Project 3: US Traffic Fatalities

02/14/2020

1 Introduction

1.1 Background

Traffic fatalities are a major source of accident deaths at all ages, and almost half of drivers and more than 40% of passengers killed in vehicle crashes have been drinking [1]. The effects of drunk driving laws such as mandatory jail sentence on road fatalities is an important topic for policymakers in order to reduce the fatality rate. Our study is based on dataset "Fatalities" from the AER package in R which is a panel dataset reporting annual state level observations on U.S. traffic fatalities for the period 1982 through 1988.

There are two primary scientific questions we are interested in. The first question is whether there is an effect of mandatory jail sentence on fatality rate or not. The second question is whether there is a causal effect between jail sentence and fatality rate. In this study, we implement exploratory data analysis, fixed effect model, model diagnostics and propensity score. In the end, we draw our conclusion and present some suggestions to policymakers based on our assumptions and analysis.

1.2 Statistical questions of interest

To answer the primary scientific question of interests, we propose to fit a fixed effects model with fatality rate as responsible variable, economic conditions and drunk driving laws as predictor variable given state fixed effects and time fixed effects. We will then run our model diagnostic to check if the assumptions of the model hold and find whether having a mandatory jail sentence is associated with reduced traffic fatalities. Then, we use propensity score matching method to measure the causal effect between mandatory jail sentence and traffic fatality.

2 Analysis Plan

2.1 Population and study design

According to the description of the dataset, this observational study includes the fatality related information in 48 states for the period 1982 through 1988. We only focus on the vehicle fatality rate and variables related with economic condition and drunk driving laws.

2.2 Descriptive Analysis

First of all, we pick up all valuable variables we are interested in. To deal with missing value, we search literatures and fill in it. Then we do data pre-processing to better interpret variables. For some variables related with making policies, we draw the scatter plot and boxplots to show the relationship between each one and fatality rate respectively.

2.3 Fixed effects model

In order to eliminate bias from unobservable that change over time but are constant over entities and control for factors that differ across entities but are constant over time, we include both individual fixed effects and

time fixed effects in the model, i.e. we assume state and year as fixed effect. By combining them, the model is given as follows:

$$y_{it} = \alpha_i + \mu_t + X_{it}\beta + \varepsilon_{it}, \quad \text{for} \quad t = 1, 2, \cdots, 7 \quad \text{and} \quad i = 1, 2, \cdots, 48,$$

where y_{it} is the variable observed for i-th state at year t, X_{it} is the time-variant regressor vector including beetax, unemp, log(income), miles, drinkage, jail and service variables. β is the matrix of parameters for each variable, α_i is the unobserved time-invariant individual effect, μ_t is the unobserved individual-invariant time effect, ε_{it} is the error term.

Assumptions of T-test is $\varepsilon_{it} \overset{i.i.d}{\sim} N(0, \sigma^2)$.

Assumptions of fixed effected model specified as follows:

- $E(u_{it}|X_{i1}, X_{i2}, \cdots, X_{i7}, \alpha_i) = 0$.
- $(X_{i1}, X_{i2}, \cdots, X_{i7}, u_{i1}, u_{i2}, \cdots, u_{i7}), i = 1, 2, \cdots, 48$ are i.i.d drawn from their joint distribution.
- Large outliers are unlikely: (X_{it}, u_{it}) have non-zero finite fourth moments.
- There is no perfect multicollinearity.
- The error for a given state are uncorrelated over time, conditional on the regressors: specifically, $cov(u_{it}, u_{is}|X_{i1}, X_{i2}, \cdots, X_{i7}, \alpha_i) = 0$ for $t \neq s$.

2.4 Causal Inference

To test the causal effect of jail on traffic fatality rate, we choose to use propensity score matching method for observational study [2].

Because there are many other predictors that may affect the likelihood of being assigned into the treated group, we use a logit model for the propensity of observations to be assigned into the treated group, since jail is a binary variable. The propensity score model is a probit/logit model with treatment D as the dependent variable and other observations X as independent variables. The propensity score is the conditional (predicted) probability of receiving treatment given pre-treatment characteristics X: $p(X) = P(D = 1|X) = E[D|X]$.

