

Abstract

The objective of this analysis was to examine if there was a relationship between [...]

Introduction

It is always a contentious topic whether we should revise the legal drinking age. This can be impacted by things such as an argued increased propensity for a spike in crime near the legal drinking age, especially crimes committed under the influence in alcohol. Previous studies have inconsistent findings when it comes to whether an increased availability of alcohol leads to an increase in certain types of offences such as assault, rape, and more.

Within a Canadian context, a study by Callaghan et al. had striking findings that suggested that the drinking age of 18 or 19 depending on province of residence could be problematic in the sense that there was a demonstrable increase of crime at the legal drinking age (Callaghan et al., 2016). The drinking ages in Canada are comparatively low to 21 years old in the U.S. Using the replication data of The Minimum Legal Drinking Age and Crime by Christopher Carpenter and Carlos Dobkin (2015), we sought to elaborate the findings of Billings (2014) in their research on the age-prohibition of alcohol and potentially increased propensity for crime. Particularly, we were interested in the crimes with High-alcohol usage from this study.

Research Question & Hypotheses

Within this research, we were interested in examining whether there is an effect of an increase in certain types of crimes after perpetrators reach the U.S. legal drinking age (21). This could be for a variety of reasons, most importantly being easier access to alcohol and how that could lead to a spike of crimes committed while the perpetrator was intoxicated.

H0: The U.S. legal drinking age has no significant effect on the incidence of certain crimes.

H1: The U.S. legal drinking age has a significant effect on the incidence of certain crimes.

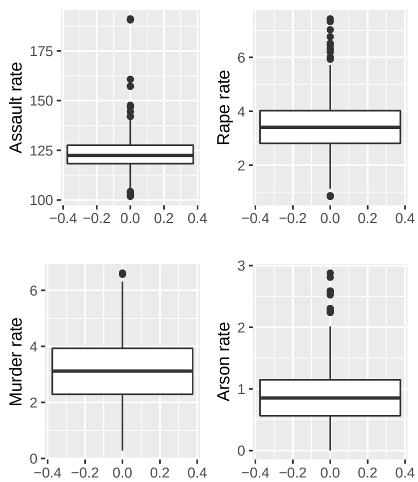
H2: The U.S. legal drinking age has a significant positive effect on the incidence of certain crimes (whereby it causes an increase in occurrence).

Dataset Description and Data Cleaning

The dataset is the replication data of The Minimum Legal Drinking Age and Crime by Christopher Carpenter and Carlos Dobkin (2015). The dataset contains 2922 observations and 144 variables. The observation is arrestee's age in days relative to 21 on the day being arrested ranging from 17 to 24. For example, -30 indicates that the arrestee is a month away from turning 21, 60 indicates that the arrestee is 21 and two months old and 0 means the arrestee is arrested on their 21st birthday. The original crimes data was retrieved from California's Monthly Arrest and Citation Register dated from 1979 to 2006, containing major crime types categorized by FBI: violent crime, alcohol-related offenses, property crime, illegal drug possession or sale, and all other offenses. The dataset contains the counts of arrests as well the arrest rates per 10,000 person-years for each crime for each observation. For our study, we followed the bandwidth of two years Carpenterr and Dobkin used; we sliced the dataset to include observations of people age 19 to 22 inclusive (from day -730 to day 729).

An additional binary variable called mlda_and_over was created based on the days_to_21 variable, classifying MLDA and over as 1, and under MLDA as 0. Finally, the dataset was filtered by the columns with the High-alcohol usage crimes under analysis (Assault, Rape, Murder, and Arson).

The distribution and descriptive statistics of the High-alcohol usage crimes is shown in Figure 1.



