FIN 550: Big Data Analytics Problem Set #2

Select whether this is an individual or group submission. No more than 3 members per group. Beyond the fact that all group members may submit the same answers, each submission must be separate work.

□ Group Submission. List group member names: Aishwary Joshi, Khushal Sharma

Problem set deliverables

You should submit the following three files as part of your problem set solution:

1. A completed version of this file, containing group member names and solutions to Problem 1.
2. A file named "Case-Executive-Summary.pdf" with the executive summary report for Problem 2.
3. An R script named "Case-Code.R" for Problem 2.

1. Difference-in-Differences (30 points)

The primary intent of medical malpractice law is to protect patients against professional negligence by a health care provider, which results in injury or death to the patient. However, proponents of medical liability reform argue that in fact these laws limit patient access to health care by driving doctors out of business or encouraging doctors not to use high-risk but potentially beneficial procedures. On June 11, 2003, Texas Governor Perry signed House Bill 4, a medical liability reform that greatly limited the amount of damages for which a physician could be held liable. You are provided the data table below, which indicates the number of doctors per 100,000 patients for Texas as well as for states neighboring the Lone Star State (nickname for TX).

State	Year	Doctors
Texas	1998	152
Texas	2002	158
Texas	2006	175
Neighbors	1998	196
Neighbors	2002	189
Neighbors	2006	180

Approach:

Data Provided:

Texas:

1998: 152 physicians per 100,000 residents
2002: 158 physicians per 100,000 residents (pre-policy)
2006: 175 physicians per 100,000 residents (post-policy)
Neighboring States (Composite):

1998: 196 physicians per 100,000 residents
2002: 189 physicians per 100,000 residents (pre-policy)
2006: 180 physicians per 100,000 residents (post-policy)
Policy Change:
On June 11, 2003, Texas enacted House Bill 4, which placed a cap on medical malpractice damages. The pre-policy period is represented by the 2002 data, while the post-policy period is represented by the 2006 data.

1. (16 points) Propose an estimate of the impact of Bill 4 on the number of doctors (per 100,000 patients) using only data for Texas.
   a. Provide a brief description of your method.

      To estimate the effect of the policy using only Texas's pre- and post-policy data, a straightforward approach is to conduct a before-and-after comparison:

      Determine the pre-policy metric (2002: 158 physicians per 100,000 population).
      Determine the post-policy metric (2006: 175 physicians per 100,000 population).

Compute the difference between the post-policy and pre-policy values (175 - 158 = 17).
This approach provides a single-difference estimate of the policy's impact, under the assumption that no other factors influenced the outcome during the period.

b. Does this estimate suggest Bill 4 increased or decreased the number of practicing physicians? By how much?

This estimation indicates that, relative to the pre-policy period, the number of physicians per 100,000 residents in Texas rose by 17 following the reform. Therefore, it seems that the policy led to an increase of 17 physicians per 100,000 individuals.

c. Discuss the key assumption required for your estimate to be valid (i.e. no bias).

The main assumption is that, without the policy, the trend in the number of physicians in Texas would have continued unchanged as it was before the reform. In other words, no other factors influenced the physician count during the period between 2002 and 2006. Put differently, the assumption implies that the policy was the sole systematic change affecting the number of doctors from the pre- to post-policy period.

d. Discuss a scenario under which this assumption would be violated.

A violation of this assumption could occur if another factor (e.g., a rise in medical school graduates, economic improvements in Texas, or unrelated healthcare policies/incentives) coincided with the reform and affected the number of physicians. For instance, if Texas implemented a loan forgiveness program for doctors in underserved areas beginning in 2003, the observed increase in physicians might be attributable to that initiative rather than the malpractice reform.

2. (16 points) Propose an estimate of the impact of Bill 4 on the number of doctors using only data for 2006.
a. Provide a brief description of your method.

This approach relies on a cross-sectional comparison focusing solely on the post-policy period. It compares the physician count in Texas in 2006 to that of its neighboring states during the same year. The underlying premise is that if the reform in Texas had a positive impact, the state should have a higher physician count relative to its neighbors after the policy was implemented.

Texas (2006): 175 physicians per 100,000 residents
Neighbors (2006): 180 physicians per 100,000 residents
Difference (Texas - Neighbors) = 175 - 180 = -5

b. Does this estimate suggest Bill 4 increased or decreased the number of practicing physicians? By how much?

