Assignment 3

Answer Key

2023-12-12

Questions

1. Carpenter and Dobkin (2011) study the effect of the legal drinking age in the United States on public health outcomes. They use a Regression Discontinuity Design where the running variable is age, and the cutoff is the legal drinking age of 21. In the dataset `a3data1.csv`, the outcomes are *all* (all deaths), *alcohol* (alcohol-related deaths), and *mva* (motor vehicle accident deaths) measured in deaths per 100,000 people. The variable *agecell* is the average age for that set of people.

Plot separate scatterplots of each of the three outcomes against age. In each graph, make sure the axes are labelled appropriately, there is an informative title, and the cutoff is marked with a vertical line. Comment on the plots and explain what you see.

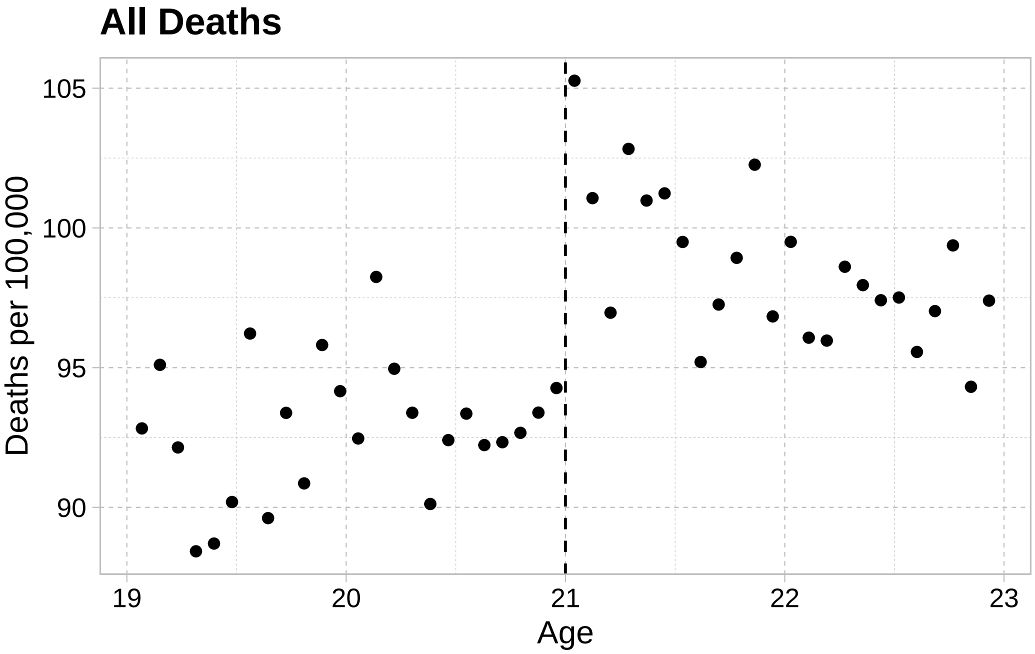
In the plot of all deaths, the relationship with age is upward sloping and there is a clear jump in the number of deaths at age 21. For alcohol-related deaths, there is also an upward sloping relationship, but the jump at age 21 is not clearly visible. Finally, for motor vehicle deaths, the relationship with age is negative, but there is a jump up at the age of 21. All of this suggests that the minimum drinking age has some effect on mortality.

```r
data <- read_csv("a3data1.csv") %>%
  filter(!is.na(all))
```