Then we match observations from treated and control groups based on their propensity scores. There are several matching methods available, like kernel, nearest neighbor. In this case, we use a genetic search algorithm to determine the weight each covariate is given[3].

Because the treatment is unbalanced, there are more observations with jail is `no` than with jail is `yes`. Therefore, we prefer ATT(average treatment effect on the treated) than ATE(Average treatment effect). ATT is the difference between the outcomes of treated and the outcomes of the treated observations if they had not been treated. The definition is $ATT = E[Y_1|X, D = 1] - E[Y_0|X, D = 1]$. The second term is a counterfactual so it is not observable and needs to be estimated.

After matching on propensity scores, we can compare the outcomes of treated and control observations to get the empirical estimation of ATT. $\hat{ATT} = \frac{1}{n_1} \sum_{i \in \{D=1\}} [y_{1,i} - \sum_j w(i,j) y_{0,j}]$. Each treated observation i is matched j control observations and their outcomes y_0 are weighed by w.

There are three assumptions on propensity score matching:

(a) Conditional independence assumption. For observational studies, the outcomes are independent of treatment, conditional on x. $y_1, y_0 \perp D|x$.

(b) Matching assumption. For each treated observation, there is a matched control observation with similar x. $0 < P(D = 1|x) < 1$.

(c) Balancing condition. Assignment to treatment is independent of the x characteristics, given the same propensity score. $D \perp x|p(x)$.

3 Results

3.1 Descriptive Analysis

The dataset we investigate has 336 rows and 34 columns. This dataset is a panel data set reporting annual state level observations on U.S. traffic fatalities, including the information of 48 states for the period 1982 through 1988. In order to investigate whether having a mandatory jail sentence is associated with reduced traffic fatalities. Besides fatal, state, year and jail variables, we should also choose variables which may be served as confounder into account. From literature [5], we consider unemp, income, miles, drinkage, beertax, service into account. Moreover, we pre-process the data as follows.

- The original data is not balanced. The value of jail and service variables in row 28 are missing. By literature[6], mandatory jail sentence and mandatory community service policy were not launched. Both of two missing values are "no".
- Because of the different population size in 48 states, investigating the traffic fatality rate is more reasonable than the number of vehicle fatalities. It is measured as the number of fatalities per 10000 inhabitants.
- Compared with the value of other variables, the magnitude of the value of variable income are pretty large. For the sake of readability of coefficients of income, log-transformation is taken.
- By observing the values of drinkage, some of them are not integers but with two decimal places, which does not make sense in practice. We present a discretized version of drinkage that classifies states into four categories of minimal drinking age: $[18, 19), [19, 20), [20, 21)$ and $[21, 22]$.

After data pre-processing, we get nine variables: fatality rate, state, year, beertax, unemp, log(income), miles, drinkage, jail, where fatality rate represents the number of fatalities per 10000 inhabitants. Beertax represents tax on case of beer. Unemp represents unemployment rate. Log(income) represents the logarithm of real per capita income. Miles represents average miles per driver. Drinkage represents age intervals of minimal drinking age. Jail represents whether there exists mandatory jail sentence in one particular state.

Besides economic conditions, we are interested in driving and alcohol policy in order to make some suggestions to policymakers. We investigate the relationship between fatality rate and beertax, jail, drinkage respectively. From Figure.1, the results is contrary to our expectations, alcohol taxes, mandatory jail sentence, mandatory community service and higher minimal drinking age are supposed to lower the rate of traffic fatalities. This is possibly due to omitted variable bias, for example, economic conditions. We can't make a conclusion just from these figures. It indicates that we need to consider other covariates into study.