This cross-sectional analysis indicates that Texas had 5 fewer physicians per 100,000 residents compared to its neighboring states in 2006. Based on this straightforward method, one could infer that Bill 4 may have resulted in a decrease of 5 physicians per 100,000 residents.

c. Discuss the key assumption required for your estimate to be valid (i.e. no bias).

The assumption underlying this method is that, apart from the policy, Texas and its neighboring states would have had the same average number of physicians in 2006. In other words, the neighboring states serve as a valid counterfactual, representing what would have occurred in Texas had the policy not been implemented. Additionally, there should be no unobserved factors systematically differing between Texas and its neighbors that could influence the supply of doctors.

d. Discuss a scenario under which this assumption would be violated.

If Texas and its neighbors differ in significant, unrelated ways—such as demographic variations, differing healthcare needs, or other state-specific healthcare policies—then the neighboring states may not serve as a valid control. For instance, if neighboring states experienced a simultaneous decline in physicians due to an unrelated policy, a post-policy comparison alone would fail to isolate the specific impact of Bill 4.

3. (16 points) Instead, construct a difference-in-differences estimate of the impact of Bill 4 on the number of doctors.

Approach:

The Difference-in-Differences (DiD) method compares the change in physician counts in Texas before and after the policy to the corresponding change in its neighboring states. This approach aims to account for shared trends that might affect both Texas and its neighbors.

Calculations:

Change in Texas (pre to post): 175 (2006) - 158 (2002) = +17
Change in Neighbors (pre to post): 180 (2006) - 189 (2002) = -9
Difference-in-Differences (DiD) Estimate:
= (Change in Texas) - (Change in Neighbors)
= 17 - (-9)
= 17 + 9
= +26

a. Does this estimate suggest Bill 4 increased or decreased the number of practicing physicians? By how much?

The Difference-in-Differences (DiD) estimate indicates that Bill 4 resulted in an increase of approximately 26 physicians per 100,000 residents. By comparing Texas's growth to the decline observed in neighboring states, this method attributes Texas's relatively stronger performance to the policy change.

b. Discuss the key assumption required for your estimate to be valid (i.e. no bias).

The primary assumption is that, in the absence of the policy, Texas and its neighboring states would have followed the same trend in physician counts over time. In other words, the difference in their outcomes would have remained stable without the intervention. This is known as the “parallel trends” assumption.

c. Discuss a scenario under which this assumption would be violated.

A violation of the “parallel trends” assumption would occur if Texas was already on a different trajectory due to factors unrelated to Bill 4. For example, if Texas had previously invested in medical education or implemented aggressive physician recruitment strategies before 2002, it could have been on a higher growth path independent of the reform. In such cases, if Texas and its neighbors were diverging for reasons other than Bill 4, the DiD estimate would be biased.

d. Set up a test to evaluate whether this key assumption appears to be plausible, using available data. Based on the results of this test, does the assumption appear to be valid?

To test the validity of the parallel trends assumption, pre-policy trends in the data can be examined. For instance, comparing the changes in physician counts from 1998 to 2002 in Texas and its neighboring states reveals the following:

Texas change (1998 to 2002): 158 - 152 = +6

Neighbors change (1998 to 2002): 189 - 196 = -7

If Texas and its neighbors exhibited similar trends before the policy (e.g., both increasing or decreasing at comparable rates), this would support the plausibility of the parallel trends assumption. However, in this case, the pre-policy trends diverge—Texas experienced an increase of 6 doctors per 100,000, while the neighbors experienced a decrease of 7 doctors per 100,000. This lack of parallel trends before the policy raises questions about the assumption's validity and casts doubt on the reliability of the DiD estimate.

If additional historical data from earlier years demonstrated more similar trends, confidence in the parallel trends assumption could increase. However, based on the limited data available, the differing pre-policy trends weaken this assumption and undermine the robustness of the DiD approach.

Summary of Estimates:

Texas-Only (Before-After): +17 doctors/100,000 (suggesting an increase)

Cross-Section Post-Only (Texas vs. Neighbors in 2006): -5 doctors/100,000 (suggesting a decrease)

Difference-in-Differences (DiD): +26 doctors/100,000 (suggesting a larger increase)

Among these, the DiD method is typically more robust as it accounts for shared trends between groups. However, its reliability depends heavily on the parallel trends assumption, which is called into question by the observed pre-policy differences.

2. Teenage Driving and Mortality (70 points)

Complete the data case, "Teenage Driving and Mortality." The case is available on Canvas. The case deliverables—an executive summary and R script—should be included with your problem set solutions.