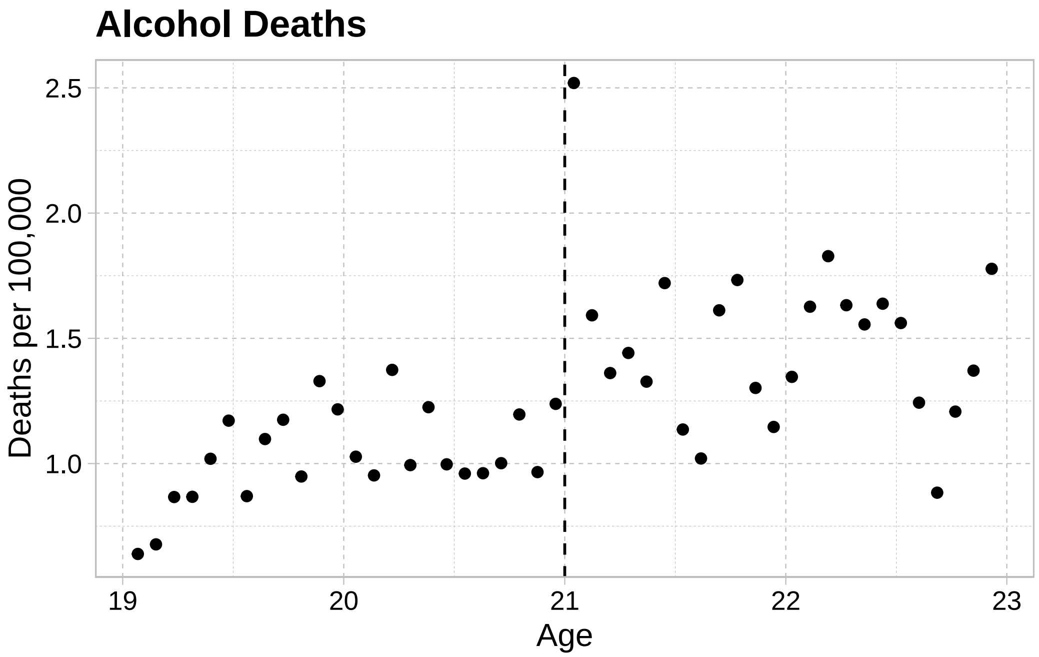
```
Rows: 50 Columns: 4
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (4): agecell, all, alcohol, mva

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

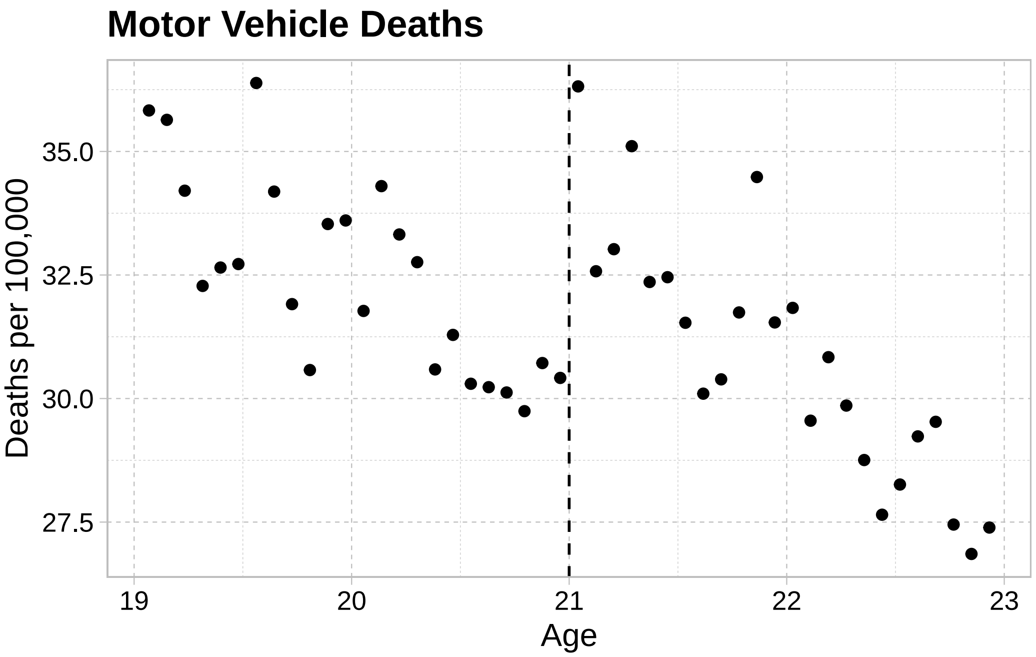
```
ggplot(data, aes(x = agecell, y = all)) +
  geom_point() +
  geom_vline(xintercept = 21, linetype = "dashed") +
  labs(x = "Age", y = "Deaths per 100,000", title = "All Deaths") +
  theme_pander(nomargin = FALSE, boxes = TRUE)
```