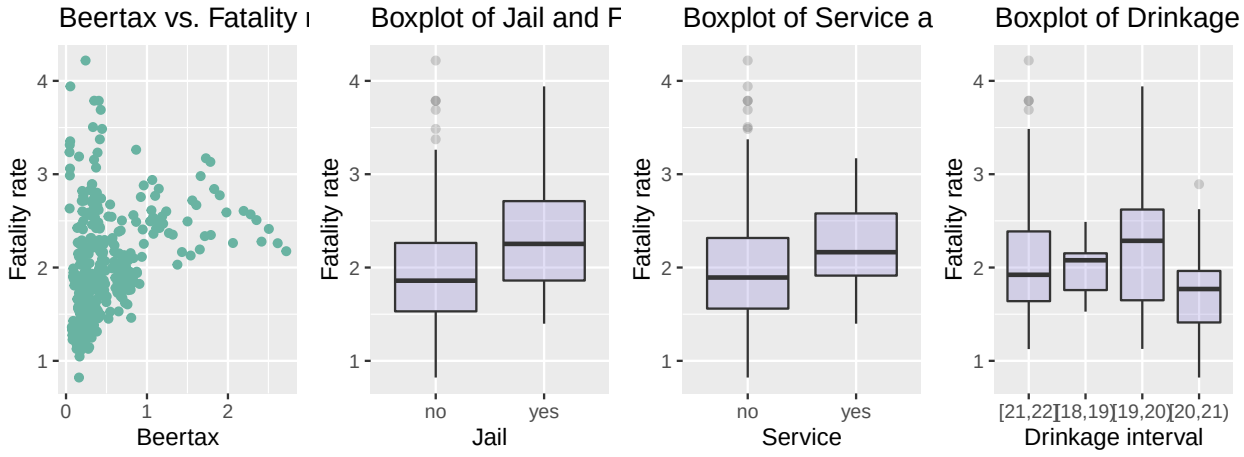

Figure.1 The Relationships between Fatality Rate and Other Variables

3.2 Fixed effects model

From calculation, the p value of F-test is less than $2.22 * 10^{-16}$, which means there is a significant linear association between fatality rate and predictor variables. For each variables, we have the null hypothesis $H_0 : \beta_i = 0 \quad \text{for} \quad i = 1, 2, \cdots, 8$ and the alternative hypothesis $H_a : \beta_i \neq 0 \quad \text{for} \quad i = 1, 2, \cdots, 8$. By t-test for every β, we can get that p values of drink, jail, breath, service and miles are greater than 0.05, and the p value of beertax, unemployment rate and income are less than 0.05. It means that beer tax, unemployment rate and income are statistically significant.

Table.1 Summary of fixed effects model

	beertax	drinkage	jailyes	breathyes	unemp	miles	log(income)	serviceyes
coefficient	-0.4607	0.0017	0.0136	0.0003	0.0625	$8.9571*10^{-6}$	1.7879	0.0337
p-value	0.0063	0.9258	0.9099	0.9952	$5.318*10^{-8}$	0.3071	1.599×10^{-6}	0.8075

3.3 Model Diagnostics

3.3.1 Zero-mean and equal variance (Homoscedasticity)

From the Residuals vs Fitted Values scatterplot (Figure.2), these points are uniformly distributed on both sides of x-axis, which means that our model satisfies the zero mean assumption and does not violate the equal variance assumption.

3.3.2 Independence

In the United States, states generally differ from each other. Every states make their own law, even though some of their situations are similar. Besides, we have chosen unobserved time-invariant individual effect as one of the variables to control the influence of different states. In addition, we think that the previous outcomes will not affect the future. Therefore, in our model, $(X_{i1}, X_{i2}, \cdots, X_{i7}, u_{i1}, u_{i2}, \cdots, u_{i7}), i = 1, 2, \cdots, 48$ are independently and identically distributed.

3.3.3 Normality

Through Q-Q Plot (Figure.2), we can see that the residuals are slightly heavy tailed compared with normal distribution. Since the p values of our variables are extremely small, slightly heavy tailed probability distribution has little influence to our result. Moreover, we do not know how to solve this situation that our model violates normality assumption.

3.3.4 Influential observations

To derive the leverage of the observations, a half-normal plot (Figure.2) is drawn to sort the observations by their leverage. And we can find that there is only one influential observations in the dataset. In order to maintain the balance of our dataset, we decide to reserve this point.