stats	days_to_21	mlda_and_over	assault	rape	murder	arson
nbr.val	1459.00	1459.00	1459.00	1459.00	1459.00	1459.00
nbr.null	1.00	729.00	0.00	0.00	0.00	64.00
nbr.na	0.00	0.00	0.00	0.00	0.00	0.00
min	−729.00	0.00	101.83	0.84	0.28	0.00
max	729.00	1.00	191.12	7.43	6.62	2.88
range	1458.00	1.00	89.30	6.59	6.34	2.88
sum	0.00	730.00	179337.11	5023.48	4610.78	1356.98
median	0.00	1.00	122.42	3.40	3.12	0.85
mean	0.00	0.50	122.92	3.44	3.16	0.93
SE.mean	11.03	0.01	0.20	0.03	0.03	0.01
CI.mean.0.95	21.64	0.03	0.39	0.05	0.05	0.03
var	177511.67	0.25	57.49	0.98	1.07	0.28
std.dev	421.32	0.50	7.58	0.99	1.04	0.53
coef.var	Inf	1.00	0.06	0.29	0.33	0.57

Figure 1: Descriptive statistics of the data

Method

Because the purpose of this paper is to validate the effect of the Minimum Legal Drinking Age (MLDA) over the most alcohol influenced crimes, we conducted Regression Discontinuity Design (RDD). The RDD estimates impact around the eligibility cutoff, in our case MLDA, as the difference between the average outcome for units on the treated side (MLDA and over) of the eligibility cutoff and the average outcome of units on the untreated (under MLDA) side of the cutoff (MLDA).

According to the researchers of our reference study (Billings, 2014), after applying their research through several surveys to inmates and other techniques, they found the fifteen most affected crimes by Alcohol consumption. For this investigation, we focused on the crimes classified as High-alcohol usage from that paper. These are Assault, Rape, Arson, and Murder, based on the role alcohol plays when an offender commits a given crime.

Ethical Issues

The findings of studies like this can have grave policy implications at a national level. One of the major ethical considerations within this study is the need to accurately and completely represent the findings of the data. As this data relates to the need to properly control and regulate alcohol, it is important that we do not overrepresent the gravity or implications of our findings. If it is true that there is an increase in certain crimes at the drinking age, there is a delicate balance between a potential increase in youth offences if the drinking age is lowered verses an economic implication if the drinking age is raised.

Additionally, it is very difficult to study the actual effect of alcohol consumption on crime rate due to the many confounding variables that could impact this relationship. Also, gathering this data can be difficult and lead to inconsistencies. Much of this data would be self-reported by the victim, police, or the perpetrator after the fact, so it may be less reliable in some cases. However, as mentioned, this is important to study as it could have serious policy implications.

Weaknesses and Limitations

Limitation in dataset

The original data was collected from California's Monthly Arrest and Citation Register (MACR) for the period 1979 to 2006 and California Health Interview Survey (CHIS). To validate the replication, we would need access to both MACR and CHIS. However, their data are not accessible to the public without permission. Therefore, we do not have the freedom in data transformation, and we are assuming the process of data collection and transformation to generate this replication dataset is reliable.

One limitation of this data set is that there is no variable for actual alcohol consumption. Although many studies have stated that there is high correlation between alcohol consumption and crime rate, we cannot know if there is significant number of individuals observed in this dataset who have committed their crimes under the influence of alcohol. Even if the usage of alcohol was proven, the effect of alcohol on crime arrest rate could be explained in a plethora of different ways. Moreover, in order to retrieve the age information, the dataset relies on arrest data instead of reported crimes data. The concern raised by the data collection is that alcohol use might be increasing the chance of being arrested, not the crime commission.

Limitation in approach

External validity of estimates is hard to assess with regression discontinuity design (cite Rohan) because the estimates are valid only for observations very near the cutoff points. In our case, the estimates can only be applied to people who are very near their 21st birthday.

The age cut-off is not unique to legal drinking age and could be contaminated by other rights and responsibilities an individual obtains once they have turned 21. For example, when an individual is 21 years old, they can legally drink at restaurants, bars, and pubs. In this case, it is difficult to clarify whether it is the alcohol consumption, the increased social interaction, or the combination of the two that increases crime incidents.

Another weakness in our approach is that "birthday celebration effect" that is observed by the author of the dataset where crime arrest rates have pronounced peaks in arrestees' 19th, 20th, 21st and 22nd birthdays. However, the birthday celebration effect is not considered in our analysis.

Finally, we used linear regression to estimate the effect of legal access to alcohol to crime arrest rates while there are other regression models available. For example, the author of the dataset used polynomial regression. The different choices in regression method resulted in different estimates.

