Project3

Rongkui Han

2/6/2020

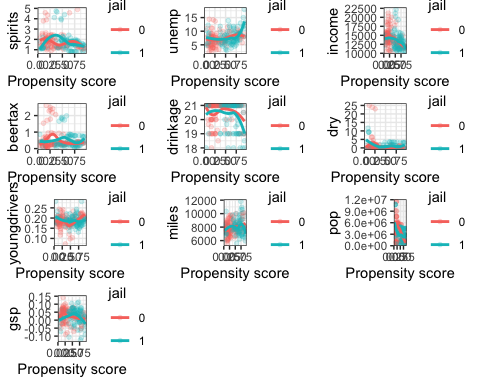
## Examining covariate balance in the matched sample

We’ll do three things to assess covariate balance in the matched sample:

* visual inspection
* t-tests of difference-in-means
* computation of the average absolute standardized difference (“standardized imbalance”)

fn\_bal <- function(dta, variable) {  
 dta$variable <- dta[, variable]  
 dta$jail <- as.factor(dta$jail)  
 support <- c(min(dta$variable), max(dta$variable))  
 ggplot(dta, aes(x = distance, y = variable, color = jail)) +  
 geom\_point(alpha = 0.2, size = 1.3) +  
 geom\_smooth(method = "loess", se = F) +  
 xlab("Propensity score") +  
 ylab(variable) +  
 theme\_bw() +  
 ylim(support)  
}  
  
grid.arrange(  
 #fn\_bal(dta\_m, "year"),  
 fn\_bal(dta\_m, "spirits"),  
 fn\_bal(dta\_m, "unemp"),  
 fn\_bal(dta\_m, "income"), #+ theme(legend.position = "none"),  
 fn\_bal(dta\_m, "beertax"),  
 #fn\_bal(dta\_m, "baptist"),  
 #fn\_bal(dta\_m, "mormon"),  
 fn\_bal(dta\_m, "drinkage"),  
 fn\_bal(dta\_m, "dry"),  
 fn\_bal(dta\_m, "youngdrivers"),  
 fn\_bal(dta\_m, "miles"),  
 #fn\_bal(dta\_m, "breath"),  
 fn\_bal(dta\_m, "pop"),  
 fn\_bal(dta\_m, "gsp"),  
 nrow = 4#, widths = c(1, 0.8)  
)

## Warning: Removed 19 rows containing missing values (geom\_smooth).



### Team ID: Team 6

#### NAME: Connor Rosenberg

#### NAME: Rongkui Han

#### NAME: Yuqing Yang

#### NAME: Nassim Ali-Chaouche

## 1.0 Introduction

#### 1.1 Background

Traffic accidents cause thousands of deaths in the United States every year. Data pertinent to US traffic fatalities from the years 1982 to 1988 can be easily accessed in the “Fatalities” dataset. The data was obtained from sources such as the US Department of Transportation Fatal Accident Reporting System (FARS) and the US Bureau of Labor Statistics. The dataset includes panel data for 48 states (Alaska and Hawaii not included), containing demographic variables such as population, income per capita, religious belief, and unemployment rate. In addition, features that are commonly associated with traffic accidents and its regulation, such as average miles per driver, percentage of young drivers, tax collected per case of beer, presence of a preliminary breath test law, and whether the state implemented mandatory jail sentences or community service for an initial drunk driving conviction, were also presented in the dataset. Finally, the number of vehicle fatalities and its numerous subsets, such as night-time or single-vehicle fatalities, were introduced. The observations were recorded for each state annually. In total, there are 336 observations recorded for 34 distinct variables.

Due to the observational nature of the data, obtaining causal effects may pose a challenge. In observational studies, treatment selection is often influenced by subject characteristics. In the context of our study, “treatment assignment” refers to whether a state has a mandatory jail sentence for an initial drunk driving conviction. It is not difficult to imagine that demographic characteristics of a state can influence both its traffic legislations as well as its traffic fatality rate, causing confounding effects that obscure the impact of legislation on traffic fatality. As a result, systematic differences in baseline characteristics between states with and without mandatory jail sentences must be taken into account when estimating its effect on outcomes. The **propensity score** is the probability of treatment assignment conditional on observed baseline characteristics. The propensity score allows one to analyze an observational study so that it mimics some of the particular characteristics of a randomized controlled trial. In particular, conditional on the propensity score, the distribution of observed baseline covariates will be similar between treated and untreated subjects, allowing the estimation of the average treatment effect (Austin, 2011). In this report, we will attempt to discover the potential causal relationship between a mandatory jail sentence for an initial drunk driving conviction and the traffic fatality rate of the state using the **propensity score matching** technique, followed by **mixed-effect ANOVA modeling**. The primary objective of this analysis is to educate State legislators on whether a mandatory jail sentence is a proper current and will result in lower automobile fatality rates.