Executive Summary

This analysis examines the causal impact of reaching the Minimum Legal Driving Age (MLDA) on mortality rates, utilizing mortality data spanning individuals within ±4 years of the MLDA. Key findings indicate a significant increase in mortality rates due to any cause and motor vehicle accidents immediately after reaching the MLDA. The average mortality rate for individuals 1–24 months above the MLDA is 64.34 deaths per 100,000 person-years, compared to 34.07 deaths for those 1–24 months below, suggesting an increase of 30.28 deaths. Scatter plots reveal that motor vehicle accidents are a primary contributor to this rise. Both nonparametric and parametric regression discontinuity (RD) designs confirm these findings, with RD estimates decreasing as bandwidth narrows, reflecting greater focus near the MLDA cutoff but reduced precision. Parametric RD estimates, accounting for linear trends, are generally smaller, suggesting nuanced effects beyond the immediate MLDA threshold. These results underscore the substantial public health risks associated with driving eligibility and highlight the potential for targeted policies to mitigate these risks.

Case 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?

Answer 1

The average mortality rates for individuals within 1–24 months above and below the MLDA were calculated as follows:

Above MLDA (1–24 months):
Average mortality rate = 64.34 deaths per 100,000 person-years

Below MLDA (1–24 months):
Average mortality rate = 34.07 deaths per 100,000 person-years

Difference:
Difference in mortality rate = 30.28 deaths per 100,000 person-years

This difference plausibly describes the causal effect of reaching the MLDA on mortality. The spike in mortality rates above the MLDA could be attributed to increased exposure to risks, such as motor vehicle accidents, as individuals gain driving eligibility. However, it is important to note that this is an observational comparison, and other confounding factors, such as differences in behavior, socioeconomic factors, or geographic differences, might influence mortality rates.

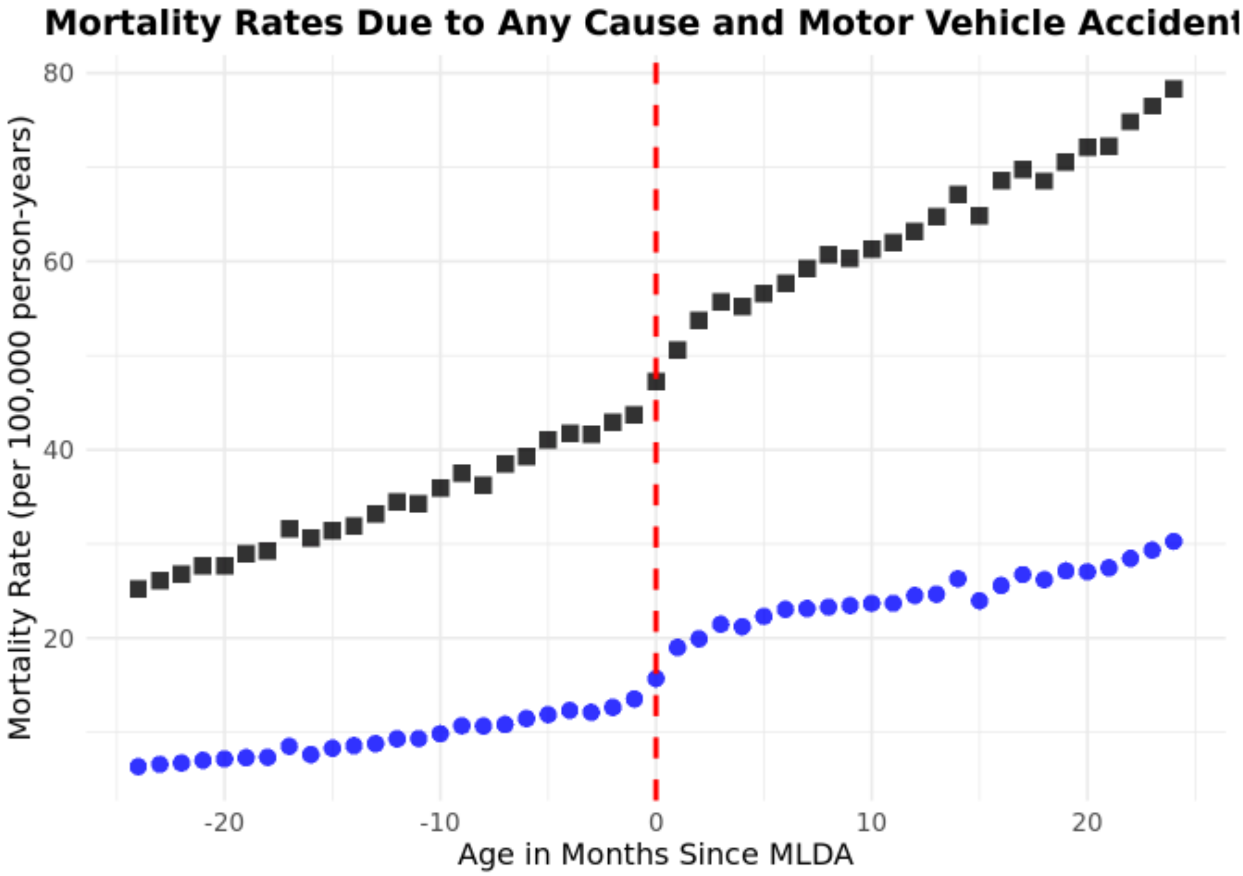
2. Create a scatter plot showing mortality rates due to (a) any cause and (b) motor vehicle accidents. Use black squares as markers for any cause of death and blue circles as markers for mortality due to motor