Analysis

Assault

Assaults accounted for most of the sharp increase in arrests exactly at age 21. As Figure X shows, it is visually clear that there is a discontinuity at the cutoff. The result in Table X confirms that there is a significant increase of 9.28 (7.9%) arrests at age 21 for assault crimes.

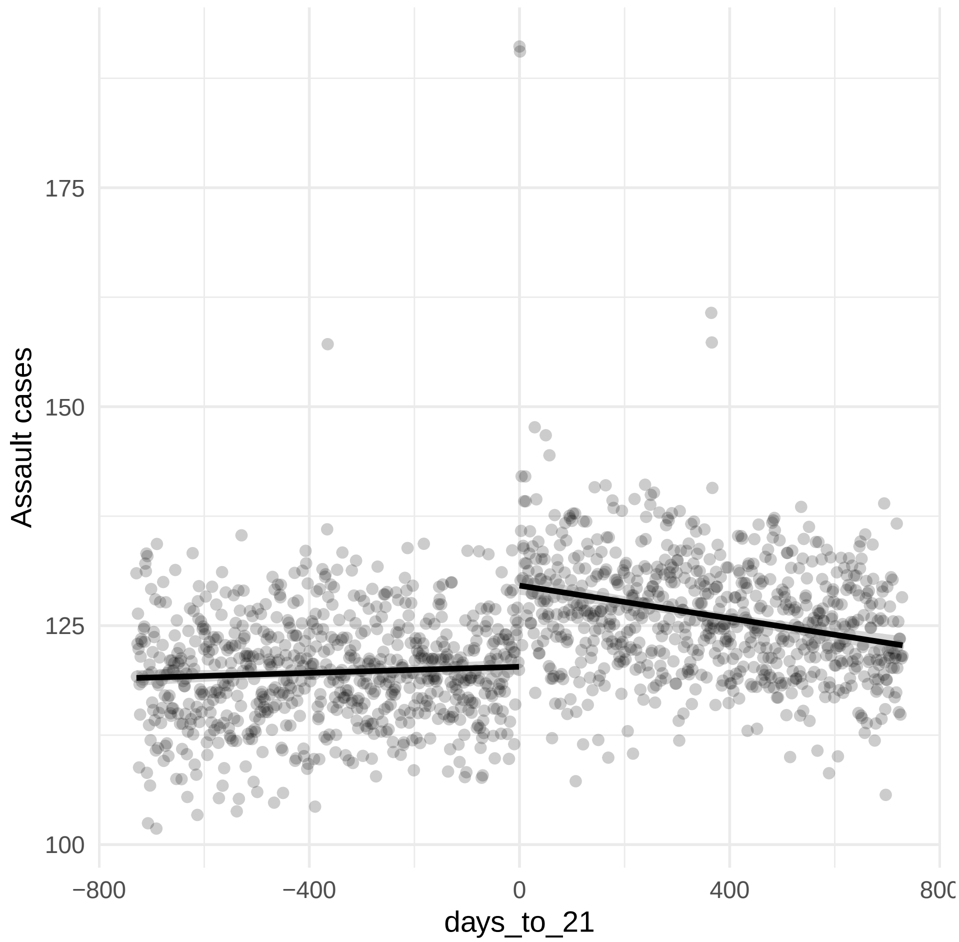


Figure 2. caption

term	estimate	std.error	statistic	p.value
(Intercept)	118	0.399	297	0
days_to_21	-0.00381	0.000846	-4.51	7.13e-06
mlda_and_over	9.28	0.713	13	9.9e-37

Table 1. Caption

Rape

The visual evidence of the discontinuity for rape arrest cases in Figure x is not as apparent as in assault crimes. However, Figure x shows that there is a statistically significant increase in rape arrest cases at the age 21 (9.1%), estimated as 0.302.

