Teenage Driving, Mortality, and Risky Behaviors*

Xing XiaoYuan Yuping Hao Meixuan Chen

26 February 2022

Abstract

This paper mainly focus on investigating effect of death rate from both car accidents and poisoning in teenagers group. After investigation, we found that teenagers aged 16 is a cutoff point for car accidents death rates, teenagers less than 16 years old has smaller death rates because of car accidents than teenagers older than 16 years old. Moreover, we found that male teenagers have much larger death rates than female teenagers. Furthermore, with respect to poisoing death rates, male teenagers have much larger death rates than female teenagers, but we don't find sigificantly upward poisoning death rates after 16 years old, just rate of poisoning death rate becomes larger for male teenagers older than 16 years old. As results, we are going to use two statistical models to investigate effect of death rates for car accidents and poisoning respective. The first model is regression discontinuity model, it is used to analyze ffect of death rates for car accidents; the second model is normal linear regression model, it is used to analyze ffect of death rates for poisoning. Finally, we conclude age and gender are both significant to impact death rates from both car accidents and poisoning.

1 Introduction

1.1 Goal and Paper Source

Goal of this assignment is making replication of the paper written by Jason Huh and Julian Reif, paper's name is called "Teenage Driving, Mortality, and Risky Behaviors" written by Huh and Reif (2021). Whole project is using R language R Core Team (2020) and RStudio Team (2020). The report is focusing on investigating effects of teenager mortality on dangerous driving in the United States. In the United States, teenagers' deaths are coming from different reasons, however, poisoning deaths and car accidents are two major death sources. As a result, we are going to investigate effect of teenagers mortality because of driving accidents and poisoning deaths.

1.2 Teenager mortality from car accidents and poisoning death

In this replication tasks, we are going to use same data sources and statistical method as original paper provided by authors. With respect to data, authors derived them from different sources like Insurance Institute for Highway Safety, National Vital Statistics, etc. However, we will use data sets cleaned by paper's authors for the rest of analysis, if anyone interested in original and cleaned dataset, please refer to *Teenage Driving, Mortality, and Risky Behaviors Data Set* (2021). With respect to statistical model, we also use linear regression with discontinuity, the regression discontinuity cutoff point is relying on teenagers' ages because teenagers should follow "minimum driving age" or "MDA" in law, teenagers in "minimum driving age" rarely drive cars so have lower car accident rate compare to teenagers not in "minimum driving age." In sample data set, the "minimum driving age" is 16 years old. On the other hand, use same statistic model to investigate mortality effect from death because of poisoning.

*Code and data are available at: https://github.com/NATHAN0472/PAPER2.git

1.3 Paper structure

Now, let's introduce the whole structure of this paper, in addition to current introduction section, the rest of paper will be divided into three main parts: the first part is about data analysis, like how teenager mortality's trends look like for different genders; the second part is what about statistical model used for this analysis; the third part is modeling and analysis results display; the last part is whole replication discussion, we will raise three discussion points and that acquire from "Teenage Driving, Mortality, and Risky Behaviors" during replications.

2 Data

Our data is from different sources, for example, minimum driving age laws from the Insurance Institute for Highway Safety for the years 1995–2014; mortality measurement using the National Vital Statistics; driving behaviors from the National Longitudinal Study of Adolescent to Adult Health. These data from different sources are assembled in here: *Teenage Driving, Mortality, and Risky Behaviors Data Set* (2021).

Before stepping into data analysis, let's clarify one thing about the most important (response) variable 'death rate,' where data set doesn't include it explicitly. 'death rate' here refers to proportion of death per 100,000 people per year, so 'death rate' is calculated as follows Wickham et al. (2019):

$$y = \frac{100000 \times \text{number of death per month}}{\text{total population}/12}$$

