ILLINOIS
Gies College of Business

Teenage Driving and Mortality

Authors: David Molitor and Julian Reif
Released: July 2022
Prepared for: FIN 550 Big Data Analytics

1 DANGERS OF DRIVING

Motor vehicle accidents killed over 40,000 Americans in 2020 and are the leading cause of death for adults under the age of 25 (CDC 2020). Adolescent drivers who lack driving experience face especially high crash risk. All state governments heavily regulate teenage driving to protect the safety of young drivers and those around them. These regulations include a minimum legal driving age, zero-tolerance drunk driving laws, and driver's education requirements.

While there is little doubt that driving increases mortality risk, estimating the magnitude of this causal effect is not straightforward. Comparing mortality rates of licensed drivers to non-licensed drivers is unlikely to yield causal estimates because a teenager's decision to obtain a license is voluntary and probably correlated with other behaviors such as illegal drug consumption that also affect mortality risk.

Alternatively, one could investigate what happens to mortality rates after changes to driving regulations. For example, a researcher could compare mortality rate trends in states that enacted a new law—such as prohibiting nighttime driving for young teenagers—to trends in other states that did not change their driving laws. This "difference-in-differences" comparison assumes that these other states serve as a reliable counterfactual for what would have happened in the states that enacted new laws. Is this assumption plausible? It is difficult to say because the legislative process is complicated and influenced by a number of different factors. If states that enacted stricter driving laws have unique teenage mortality patterns—for example, if they enacted laws in response to a public outcry over a growing number of motor vehicle fatalities—then this approach will yield biased estimates. This case study therefore considers a different empirical approach.

2 REGRESSION DISCONTINUITY

Regression Discontinuity (RD) is a research design that can overcome concerns about selection bias and estimate causal effects in certain non-experimental settings where an arbitrary "cutoff" rule causes individuals on either side of the cutoff threshold to be treated differently. For example, a school might only admit students who score above a certain threshold value on an admission test, or an individual might be eligible for free health insurance only if their income is below a certain amount. If individuals are unable to precisely manipulate on which side of the cutoff they lie, then individuals close to either side of the cutoff will receive different treatment—such as admission to a school—but otherwise be similar on average, producing a natural experiment. Under these conditions, comparing outcomes for individuals just above the cutoff to those just below the cutoff will produce causal estimates of the treatment effect for individuals near the threshold.

In practice, estimating a regression discontinuity requires the analyst to make two important choices. The first is a **sample choice** about which observations are near enough to the cutoff to draw valid conclusions about how outcomes change at the cutoff. The second is a **curve choice** for modeling how outcomes change for individuals who are nearer to or farther from the cutoff.

Terminology regarding sample and curve choices

- Sample choice. When asked to use a certain **bandwidth** B, keep observations within a distance B of the cutoff. When asked to estimate a **"donut" RD**, drop the observation at the cutoff.

- Curve choice. For this assignment, interpret the term "**non-parametric**" to mean that you should just calculate a difference in raw means on either side of the cutoff. Interpret "**parametric**" to mean that you should allow for linear trends on either side of the cutoff.

3 THE TEENAGE DRIVING STUDY

Teenagers cannot apply for a driver's license until they meet their state's minimum legal driving age. These laws thus create a large difference in the number of teenage drivers on either side of the minimum legal driving age threshold. Huh and Reif (2021) use an RD research design to study how mortality rates of teenagers just above this threshold compare to mortality rates of those just below the threshold. If the only difference between these two groups is their driving eligibility, then observed differences in their mortality rates can be attributed to the effect of driving.

The analysis in Huh and Reif (2021) is based on death certificate record data that include information on decedents' month and year of death, state of residence, cause of death, race, and sex. The authors use these data to compare death rates across teenage groups that differ in age by as little as one month. Because teenage mortality rates are very low (less than 1 per 1000 per year), obtaining precise estimates requires a large sample size. The authors were able to obtain death certificate records for every US death that occurred between 1983 and 2014, during which time there were 501,193 teenage deaths.

The authors also collected information on state minimum legal driving ages for the 1983–2014 sample period. The most common minimum legal driving age is 16 years, but this threshold varies across states and sometimes changes following the passage of a new driving law. During the time period studied by the authors, the lowest minimum legal driving age is 14 years and the highest is 18 years. The authors matched each decedent in their dataset to the law that was in force in the decedent's state of residence at the time of death, which allows them to calculate how many months away from the minimum legal driving age each decedent was at the time of death.

Here are a few resources that offer additional background on the study:

- Academic publication: https://julianreif.com/research/reif.aeri.2021.driving.pdf
- GitHub repository containing public use data: https://github.com/reifjulian/driving
    - Examples of how to load the data in R are included on the homepage README

4 REFERENCES

CDC (2020). Wisqars (web-based injury statistics query and reporting system).