3.3.5 Multicollinearity

A popular way of diagnosing multicollinearity is through the calculation of variance inflation factors (VIFs). The VIF score indicates the proportion by which the variance of an estimator increases due to the inclusion of a particular covariate. From Table.2, the VIFs of the variables in the model ranges from 1.0 to 1.6, which

means that multicollinearity does not exist in our model. Therefore, the model coefficients are stable and unbiased.

Table.2 Variance Inflation Factors of Variables

beertax	drinkage	jailyes	breathyes	unemp	miles	log(income)	serviceyes
1.0505	1.0350	4.1313	1.0316	1.5182	1.0148	1.5483	4.1642

3.3.6 Uncorrelation

There is only one error for a given state each year, so the correlation value for a given state over time does not exist. Therefore, we can just assume that the model meets the uncorrelation assumption.

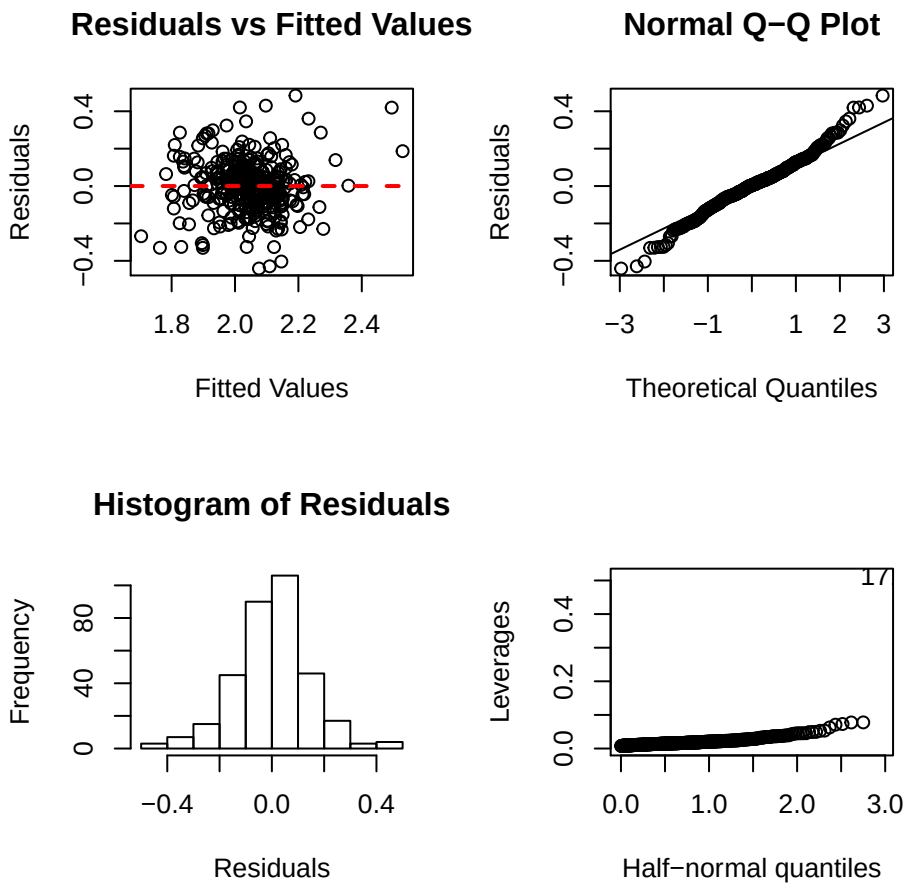

Figure.2 Plots of Model Diagnostics

3.4 Causal Inference

In this case, the jail sentence can be independent with fatality rate if other variables like mandatory community and minimum legal drinking age are given. There are more observations where jail is `no` than `yes`. So for treatment of jail sentence, there will always be matched control observation. Also, if propensity

score is given, the jail sentence is independent with other observations. So all assumptions are valid in this observation study.