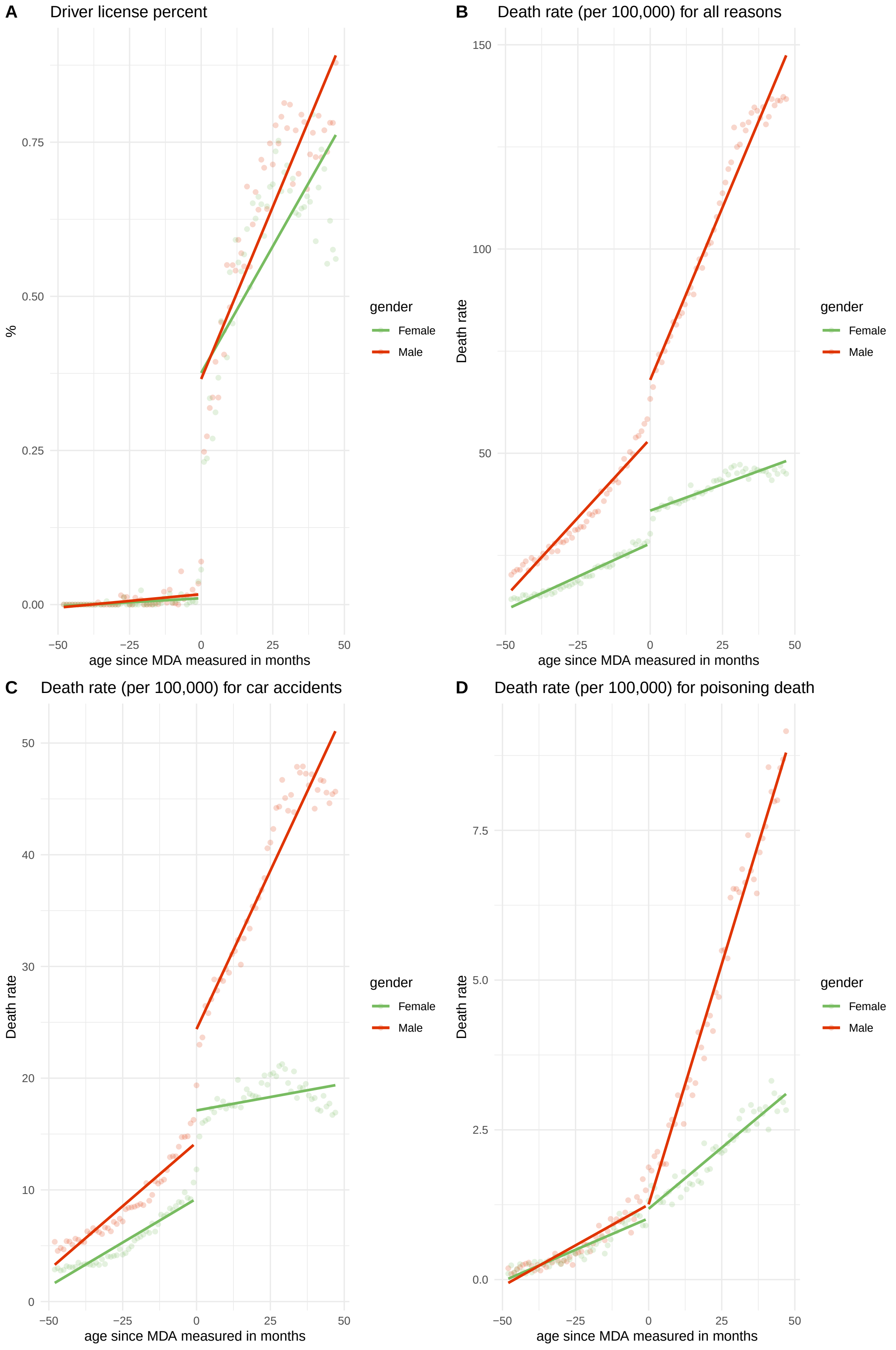
Where y is 'death rate,' 'number of deaths per months' and 'total population' are included in the data set, 12 is referring to number of months per year. 'death rate' from different reasons will be calculated same way.

In addition, another important (independent) variable is 'age since MDA,' as introduced before, the minimum driving age or MDA is 16 years old in the sample, so 'age since MDA measured in months' will be calculated as follows:

$$z = \text{teenager's age} - 16$$

Where z is 'age since MDA measured in months,' if it is greater than 0 then observed teenager's age greater than 16 years old, if it is less than 0 then observed teenager's age less than 16 years old.

Now, let's look at several data visualization results for teenagers mortality because of car accidents and poisoning death Wickham (2016).