```
ggplot(data, aes(x = agecell, y = alcohol)) +
  geom_point() +
  geom_vline(xintercept = 21, linetype = "dashed") +
  labs(x = "Age", y = "Deaths per 100,000", title = "Alcohol Deaths") +
  theme_pander(nomargin = FALSE, boxes = TRUE)
```



```
ggplot(data, aes(x = agecell, y = mva)) +
  geom_point() +
  geom_vline(xintercept = 21, linetype = "dashed") +
  labs(x = "Age", y = "Deaths per 100,000", title = "Motor Vehicle Deaths") +
  theme_pander(nomargin = FALSE, boxes = TRUE)
```



2. Create a treatment variable called *mlda* that equals 1 when age is greater than or equal to 21, and zero otherwise. Using the outcome *all*, estimate the following regressions: a) a simple OLS regression of the outcome on mlda, b) regression discontinuity model that is linear in age with the same slope on both sides of the cutoff, c) regression discontinuity model that is linear in age with different slopes on each side of the cutoff, d) regression discontinuity model that is quadratic in age with different slopes on each side of the cutoff. Report the results in a professional-looking output table using the `modelsummary` package. Comment on the results.

In the most basic model, it appears that across the cutoff deaths increase by about 5 per hundred-thousand when people turn 21. This is consistent as we make the model more flexible, though the size of the coefficient increases gradually to about 9.5 deaths per hundred-thousand.

```
data %<>% mutate(mlda = ifelse(agecell >= 21, 1, 0)) %>%
  mutate(magecell = agecell - 21)

model1 <- feols(all ~ mlda,
                data = data, vcov = "HC1")
model2 <- feols(all ~ magecell + mlda,
                data = data, vcov = "HC1")
model3 <- feols(all ~ magecell + mlda + magecell*mlda,
                data = data, vcov = "HC1")
```

	Model A	Model B	Model C	Model D
(Intercept)	92.803***	91.841***	93.618***	93.073***
	(0.487)	(0.709)	(0.628)	(0.780)
mlda	5.740***	7.663***	7.663***	9.548***
	(0.730)	(1.514)	(1.273)	(1.830)
magecell		−0.975	0.827	−0.831
		(0.664)	(0.721)	(2.850)
magecell × mlda			−3.603**	−6.017
			(1.124)	(4.528)
				−0.840
				(1.541)
mlda × I(magecell^2)				2.904
				(2.257)
Num.Obs.	48	48	48	48
R2	0.573	0.595	0.668	0.682
RMSE	2.48	2.41	2.19	2.14

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

```
model4 <- feols(all ~ magecell + mlda + magecell*mlda + I(magecell^2) +
                  I(magecell^2)*mlda, data = data, vcov = "HC1")