Firstly, we build on a logistic regression on treatment `jail` with other predict variables to get propensity score in each case. Then we use GenMatch function in R to get the weight matrix. In the end we use Match function to get the estimation of ATT in this case[4]. The estimator is 0.19 and the p value is 0.07. When significance level is 0.05, we can not reject the null hypothesis that the estimation of ATT is 0. Thus, we can not accept the statement that there is causal effect of jail sentence on traffic fatality rate. By sensitivity test, if we adjust parameter gamma, which is odds of differential assignment to treatment due to unobserved factors, the bound of estimation can easily contain 0. Therefore, this causal effect is not significantly influential.

4 Conclusion

By fixed effect model, neither stiff punishments nor increases in the minimal legal drinking age have important effects on fatalities. In contrast, there is some evidence that increasing alcohol taxes, as measured by the real tax on beer, does reduce traffic deaths. Good economic conditions are associated with higher fatalities, perhaps because of increased traffic density when the unemployment rate is low or greater alcohol consumption when income is high[5]. Miles also has a significant positive estimate coefficient, and there is no doubt that the more people driving, the more possibility they would run into a car accident.

For casual inference, there is no significant evidence to show a mandatory jail sentence has causal effect on fatality rate. What is worse, the mandatory jail is slightly related to a higher fatality rate. It may be interpreted by a negative mentality on this policy, or a low cost to violate this regulation.

Here are suggestions for policymakers:

(a) Raising beer tax properly can effectively reduce the vehicle fatality rate. In detail, the effect of a $2.17 increase (in 1988 dollars) in the beer tax is a decrease in the expected fatality rate by 1 death per 10,000.

(b) There is no significant casual effect of jail on vehicle fatality rate. The policymaker should introspect on the rationalization of law in drinking and driving.

5 Discussion

The distribution of residuals is a little heavy-tailed and violates the normality assumption. However, if we use any transformation on response variable, the model will be less interpretable especially for observational study. Thus, the t test is less reliable in this case. We need further method like non-parameter test to avoid normality.

The p-value in linear regression model only means association, so we can't make a causal statement for quantitative variable such as beer tax. We need further model to study the causal effect of beer tax on fatality rate.

For GenMatch function in R, the parameter pop.size should be considered. In our analysis, we just use the default value 100. We should find an optimal one by further study.

6 Appendix. Reference

[1] Zobeck, Terry S., Frederick S. Stinson, Bridget F. Grant, and Darryl Bertolucci. 1993. Trends in Alcohol-Related Fatal Traffic Crashes, United States: 1979-91, Washington, D.C.: National Institue of Alcohol Abuse and Alcoholism, Surveeillance Report #26, November.

[2] Ani Katchova. 2013. Propensity Score Matching. https://sites.google.com/site/econometricsacademy/econometrics-models/propensity-score-matching

[3] Alexis Diamond; Jasjeet S. Sekhon. 2012. Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. Review of Economics and Statistics.

[4] Jasjeet Singh Sekhon. Package 'Matching'. https://cran.r-project.org/web/packages/Matching/Matching.pdf

[5] James H. Stock; Mark W. Watson. 2007. Introduction to Econometrics, 2nd Edition. Pearson.

[6] NATIONAL HIGHWAY TRAFFIC SAFETY ADMINISTRATION. Digest of State Alcohol Highway Safety-Related Legislation. https://nhtsa.dr.del1.nhtsa.gov/Driving-Safety/Impaired-Driving/Digest-of-State-Alcohol-Highway-Safety%E2%80%93Related-Legislation

7 Appendix II. Group Partners

This Document is the project 3 of Team 7 in STA 207, Winter 2020.

1. Bingdao Chen bdchen@ucdavis.edu contribution: descriptive analysis and model establishment
2. Yahui Li yhuli@ucdavis.edu contribution: casual inference and conclusion
3. Zihan Wang zihwang@ucdavis.edu contribution: fixed effects model
4. Jian Shi jnshi@ucdavis.edu contribution: model diagnose

The repository in Github is on https://github.com/yhli097/STA207Project3.git