#### 1.2 Questions of Interest

* Are there demographic features that correlate with a state’s mandatory jail sentence law?
* Is a state’s mandatory jail sentence law associated with its annual traffic fatality rate, without adjusting for potential covariating demographic variables?
* Is a state’s mandatory jail sentence law associated with its annual traffic fatality rate, after adjusting for potential covariating demographic variables?
* Can we draw a causal conclusion regarding the relationship between a state’s mandatory jail sentence law and its annual traffic fatality rate?

## 2.0 Analysis Plan

From data collected by the National Highway Traffic Safety Administration’s FARS, we plan to conduct a propensity score analysis followed by mixed-effect ANOVA modeling to isolate the average effect of required jail time on automobile fatality rates.

#### 2.1 Population and study design

The response variable used in this analysis is the yearly traffic fatality rate per 10,000 population. This statistic will be calculated by dividing a state’s total traffic fatality count within a given year by its population of the year, multiplied by 10,000.

The “Fatalities” dataset is a longitudinal dataset, also know as “panel dataset”, with a “year” index that delineates the temporal order of entries. In a typical panel dataset, each individual has multiple entries, some of which correspond to pre-treatment records while others post-treatment. In this regard, the vehicle fatality dataset is atypical, because only six out of 48 states had pre- and post-treatment records (*Figure X*). Most states did not change their mandatory jail sentence policy between 1982 and 1988. Due to this limitation, common panel data analysis techniques are not applicable. In response, we will follow a bench-marked pipeline for analyzing non-panel observational data under a propensity score matching framework (Smith, 1997), while closely monitoring the behavior of the temperal variable throughout the analysis, and address any potential issues that might arise.

#### 2.2 Descriptive Analysis

#### 2.3 Propensity Score Analysis

##### 2.3.1 Propensity Score Estimation

We will estimate the propensity score through a logistic regression model. The dependent variable of the logistic regression model is a binary variable indicating treatment status, whether or not a State has a mandatory jail sentence. There is a lack of consensus in the applied literature as to which variables to include in the propensity score model (Austin, 2001). Brookhart et al. (2006) suggested that for practical purposes, it is safe to include all potential confounding covariables in propensity score estimation. In this study, 14 independent variables were included in the logistic regression model to account for demographic characteristics that could potentially influence whether a state mandates such a jail sentence. These variables include year, population, gross state product (GSP), spirits consumption, unemployment rate, per capita income, tax on a case of beer, percentage of baptists, percentage of Mormons, minimum drinking age, percent residing in dry counties, percentage of drivers younger than 24, average miles per driver, and preliminary breath test upon initial drunk driving conviction.

The output of this model is the propensity score, which equals to the probability that a State has a mandatory jail sentence given the set of covariates. The logistic regression model we used to estimate the propensity score is as follows:

Where , is the indicator variable for mandatory jail sentence upon initial drunk driving conviction. when the state has mandatory jail sentence, and other wise. is a vector of length 14, indicating the realized value of the 14 independent variables of the i-th subject in the logistic regression model. .

##### 2.3.2 Matching

To match observations with mandatory jail sentences to observations without, we will use the nearest neighbor matching algorithm based upon propensity score.

##### 2.3.3 Examining covariate balance in the matched sample

We must assess the covariate balance in our matched treated and untreated sample sets to ensure a near-random distribution covriates within each set. We will perform visual inspections and t-tests to check for difference in means.

##### 2.3.4 Estimate Treatment Effect

To estimate the effect of mandatory jail sentences on a State’s traffic fatality rate, we will fit the following linear regression relating the binary treatment variable to Fatality Rate:

for and .