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.

<u>Answer 2</u>

A scatter plot was created to visualize mortality rates due to any cause and motor vehicle accidents for individuals within ±24 months of MLDA. The plot includes:

- Black squares representing mortality rates due to any cause.
- Blue circles representing mortality rates due to motor vehicle accidents. - A vertical dashed red line at agemo_mda = 0 (the MLDA cutoff).



The scatter plot shows a clear increase in mortality rates immediately after reaching the MLDA, particularly due to motor vehicle accidents. This highlights the direct risk associated with becoming eligible to drive.

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.

Answer 3

Non-parametric RD estimates were calculated by omitting the partially treated observation (agemo_mda == 0) and performing linear regressions for four bandwidths: 48, 24, 12, and 6 months. The regression equation used was:

$$\text{MortalityRate} = \beta 0 + \beta 1.I(\text{agemo\_mda}>0) + \epsilon$$

Here, β1 represents the RD estimate.

Bandwidth (months)	RD Estimate (All-Cause Mortality)	RD Estimate (Motor Vehicle Accidents)
48	48.84	21.45
24	30.28	15.29
12	19.07	11.18
6	13.17	8.84

As the bandwidth decreases, the RD estimates for both all-cause mortality and motor vehicle accident mortality become smaller. This is because smaller bandwidths focus more closely on observations near the MLDA cutoff, reducing noise but also decreasing the sample size, which can reduce the precision of the estimates.

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?

Answer 4

Parametric RD estimates were calculated by allowing for linear trends on either side of the cutoff using the following regression equation:

$$\text{MortalityRate}= \beta 0 + \beta 1.I(\text{agemo\_mda}>0) + \beta 2.\text{agemo\_mda} + \beta 3.(\text{agemo\_mda}.I(\text{agemo\_mda}>0)) + \epsilon$$

Here, β1 represents the RD estimate.

Bandwidth (months)	RD Estimate (All-Cause Mortality)	RD Estimate (Motor Vehicle Accidents)

48	11.89	9.53
24	6.88	6.55
12	6.61	5.97
6	6.01	4.87

As with the non-parametric RD, the estimates decrease as the bandwidth becomes smaller. The parametric RD estimates are generally lower than the non-parametric RD estimates, as the inclusion of linear trends accounts for gradual changes in mortality rates away from the cutoff, which may reduce the estimated effect. Precision improves as larger bandwidths are used, but this comes at the cost of potentially including observations farther from the cutoff that may not be comparable.

Comparison of Non-parametric and Parametric Estimates

- Non-parametric estimates are generally higher because they rely on fewer assumptions about the functional form of the relationship between age and mortality rates.
- Parametric estimates adjust for linear trends, resulting in lower RD estimates.
- Both methods show that mortality rates increase significantly above the MLDA, with motor vehicle accidents being a major contributor.

Conclusion

- **Causal Effect:** Both non-parametric and parametric RD estimates suggest a significant causal effect of reaching the MLDA on mortality, particularly due to motor vehicle accidents.
- **Precision vs. Bandwidth:** Smaller bandwidths improve the validity of causal inference but reduce precision due to smaller sample sizes.
- **Policy Implication:** The findings support the need for policies that mitigate risks associated with driving eligibility, such as graduated licensing or education programs.

R Code given in other file