modelsummary(list("Model A" = model1, "Model B" = model2, "Model C" = model3,
                  "Model D" = model4),
             gof_omit = "IC|Log|Adj|p\\.value|statistic|F|Std",
             stars = TRUE)
```

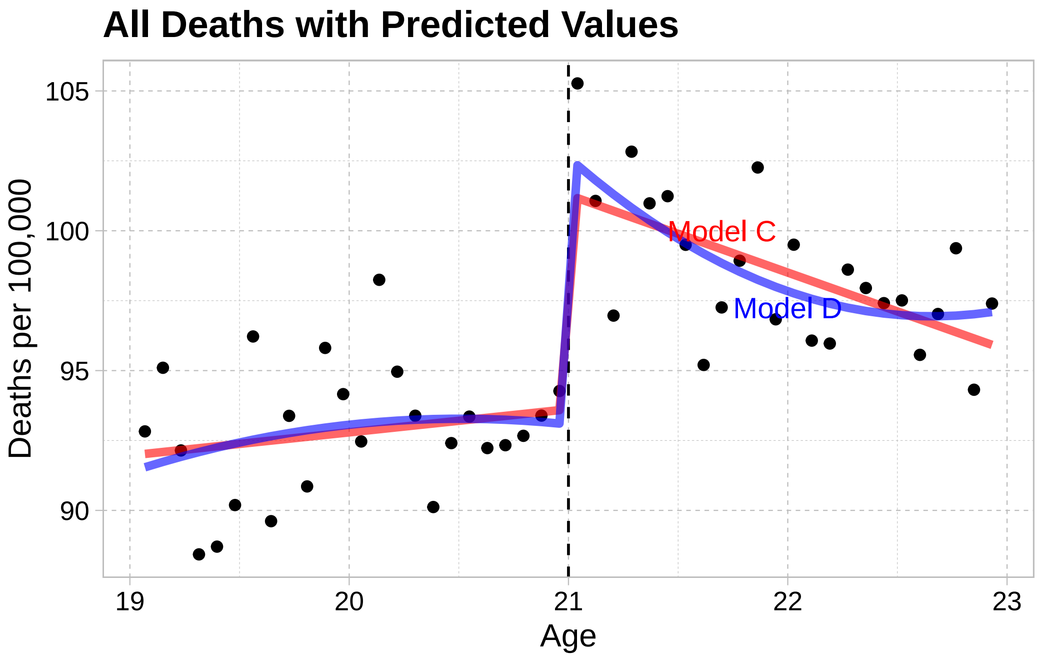
3. Plot the predicted values from (c) and (d) on top of a scatterplot of the data for the *all* outcome. Make sure the axes are labelled appropriately, there is an informative title, and the cutoff is marked with a vertical line. Explain which model fits the data better and why.

Based on an eyeball comparison, both seem to fit the data fairly well given how widely spread out the data are, and in reality it doesn't matter because they produce similar estimates of the coefficient.

```
# add legend labels

ggplot(data, aes(x = agecell, y = all)) +
  geom_point() +
  geom_vline(xintercept = 21, linetype = "dashed") +
  labs(x = "Age", y = "Deaths per 100,000", title = "All Deaths with Predicted Values") +
  theme_pander(nomargin = FALSE, boxes = TRUE) +
  geom_line(aes(x = agecell, y = predict(model3)),
            color = "red", size = 1.5, alpha = 0.6) +
  geom_line(aes(x = agecell, y = predict(model4)),
            color = "blue", size = 1.5, alpha = 0.6) +
  annotate("text", x = 21.7, y = 100, label = "Model C", color = "red") +
  annotate("text", x = 22, y = 97.25, label = "Model D", color = "blue")
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```



4. The Regression Discontinuity method requires that the untreated potential outcome is continuous across the cutoff. Explain what this means in the context of the Carpenter and Dobkin (2011) study. Do you think this assumption is reasonable? Why or why not?

The treatment here is being legally allowed to drink alcohol, which occurs when people are 21 years or over. The untreated potential outcome is the number of deaths that would occur if people were not allowed to drink alcohol, and we have to assume this would increase continuously if people were not allowed to drink when they turn 21. I think this is a very reasonable assumption to make, in particular because there are no obvious other things that change when people turn 21 that would affect the number of deaths.