A
Driver license percent
%
0.00
0.25
0.50
0.75
−50
−25
0
25
50
age since MDA measured in months
gender
Female
Male
B
Death rate (per 100,000) for all reasons
Death rate
50
100
150
−50
−25
0
25
50
age since MDA measured in months
gender
Female
Male
C
Death rate (per 100,000) for car accidents
Death rate
0
10
20
30
40
50
−50
−25
0
25
50
age since MDA measured in months
gender
Female
Male
D
Death rate (per 100,000) for poisoning death
Death rate
0.0
2.5
5.0
7.5
−50
−25
0
25
50
age since MDA measured in months
gender
Female
Male

These four plots Kassambara (2020) show something about the mortality and poisoning death data sets. Plot A shows relationship of age since minimum drive age and percentage of observed teenagers who have driver's license. There is an obvious cutoff point at age = 16 or MDA = 0, for teenagers younger than 16 years old, holding driver's license percentage is almost zero, but there is still an upward treading for MDA from -50 months to 0 months, and when MDA greater than zero, upward trend becomes more obvious that when teenagers become older and older, percent of having a driver's license becomes higher and higher for both female and male.

The plot B shows relationship of age since minimum drive age and death rates. It is very obvious that teenagers' death rates keep increasing as age since MDA increasing, but start from 16 years old, death rates are significantly higher than death rates among teenagers less than 16 year old. Also, in general, male's death rates are higher than female's death rates for all ages.

The plot C shows relationship of age since minimum drive age and death rates purely come from car accidents. Still, teenagers' death rates because of car accidents keep increasing as age since MDA increasing, but start from 16 years old, death rates because of car accidents are significantly higher than death rates among teenagers less than 16 year old.

The final plot D shows relationship of age since minimum drive age and death rates purely come from poisoning deaths. For female teenagers, the relationship between age and poisoning death is upward trending for all ages. For male teenagers, the relationship between age and poisoning death is still upward trending for all ages, but after 16 years old, slope of death rates increases dramatically. However, compare to the first three plots, plot D doesn't show death rates have significant change for teenagers aging 16.

3 Model

After doing some data analysis and data visualization, in this section, let's introduce statistical models to analyze mortality rate because of car accidents and poisoning deaths.

The first model is to investigate effects of death rates from car accidents. As we analyzed in previous section, cutoff point at 16 years old is very obvious, car accident death rates are dramatically increase for teenagers 16 years old or older. So the regression discontinuity model should consider a cutoff point at 16 years old. For more details about regression discontinuity model, please refer to the article written by Lee and Lemieux (2010).

$$Y_i = \beta_0 + \beta_1 MDA_i + \beta_2 cutoff_i + \beta_3 (MDA_i \times cutoff_i) + \beta_4 D_i + \beta_5 Gender_i$$

Where Y_i is response variable, stands for death rates from car accidents for observed teenager i, MDA_i stands for minimum driving age for observed teenager i, $cutoff_i$ stands for a indicator variable, if $MDA_i < 0$ then $cutoff_i = 0$, if $MDA_i \geq 0$ then $cutoff_i = 1$, $(MDA_i \times cutoff_i)$ is interaction effect between MDA and cutoff predictors, D_i is another indicator predictor, if $MDA_i = 0$ then $D_i = 1$ and $D_i = 0$ otherwise. Adding D_i can remove the bias of $cutoff_i$ when $MDA_i = 0$, because when $MDA_i = 0$, it is bias to set $cutoff_i = 1$, $Gender_i$ is gender of observed teenager i. β_0 to β_4 are coefficients for corresponding predictors.

Another model is regular linear regression model without discontinuity point at 16 years old, it will be used to analyze effect of poisoning death rates by different ages or MDA. This is the method that we chose differently from original article. The reason we use linear regression model without discontinuity is because from plot D, $MDA = 0$ is not really a cutoff point like plot A to plot C. As a result, model should be like follows:

$$Z_i = \beta_0 + \beta_1 MDA_i + \beta_5 Gender_i$$

Where Z_i is poisoning death rates for observation i, MDA_i stands for minimum driving age for observed teenager i same as before, $Gender_i$ is gender of observed teenager i like before. β_0 and β_1 are still estimated coefficients in the regression model.

Table 1: Regression Discontinuity Model Summary

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.0396	1.1230	5.3782	0.0000
agemo_mda	0.1927	0.0377	5.1144	0.0000
cutoff1	9.4696	1.5081	6.2792	0.0000
D1	-5.6100	3.7722	-1.4872	0.1387
genderMale	11.3888	0.7382	15.4268	0.0000
agemo_mda:cutoff1	0.1007	0.0542	1.8602	0.0644

Table 2: Regression without Discontinuity Model Summary

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.3558	0.1114	12.1727	0
agemo_mda	0.0631	0.0028	22.1872	0
genderMale	1.4815	0.1575	9.4062	0

In the next section, let's show some analysis results about death rates from these two regression model.