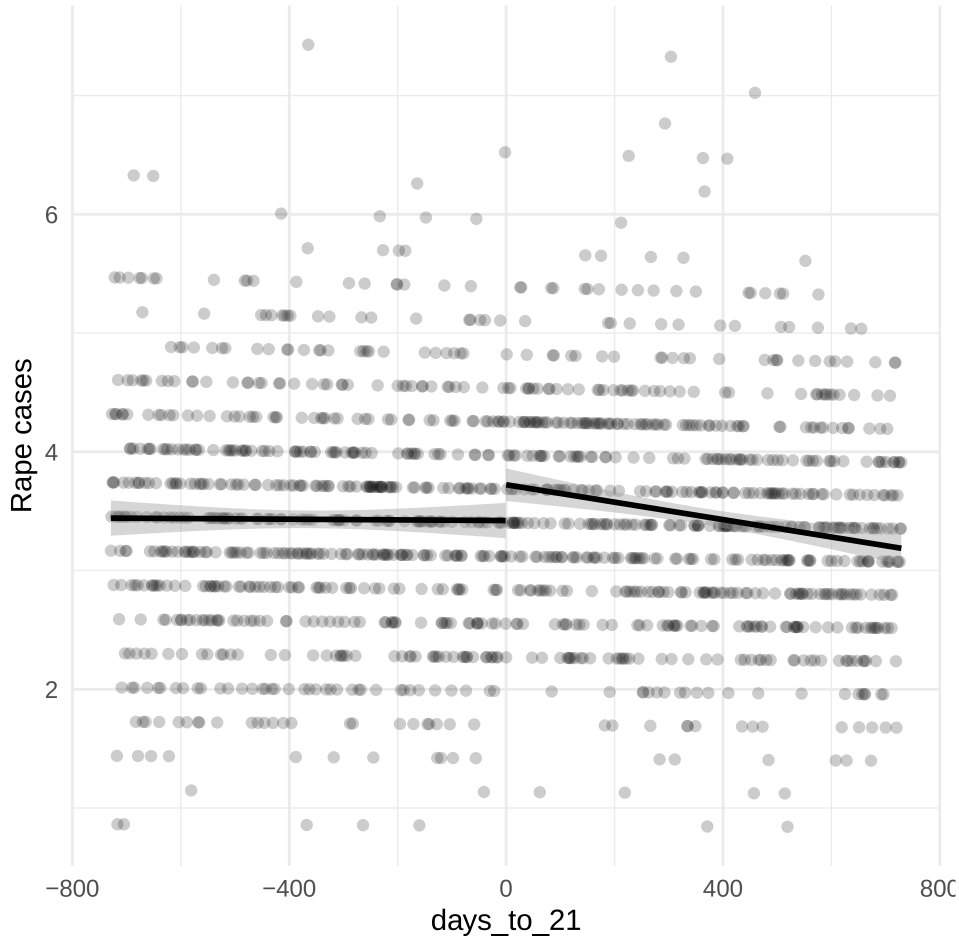


Figure 3. Effects of legal access to alcohol on rape crime arrests.

term	estimate	std.error	statistic	p.value
(Intercept)	3.29	0.0579	56.8	0
days_to_21	-0.000382	0.000123	-3.11	0.00194
mlda_and_over	0.302	0.104	2.92	0.00358

Table 2. Result of linear regression that estimates the effect of legal access to alcohol on rape crime arrests

Murder

In Figure x. there is no clear visual evidence exhibited. However, in the corresponding regression estimate shown in Table x, the increase in murder arrest rate at age 21 is statistically significant. There is a 0.269 (8.9%) increase in murder arrest rate at the threshold.