Where  
- represents the yearly traffic fatality rate of a given State in a given year.  
- represents the overall sample mean of yearly traffic fatality rates.  
- represents the fixed effect of mandatory jail sentence law. when the state has mandatory jail sentence, and other wise.  
- represents the random effect of State , and  
- the residuals.

The model is constrained such that , , and .

#### 2.4 Model Diagnostics

We will use Q-Q plot, histogram and the Shapiro-Wilk test inspect the normality of residuals. A scatter plot and a fitted-value-versus-residual scatter plot and the Levene test will be used to examine equality of residual variance. Independence of residuals and outlying data points will also be discussed.

## 3.0 Results

#### 3.1 Descriptive Analysis

#### 3.2 Propensity Score Analysis

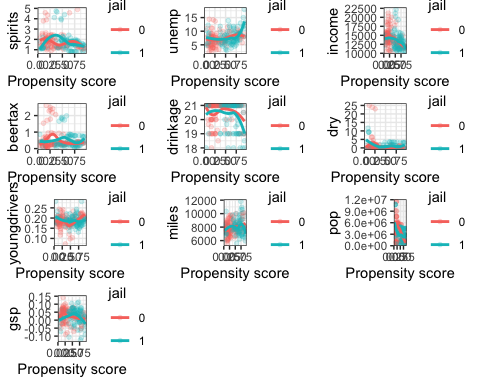
##### 3.2.1 Propensity Score Estimation

A logistic regression model was fit to estimate propensity score for matching treated and untreated samples. Distribution of the estimated propensity scores are displayed in *Figure Y*.

*Figure T*: Distribution of quantitative dependent variables by propensity score:

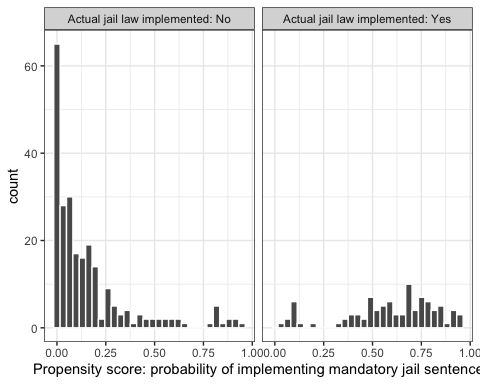
fn\_bal <- function(dta, variable) {  
 dta$variable <- dta[, variable]  
 dta$jail <- as.factor(dta$jail)  
 support <- c(min(dta$variable), max(dta$variable))  
 ggplot(dta, aes(x = distance, y = variable, color = jail)) +  
 geom\_point(alpha = 0.2, size = 1.3) +  
 geom\_smooth(method = "loess", se = F) +  
 xlab("Propensity score") +  
 ylab(variable) +  
 theme\_bw() +  
 ylim(support)  
}  
  
grid.arrange(  
 #fn\_bal(dta\_m, "year"),  
 fn\_bal(dta\_m, "spirits"),  
 fn\_bal(dta\_m, "unemp"),  
 fn\_bal(dta\_m, "income"), #+ theme(legend.position = "none"),  
 fn\_bal(dta\_m, "beertax"),  
 #fn\_bal(dta\_m, "baptist"),  
 #fn\_bal(dta\_m, "mormon"),  
 fn\_bal(dta\_m, "drinkage"),  
 fn\_bal(dta\_m, "dry"),  
 fn\_bal(dta\_m, "youngdrivers"),  
 fn\_bal(dta\_m, "miles"),  
 #fn\_bal(dta\_m, "breath"),  
 fn\_bal(dta\_m, "pop"),  
 fn\_bal(dta\_m, "gsp"),  
 nrow = 4#, widths = c(1, 0.8)  
)

## Warning: Removed 19 rows containing missing values (geom\_smooth).



*Figure Y*: Propensity score distribution across two treatment groups

prs\_df %>%  
 mutate(jail = ifelse(jail == 1, labs[1], labs[2])) %>%  
 ggplot(aes(x = pr\_score)) +  
 geom\_histogram(color = "white", bins = 30) +  
 facet\_wrap(~jail) +  
 xlab("Propensity score: probability of implementing mandatory jail sentence") +  
 theme\_bw()



##### 3.2.2 Matching

Nearest neighbor matching algorithm resulted in a new dataset of 188 entires, with 94 original entries with mandatory jail sentence matched with their respective non-treated entry with the most similar propensity score.