4 Results

In this section, we are going to show summary of both fitted regression discontinuity model for car accidents death rates and fitted normal regression model for poisoning death rates.

From table 1 about regression discontinuity summary, all predictors are significant except D, that is because D is an indicator predictor to remove bias when classifying MDA=0 into cutoff category '1.' So, we can say all predictors are significant to impact death rates from car accidents. In table 1, we printed out estimated coefficient for each predictor or interaction effect, now let's interpreting them. When female teenagers at 16 years old, expected death rates from car accidents are around 6%, for one year age increase, expected death rates from car accidents are increasing by 0.19% keep all other effects constant in the model, teenagers older than 16 years old have average 9.46% more death rates than teenagers younger than 16 years old keep all other effects constant in the model, male teenagers have average 11.38% more death rates than female teenagers keep all other effects constant in the model, when teenagers older than 16 years old, for one year age increase, expected death rates from car accidents are increasing by 0.1% keep all other effects constant in the model.

Table 2 shows estimated coefficient for regression without discontinuity Model, this model is much easier than regression discontinuity model, only two predictors are involved 'Age' and 'gender,' both of them are significant to impact poisoning death rate. Now, let's interpret model. When a teenager is 16 years old and she is a female, her expected death rate from poisoning is about 1.35%, for one year age increase, expected death rates from poisoning are increasing by 0.063% keep all other effects constant in the model, male teenagers have average 1.48% more death rates because of poisoning than female teenagers keep all other effects constant in the model.

5 Discussion

We are almost finishing the paper replication work for "Teenage Driving, Mortality, and Risky Behaviors," after do some data analysis and fit two regression models, we can have some idea about relation between teenagers' death rates and age or genders. In short, firstly, male's death rates are higher than female's death rates in general, secondly, 16 years old is a noticeable teenager age, teenagers after 16 years old have much higher death rates than teenagers before 16 years old. In this process, we learned something from this paper about how to investigate effect of death rates by using regression discontinuity model. However, there are still some points that we did differently compare to original paper, let's do three discussion points about this replication work.

5.1 First discussion point about regression discontinuity model.

This is the first time to know the regression discontinuity model, linear regression model is the most common model that we used before. In this regression discontinuity model, we manually split the analyzed data into several parts by creating a new indicator function. In this case, we are able to conclude more precise effect for death rates because of car accidents. Based on explanatory data analysis in Data section, it is obvious to see that car accidents death rates changed significantly from teenagers younger than 16 years old to teenagers older than 16 years old, That is because the U.S government don't allow teenagers under 16 years old drive alone, that law brings down the car accidents death rates significantly for teenagers under 16 years old.

In order to tell regression discontinuity model performs better than normal regression model for car accidents death rates, we decided to fit a normal regression model by using age and gender predictors and compare these two models by likelihood ratio test Zeileis and Hothorn (2002). Likelihood ratio test is a hypothesis testing with null and alternative hypothesis:

$$H_0 : \text{simple model is better than complex model}$$
$$H_a : \text{simple model is not better complex model}$$

Based on the context of analysis, null and alternative hypotheses are:

$$H_0 : \text{normal regression model is better than regression discontinuity model}$$
$$H_a : \text{normal regression model is not better than regression discontinuity model}$$

Likelihood Ratio Test Results 1.1				
Model 1: y_MVA ~agemo_mda + gender Model 2: y_MVA ~agemo_mda + cutoff+ agemo_mda * cutoff + D + gender				
DF	LogLik	DF	Chisq	P value
4	-603.00			
7	-582.76	3	40.498	**<8.355e-09**

From the likelihood ratio test result 1.1, p value of likelihood ratio test is closing to zero, that means there strong evidence to reject null hypothesis and conclude that regression discontinuity model performs better than normal regression model when analyzing car accidents death rates.

Among predictors in regression discontinuity model, age and gender are both important predictors that impact car accidents death rates. Also, another important thing is in addition creating an indicator predictor to identify whether teenagers younger or older than 16 years old, we also learned from original paper to create another indicator predictor to tell whether teenagers is exact 16 years old, because we classify an observed teenager who is exact 16 years old to teenager older than 16 years old group, adding another indicator predictor to tell whether teenagers is exact 16 years old can remove bias of that classification.