Huh, Jason, and Julian Reif (2021). "Teenage Driving, Mortality, and Risky Behaviors." *American Economic Review: Insights,* 3(4), 523-39.

5 CASE DESCRIPTION

The public use data file "/data/mortality/derived/all.dta" in the GitHub repository for Hu and Reif (2021) contains data derived from National Vital Statistics mortality records and minimum legal driving age (MLDA) laws over the period 1983–2014. The data cover people who are within +/- 4 years of the MLDA and have been collapsed into bins, each corresponding to an age (in months) since MLDA, as given by the variable `agemo_mda`. Negative values of `agemo_mda` correspond to people too young to drive, and positive values correspond to those who are past the MLDA. The value `agemo_mda==0` corresponds to people in the calendar month they become eligible to drive, so on average, people observed in this month are eligible to drive for half the month (i.e., partially treated that month).

The other key variables in "all.dta" are `pop`, the total number of people observed in each bin, and the variables beginning with prefix "cod", which describe the total number of people in that bin who die from a given cause of death.

Note the following

- Use the "all.dta" data to answer the following questions. Load the data into R using `haven::read_dta()`.
- Calculate mortality rates in units of deaths per 100,000 person-years. For example, the rate of death per 100,000 person-years from any cause can be calculated as `cod_any = 100000 * cod_any / (pop / 12)`.

Answer the following questions

1. Calculate mortality rates due to any cause for individuals in the sample who are 1–24 months above the MLDA and for those who are 1–24 months below the MLDA. Does this difference between these two groups plausibly describe the causal effect of reaching the MLDA on mortality? Why or why not?

2. Create a scatter plot showing mortality rates due to (a) any cause and (b) motor vehicle accidents. Use black squares as markers for any cause of death and blue circles as markers for mortality due to motor vehicle accidents. Limit the plot to people who are within 2 years of the MLDA. Add a vertical line at the age at which driving eligibility begins.

3. Non-parametric "donut" RD. Calculate a non-parametric RD estimated effect of driving on mortality rates due to (a) any cause and (b) motor vehicle accidents. Calculate these estimates using four different bandwidths: 48, 24, 12, and 6 months. Omit the partially-treated observation `agemo_mda==0` from the estimation to generate what is called a "donut" RD. Use linear regression to calculate all these values, and report and describe this equation in your answer below. Report the results in a three-column table with 4 rows (one row per bandwidth). Column (1) should report the bandwidth, column (2) the RD estimate for all-cause mortality, and column (3) the RD estimate for motor vehicle accident mortality. Discuss whether/why point estimates and their precision change as the bandwidth becomes smaller.

4. Parametric "donut" RD. Calculate a parametric RD estimated effect of driving on mortality rates due to (a) any cause and (b) motor vehicle accidents. Allow for linear trends on either side of the cutoff. Calculate these estimates using four different bandwidths: 48, 24, 12, and 6 months. Omit the partially-treated observation `agemo_mda==0` from the estimation to perform a "donut" RD. Use

linear regression to calculate all these values, and report and describe this equation in your answer below. Report the results in a three-column table with 4 rows (one row per bandwidth). Column (1) should report the bandwidth, column (2) the RD estimate for all-cause mortality, and column (3) the RD estimate for motor vehicle accident mortality. Discuss whether/why point estimates and their precision change as the bandwidth becomes smaller. How do these parametric estimates compare to the non-parametric RD estimates?

6 CASE DELIVERABLES

You should **deliver the following two files** as part of your final project:

1. **Executive Summary**. Produce an executive summary report of your approach and findings in PDF format. The document should have three sections: (I) a concise overview of the case and objectives; (II) answers to each questions asked in the Case Description above; and (III) a conclusion that draws lessons for policymakers and stakeholders. Your executive summary needs to be thoughtful, but it should not be more than 3 double-spaced pages, font 12, in length.

2. **Script**. Along with your executive summary, turn in a script that contains programming code that performs your analysis. It should be created such that if you were to give your script to someone else, they could run the script and generate exactly the same results. To help us to understand your code, annotate your code with brief comments and follow recommended programming style practices.

7 CASE GRADING

1. **Executive Summary (60%)**. Your grade on this component will be based primarily on your ability to clearly communicate your objectives, methodologies, and results. Avoiding jargon, correct grammar, and proper sentence and paragraph structure will all be considered.

2. **Script (40%)**. Your grade on this component will be based on two factors. First, how easy is it to follow your script? Clearly commenting your code and using informative naming conventions will help. Second, are your results replicable? If the raw data files are in the same folder as your script, we should be able to run your code without any modification and exactly replicate all of your results.