5. We will now examine the effects of the minimum legal drinking age on deaths using a Difference in Differences framework. In 1975, the state of Alabama changed its legal drinking age from 21 to 19, while many other states kept their legal drinking age at 21. The data `a3data2.csv` contains information for Alabama and several other states over the years 1970 to 1985, and we would like to use this to estimate the effect of the change in the legal drinking age on deaths, as measured by mortality rates.

Plot the time series of mortality rates for all states in the data on the same graph. Make sure the axes are labelled appropriately, there is an informative title, and the time period where the policy change occurs is marked with a vertical line. Comment on the general patterns and trends in the data, and whether you think there is any obvious effect of the Alabama policy.

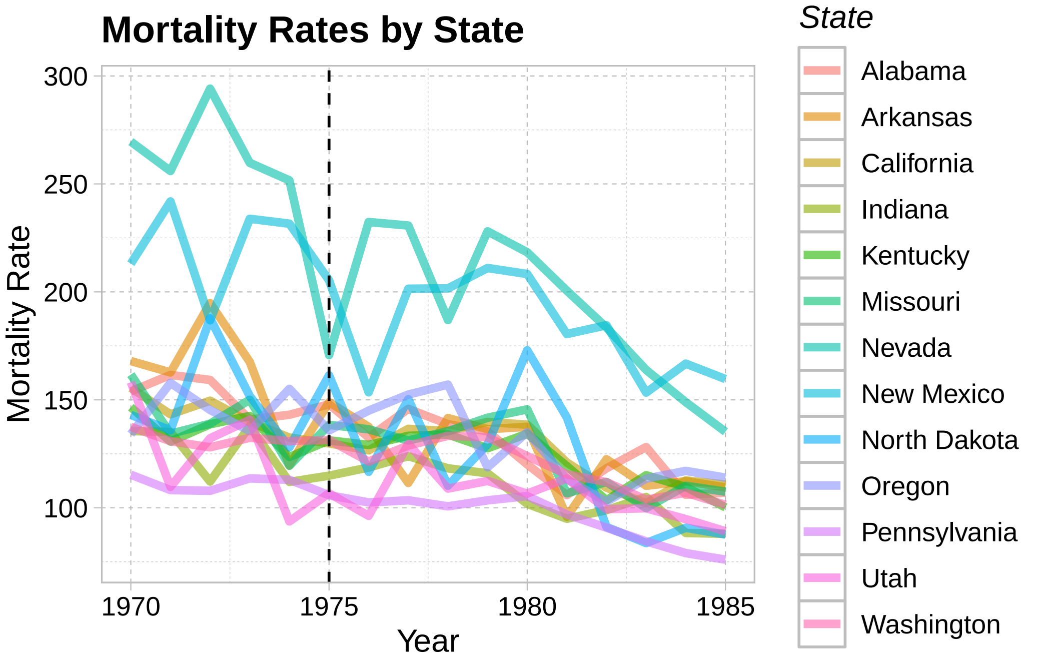
Mortality rates are clearly on the decline across states, and there is no obvious effect of the Alabama policy change. The good news is that states appear to have similar trends prior to the policy change, so there is a good chance that difference in differences would work.

```r
data2 <- read_csv("a3data2.csv")
```

```
Rows: 208 Columns: 4
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (1): statename
dbl (3): year, state, mrate

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
ggplot(data2, aes(x = year, y = mrate, color = as.factor(statename))) +
  geom_line(size = 1.5, alpha = 0.6) +
  geom_vline(xintercept = 1975, linetype = "dashed") +
  labs(x = "Year", y = "Mortality Rate", title = "Mortality Rates by State") +
  theme_pander(nomargin = FALSE, boxes = TRUE) +
  scale_color_discrete(name = "State")
```



6. Estimate the following regressions: a) a simple OLS regression of the mortality rate on a dummy for Alabama, b) a difference in differences specification with a dummy for Alabama, a dummy for the time period after treatment, and the interaction; c) a difference in differences specification a full set of state dummies, a full set of time dummies, and the same interaction from (b). Report the results in a professional-looking output table using the `modelsummary` package. Comment on the results.