5.2 Second discussion point about teenager's death rates discussion

From both death rates, male teenagers is much higher than female teenagers. From fitted models, we can tell exact quantitative numbers, male teenagers have average 11.38% more car accidents death rates than female teenagers; male teenagers also have average 1.48% more death rates because of poisoning than female teenagers. Now, need to think about why males teenagers death rates higher than female teenagers.

According to journey “Boys Have Higher Death Rates From Many Causes, Study Shows” Norton (2013) written by Amy Norton, she also pointed out that male teenagers between age 16 to age 19 have higher death rates than female teenagers in many perspective like suicide, car accidents, homicide, etc. She explained the reason from genetics perspective, she pointed out that male teenagers have some inherent that develop serious disease, but female teenagers have less probability have that inherent. Once teenagers have this inherent, they may have higher probability to die in any accidents.

In addition to that, Jason Huh and Julian Reif Huh and Reif (2021) also pointed showed teenagers death rates become higher as they are eligible to drive a car.

So what we suggest are parents should be more attention to their children driving especially after they are eligible to drive for both male and female. From all analysis before, for both male and female teenagers, death rates from car accidents are highest after their 16 years old. Before 16 years old, poisoning death is more serious than car accident death, so parent should be more attention to their children health care before 16 years old.

5.3 Third discussion point about some improvments I did compare to original paper

The final discussion point we want to raise here is about things that we did differently compare to original paper, because for some points, we feel some improvements are needed for original paper.

The improvement suggestion is about in regression models to analyze death rates, Jason Huh and Julian Reif didn’t include gender as part of predictors. However, based on data analysis and model analysis in our paper and original paper, gender is a very important predictor that impact death rates because of both car accidents and poisoning, in both models, they suggested that male teenagers have higher death rates than female teenagers. Similarly, we still use likelihood ratio test to tell whether adding gender as part of predictors makes regression discontinuity better with following hypotheses:

H_0 : regression discontinuity model without gender is better than regression discontinuity model with gende

H_a : regression discontinuity model without gender is not better than regression discontinuity model with gende

<table>
<tr><th colspan="5">Likelihood Ratio Test Results 1.2</th></tr>
<tr><td colspan="5">Model 1: y_MVA ∼agemo_mda + cutoff + agemo_mda * cutoff + D
Model 2: y_MVA ∼agemo_mda + cutoff + agemo_mda * cutoff + D + gender</td></tr>
<tr><td>DF</td><td>LogLik</td><td>DF</td><td>Chisq</td><td>P value</td></tr>
<tr><td>6</td><td>-661.86</td><td></td><td></td><td></td></tr>
<tr><td>7</td><td>-582.76</td><td>1</td><td>158.2</td><td><b><2.2e-16</b></td></tr>
</table>

According to likelihood ratio test table 1.2, p value is closing to zero, that means we have strong evidence to reject null hypothesis and conclude that add gender to regression discontinuity model is better.