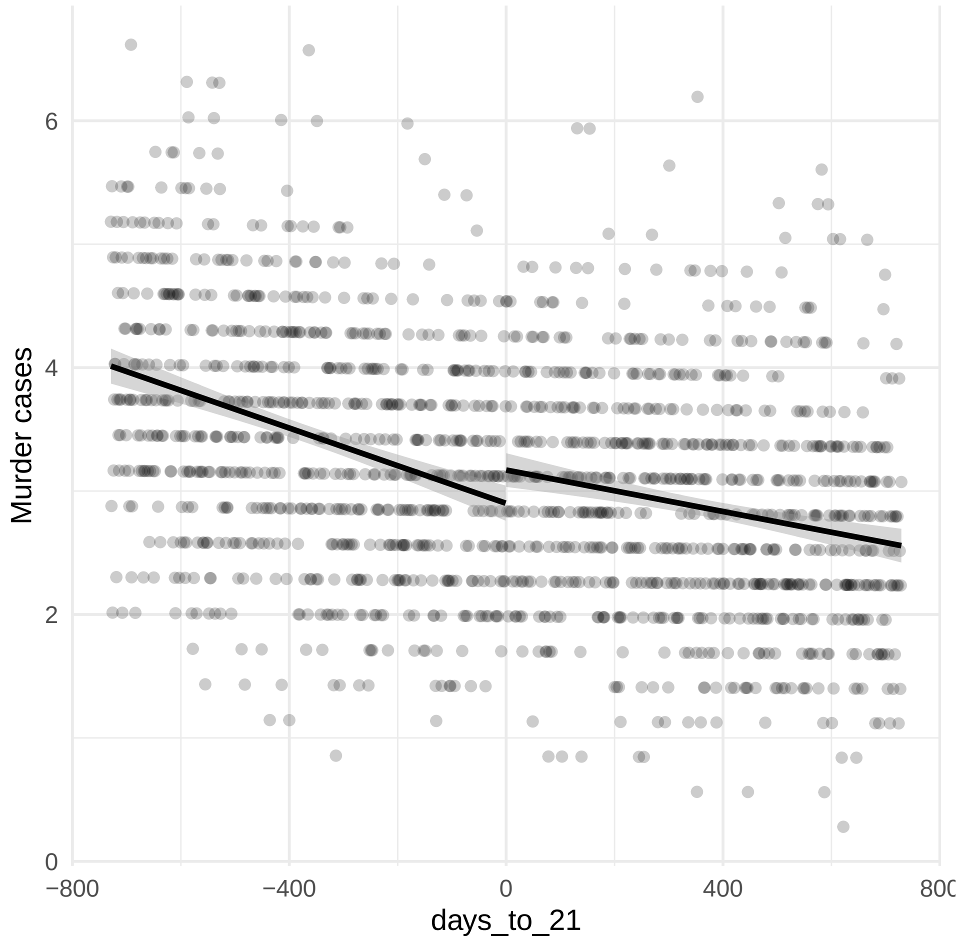


Figure 4. Effects of legal access to alcohol on murder crime arrests

term	estimate	std.error	statistic	p.value
(Intercept)	3.03	0.0564	53.7	0
days_to_21	-0.00118	0.00012	-9.88	2.56e-22
mlda_and_over	0.269	0.101	2.67	0.00763

Table 3. Result of linear regression that estimates the effect of legal access to alcohol on murdercrime arrests

Arson

Comparing to the other crimes, the increases in arrests for arson crime is substantially smaller. Arson crime arrest increases by 0.175 which is not a statistically significant increase. However, we hesitate to exclude the correlation because the number of arson arrest cases are very small, and the result might be limited by statistical precision in this analysis.