##### 3.2.3 Examining covariate balance in the matched sample

##### 3.2.4 Estimate Treatment Effect

#### 3.3 Model Diagnostics

##### 3.3.1 Goodness-of-fit of logistic regression model

Unlike linear regression with ordinary least squares estimation, there is no statistic which explains the proportion of variance in the dependent variable that is explained by the predictors. The most notable pseudo metric commonly used for assessing goodness-of-fit in logistic regression is McFadden’s , which is defined as where is the log likelihood value for the fitted model and is the log likelihood for the null model with only an intercept as a predictor. The propensity-score-generating logistic regression model used in this study has a McFadden’s of 0.39, indicating effective removal of a large portion of confounded samples through the propensity score matching process.

##### 3.3.2 Mixed Effect model

## 4.0 Discussion

### Propensity Score Matching

In this report, we highlight the usage of propensity score matching for isolating average treatment effect of implementing mandatory jail sentence upon initial conviction of driving under influence. Rosenbaum and Rubin (1983) defined treatment assignment to be strongly ignorable if the following two conditions hold: (a) treatment assignment is independent of the potential outcomes conditional on the observed baseline covariates, and (b) every subject has a nonzero probability to receive either treatment. They demonstrated that if treatment assignment is strongly ignorable, conditioning on the propensity score allows one to obtain unbiased estimates of average treatment effects. We achieved these conditions through propensity score matching.

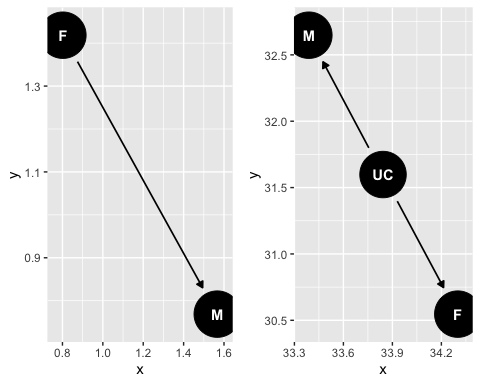
### Causal Inference

We cannot confidently draw causal conclusions using the result from this analysis. This is because of the violation of these assumptions necessary for causal inference:

1. The stable unit treatment value assumption (SUTVA):  
   SUTVA states that the treatment assignment of one experimental unit cannot interfere with the outcome of a separate experimental unit. This is violated in this dataset because
2. temporal correlations across records taken from the same State, and
3. spacial correlations among states that are geographically close to each other.  
   The implementation of mandatory jail sentence in one state is very likely to influence the vehicle fatality of an adjacent state. And the historic state of the legislation is very likely to influence the vehicle fatality of the same state in the years to come.
4. Exogeneity: The exogeneity assumption states that the independent variable (implementation of mandatory jail sentence) cannot be dependent on the dependent variable (vehicle fatality rate). This assumption is likely to be violated in this dataset because legislations can arise from existing conditions (*Figure Za*).
5. Ignorabilty of unobserved potential confounding variables: Although this expansive dataset captures many prominent variables that are associated with vehicle fatality rate, many more economic and social factors come into play in impacting the interactions between the implementation of mandatory jail sentence and vehicle fatality rate. We cannot confidently exclude the possiblity of the existence of many of such variables (*Figure Zb*).

*Figure Z: directed acyclic graphs of scenarios where direction of causal inference changes. a. Violation of exogeneity. b. Violation of ignorability of unobserved confounding variables. M: mandatory jail sentence; F: vehicle fatality rate; UC: unobserved confounding variable*

dag1 = dagitty("dag{ F -> M }")  
dag2= dagitty("dag{ M <- UC -> F }")  
grid.arrange(ggdag(dag1), ggdag(dag2), nrow = 1)



## 5.0 Reference

Austin P. C. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate behavioral research, 46(3), 399–424. <doi:10.1080/00273171.2011.568786>  
Brookhart M.A., Schneeweiss S., Rothman K.J., Glynn R.J., Avorn J., Stürmer T. (2006). Variable selection for propensity score models. American Journal of Epidemiology. 163, 1149–1156.  
Rosenbaum P.R., Rubin D.B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika. 70:41–55. Smith, H. L. (1997). 6. Matching with Multiple Controls to Estimate Treatment Effects in Observational Studies. Sociological methodology, 27(1), 325-353.