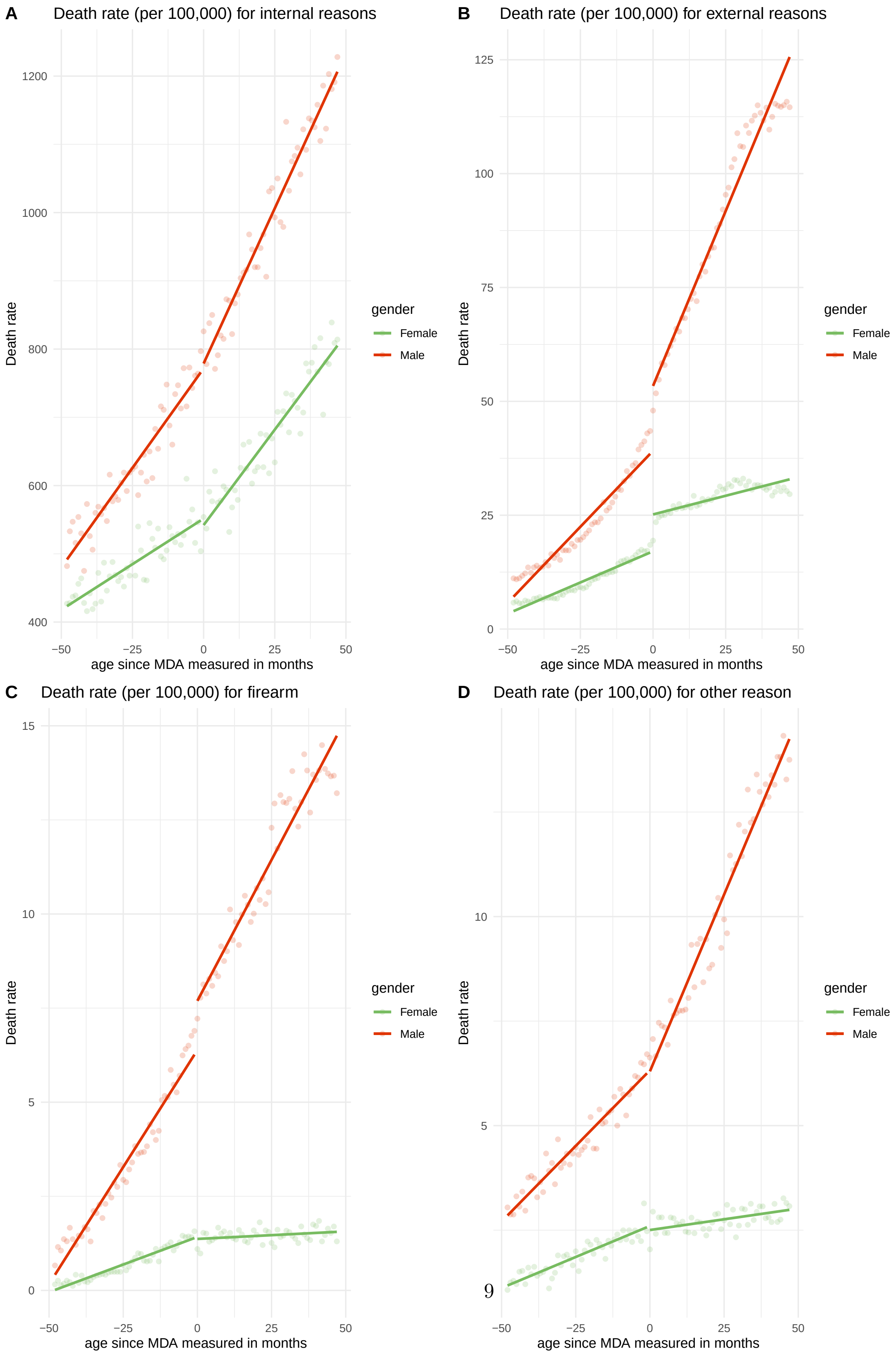
5.4 Weaknesses and next steps

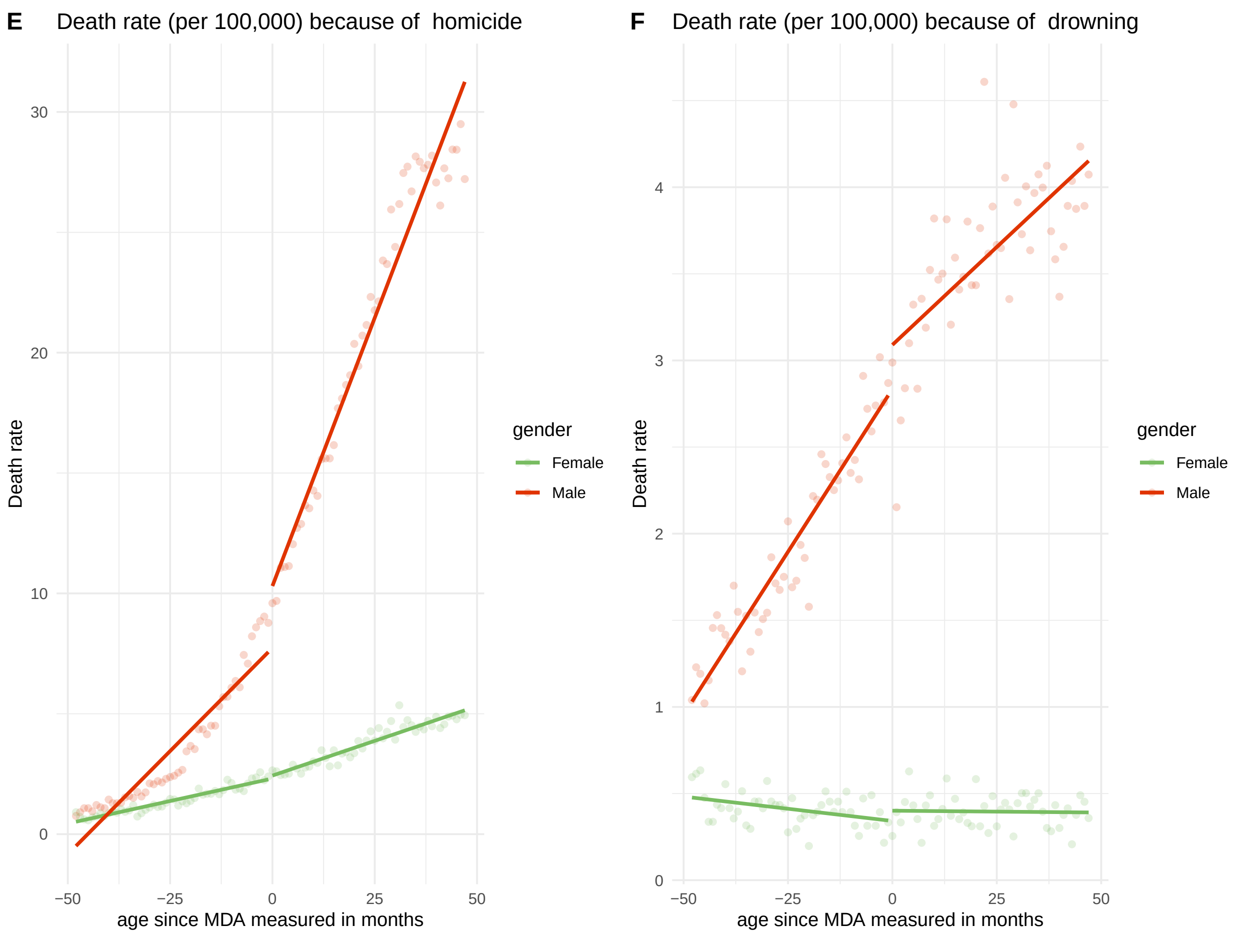
However, this paper still has some weaknesses. One is about data assumption check. When fitting a linear regression model, the model should satisfy few linear regression model assumptions, for example, linearity means response variable and predictors should have linear relationship, homodasticity means regression model residuals should be constant all the way, observations should be independent each other, if users want to make a inference for fitted regression models, then need to make sure residuals are following normal distribution. However, in both “Teenage Driving, Mortality, and Risky Behaviors” and this paper, model assumptions are missing after fitting models before making conclusions. From explanatory data analysis part, we are able to see regression model with discontinuity is a proper model, but as next step, model assumption check in detail is necessary by checking residual plot, quantile to quantile plot, et.

Another weakness is in both “Teenage Driving, Mortality, and Risky Behaviors” and this paper, they only consider age and gender as two predictors that impacting death rates. However, there are many other available variables in the data set may have impact to death rates. As an improvement or next step, it is necessary to check whether other variables have significant impact to death rates, we can still use likelihood ratio test for all possible models comparisons.