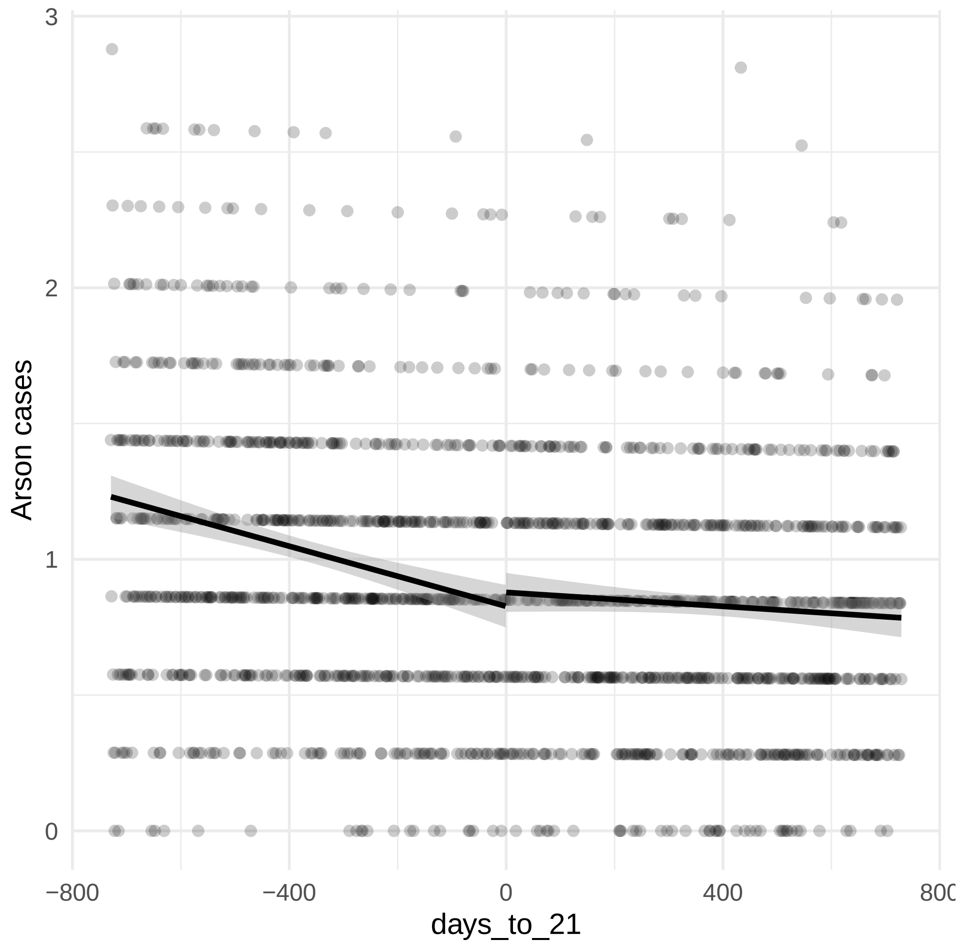


Figure 5. Effects of legal access to alcohol on arson crime arrests

term	estimate	std.error	statistic	p.value
(Intercept)	0.905	0.0304	29.8	8.29e-153
days_to_21	-0.00034	6.44e-05	-5.28	1.49e-07
mlda_and_over	0.0506	0.0543	0.933	0.351

Table 4. Result of linear regression that estimates the effect of legal access to alcohol on arson crime arrests