In model A if we simply estimate the treatment effect in the absense of a difference in differences specification, there is some downward bias in the results. Once we use the other states as a control group, the effect is roughly 0.93, which means that deaths per million people increase by about 1 person. The effect is not statistically different from zero.

```
data2 %<>% mutate(after = ifelse(year >= 1975, 1, 0),
                  alabama = ifelse(statename == "Alabama", 1, 0))

model5 <- feols(mrate ~ alabama, data = data2, vcov = "HC1")
model6 <- feols(mrate ~ alabama*after,
                data = data2, vcov = "HC1")
model7 <- feols(mrate ~ alabama:after | statename + year,
                data = data2, vcov = "HC1")

modelsummary(list("Model A" = model5, "Model B" = model6, "Model C" = model7),
             gof_omit = "IC|Log|Adj|p\\.value|statistic|F|Std",
```

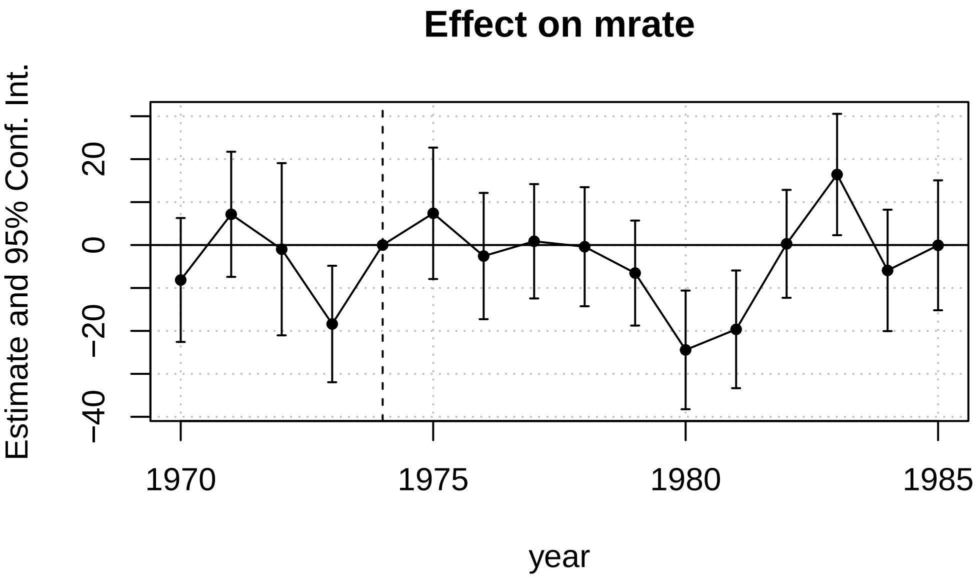
	Model A	Model B	Model C
(Intercept)	136.969***	155.289***	
	(2.862)	(5.876)	
alabama	−2.928	−3.569	
	(5.303)	(6.982)	
after		−26.648***	
		(6.572)	
alabama × after		0.931	0.931
		(8.826)	(5.737)
Num.Obs.	208	208	208
R2	0.000	0.104	0.858
R2 Within			0.000
RMSE	38.24	36.20	14.41

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

```
             stars = TRUE)
```

7. Estimate an "event study" regression where you include a full set of state dummies, a full set of time dummies, and an interaction between the Alabama dummy and all of the time period dummies except the one for 1974. Plot the results using `iplot`. Comment on whether the requirements for Difference in Differences to estimate a causal effect are met.

```
model8 <- feols(mrate ~ i(year, alabama, "1974") | statename + year,
                data = data2, vcov = "HC1")
iplot(model8, pt.join = TRUE)
```



It looks like there is a big dip in the mortality rate in Alabama relative to other states just before the treatment in 1973, which could cause issues with the common trends assumption.