Appendix

A More data visualization for other death rates




E
Death rate (per 100,000) because of homicide
Death rate
30
20
10
0
−50
−25
0
25
50
age since MDA measured in months
gender
Female
Male
F
Death rate (per 100,000) because of drowning
Death rate
4
3
2
1
0
−50
−25
0
25
50
age since MDA measured in months
gender
Female
Male

References

Huh, Jason, and Julian Reif. 2021. "Teenage Driving, Mortality, and Risky Behaviors." *Journal of American Economic Review: Insights* 3(4) (39): 523. https://doi.org/10.1257/aeri.20200653.

Kassambara, Alboukadel. 2020. *Ggpubr: 'Ggplot2' Based Publication Ready Plots*. https://CRAN.R-project.org/package=ggpubr.

Lee, David S., and Thomas Lemieux. 2010. "Regression Discontinuity Designs in Economics."

Norton, Amy. 2013. "Boys Have Higher Death Rates from Many Causes, Study Shows."

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

RStudio Team. 2020. *RStudio: Integrated Development Environment for r*. Boston, MA: RStudio, PBC. http://www.rstudio.com/.

Teenage Driving, Mortality, and Risky Behaviors Data Set. 2021. https://www.openicpsr.org/openicpsr/project/133501/version/V1/view?path=/openicpsr/133501/fcr:versions/V1/data&type=folder.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Zeileis, Achim, and Torsten Hothorn. 2002. "Diagnostic Checking in Regression Relationships." *R News* 2 (3): 7–10. https://CRAN.R-project.org/doc/Rnews/.