Appendix A

```
knitr::opts_chunk$set(echo = FALSE, include = FALSE)
# import libraries
library(broom)
library(tidyverse)
library(gridExtra)
library(pastecs)
library(readxl)
library(dplyr)
library(tidyr)
library(ggplot2)
library(reshape2)
library(stringr)
library(knitr)
library(huxtable)
library(jtools)

setwd("~/Experimental Design for Data Science/ProblemSet5")
drinking <- read.csv("P01 Age Profile of Arrest Rates 1979-2006.csv")
drinking <- janitor::clean_names(drinking)

head(drinking)

# Create a dataframe with the High-alcohol usage crimes
top4_crimes_19_to_23 <- filter(drinking, days_to_21 > -730 &
    days_to_21 < 730)
top4_crimes_19_to_23 <- select(top4_crimes_19_to_23, days_to_21,
    assault_r, rape_r, murder_r, arson_r)

# Create the dummy variable
top4_crimes_19_to_23 <- top4_crimes_19_to_23 %>% mutate(mlda_and_over = if_else(days_to_21 <
    0, 0, 1))


# Compute descriptive statistics - boxplots
p1 <- ggplot(top4_crimes_19_to_23) + aes(y = assault_r) + geom_boxplot() +
    labs(x = "", y = "Assault rate")
p2 <- ggplot(top4_crimes_19_to_23) + aes(y = rape_r) + geom_boxplot() +
    labs(x = "", y = "Rape rate")
p3 <- ggplot(top4_crimes_19_to_23) + aes(y = murder_r) + geom_boxplot() +
    labs(x = "", y = "Murder rate")
p4 <- ggplot(top4_crimes_19_to_23) + aes(y = arson_r) + geom_boxplot() +
    labs(x = "", y = "Arson rate")
grid1 <- grid.arrange(p1, p2, p3, p4, ncol = 2, nrow = 2)

# Select only top4
only_top4 <- top4_crimes_19_to_23

# Compute descriptive statistics - table
tt1 <- ttheme_default()

stats_table <- stat.desc(top4_crimes_19_to_23)
```

```r
stats_table <- round(stats_table, 2)
stats_table <- mutate(stats_table, stats = row.names(stats_table))
stats_table <- select(stats_table, stats, days_to_21, mlda_and_over,
    assault_r, rape_r, murder_r, arson_r)
colnames(stats_table) <- c("stats", "days_to_21 ", "mlda_and_over",
    "assault", "rape", "murder", "arson")
grid2 <- grid.arrange(tableGrob(stats_table, theme = tt1, rows = NULL),
    ncol = 1, nrow = 1)

# Join the boxplots and the table
grid.arrange(arrangeGrob(grid1, ncol = 1, nrow = 1), arrangeGrob(grid2,
    ncol = 1, nrow = 1), heights = c(15, 1), widths = c(1, 2),
    bottom = "Figure 1: Descriptive statistics of the data")

# Plot assault cases
top4_crimes_19_to_23 %>% ggplot(aes(x = days_to_21, y = assault_r)) +
    geom_point(alpha = 0.2) + geom_smooth(data = top4_crimes_19_to_23 %>%
    filter(days_to_21 < 0), method = "lm", color = "black") +
    geom_smooth(data = top4_crimes_19_to_23 %>% filter(days_to_21 >=
        0), method = "lm", color = "black") + theme_minimal() +
    labs(x = "days_to_21", y = "Assault cases")


# Regression discontinuity model for assault crimes
lm(assault_r ~ days_to_21 + mlda_and_over, data = top4_crimes_19_to_23) %>%
    tidy()


# Plot rape cases
top4_crimes_19_to_23 %>% ggplot(aes(x = days_to_21, y = rape_r)) +
    geom_point(alpha = 0.2) + geom_smooth(data = top4_crimes_19_to_23 %>%
    filter(days_to_21 < 0), method = "lm", color = "black") +
    geom_smooth(data = top4_crimes_19_to_23 %>% filter(days_to_21 >=
        0), method = "lm", color = "black") + theme_minimal() +
    labs(x = "days_to_21", y = "Rape cases")


# Regression discontinuity model for Rape crimes
lm(rape_r ~ days_to_21 + mlda_and_over, data = top4_crimes_19_to_23) %>%
    tidy()

# Plot murder cases
top4_crimes_19_to_23 %>% ggplot(aes(x = days_to_21, y = murder_r)) +
    geom_point(alpha = 0.2) + geom_smooth(data = top4_crimes_19_to_23 %>%
    filter(days_to_21 < 0), method = "lm", color = "black") +
    geom_smooth(data = top4_crimes_19_to_23 %>% filter(days_to_21 >=
        0), method = "lm", color = "black") + theme_minimal() +
    labs(x = "days_to_21", y = "Murder cases")

# Regression discontinuity model for Murder crimes
lm(murder_r ~ days_to_21 + mlda_and_over, data = top4_crimes_19_to_23) %>%
    tidy()
```

```r
# Plot arson cases
top4_crimes_19_to_23 %>% ggplot(aes(x = days_to_21, y = arson_r)) +
    geom_point(alpha = 0.2) + geom_smooth(data = top4_crimes_19_to_23 %>%
    filter(days_to_21 < 0), method = "lm", color = "black") +
    geom_smooth(data = top4_crimes_19_to_23 %>% filter(days_to_21 >=
        0), method = "lm", color = "black") + theme_minimal() +
    labs(x = "days_to_21", y = "Arson cases")


# Regression discontinuity model for Arson crimes
lm(arson_r ~ days_to_21 + mlda_and_over, data = top4_crimes_19_to_23) %>%
    tidy()
```

References

David Hugh-Jones (2020). huxable: Easily Create and Style Tables for LaTeX, HTML and Other Formats. R package version 4.7.1. https://cran.r-project.org/web/packages/huxtable

Hadley Wickham and Lionel Henry (2019). tidyr: Tidy Messy Data. R package version 1.0.0.

https://CRAN.R-project.org/package=tidyr

Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2018). dplyr: A Grammar of Data Manipulation. R package version

0.7.6. https://CRAN.R-project.org/package=dplyr

R Core Team (2019). R: A language and environment for statistical computing.

R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software,

4(43), 1686, https://doi.org/10.21105/joss.01686

Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

Yihui Xie (2019). knitr: A General-Purpose Package for Dynamic Report Generation in R. R

package version 1.25.

Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC.

ISBN 978-1498716963