Project3

Rongkui Han

2/6/2020

0.1 Propensity Score matching

```
library(MatchIt)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(lme4)
```

```
## Loading required package: Matrix
```

```
library(AER)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
## Registered S3 methods overwritten by 'car':
##   method                          from
##   influence.merMod                lme4
##   cooks.distance.influence.merMod lme4
##   dfbeta.influence.merMod         lme4
##   dfbetas.influence.merMod        lme4
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## Loading required package: lmtest
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Loading required package: sandwich
```

```
## Loading required package: survival
```

```
data("Fatalities")
data = Fatalities

#Impute values for missing CA data
data[28,15] = as.factor("no")
data[28,16] = as.factor("no")


#head(data)
#dim(data)
data['fr'] = data$fatal/data$pop*10000
```

0.1.1 Propensity score estimation

We estimate the propensity score by running a logit model (probit also works) where the outcome variable is a binary variable indicating treatment status. What covariates should you include? For the matching to give you a causal estimate in the end, you need to include any covariate that is related to **both the treatment assignment and potential outcomes**. I choose just a few covariates below—they are unlikely to capture all covariates that should be included. You'll be asked to come up with a potentially better model on your own later.

```
data$jail = ifelse(data$jail == 'yes', 1, 0)
m_ps = glm(jail ~ year + spirits + unemp + income + beertax + baptist + mormon + drinkage + dry + young
summary(m_ps)
```

```
##
## Call:
## glm(formula = jail ~ year + spirits + unemp + income + beertax +
##     baptist + mormon + drinkage + dry + youngdrivers + miles +
##     breath + pop + gsp, family = binomial(), data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3873  -0.5675  -0.2154   0.5790   2.5088
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   9.307e-01  5.635e+00   0.165 0.868818
## year1983      6.173e-01  7.023e-01   0.879 0.379434
## year1984      1.747e+00  7.878e-01   2.218 0.026575 *
## year1985      2.445e+00  7.508e-01   3.257 0.001128 **
## year1986      2.822e+00  8.370e-01   3.372 0.000747 ***
## year1987      3.180e+00  9.162e-01   3.471 0.000518 ***
## year1988      3.672e+00  1.026e+00   3.578 0.000347 ***
## spirits       2.975e-01  3.034e-01   0.980 0.326873
## unemp         5.322e-01  1.219e-01   4.367 1.26e-05 ***
## income        5.067e-06  1.409e-04   0.036 0.971313
```

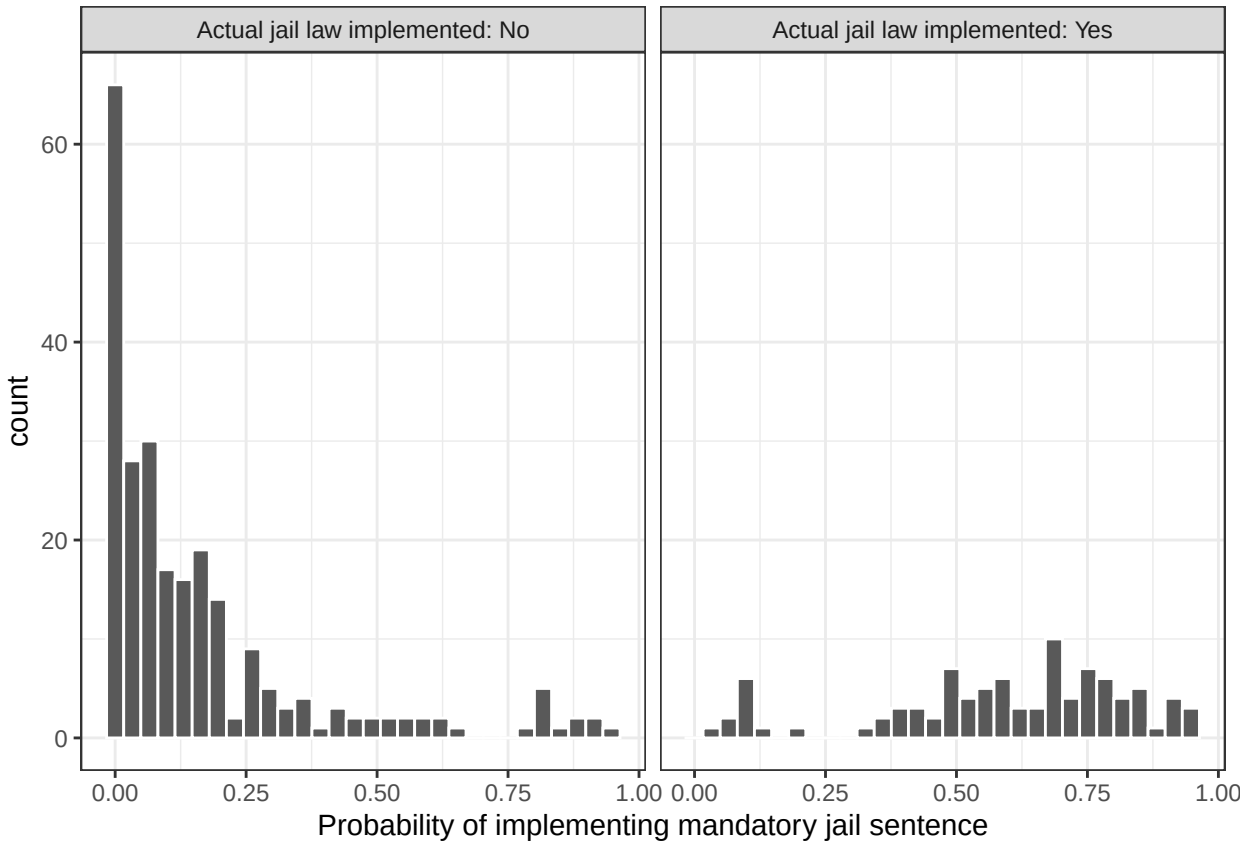
```
## beertax      -1.756e-02  4.622e-01  -0.038 0.969699
## baptist      -2.878e-02  2.761e-02  -1.042 0.297270
## mormon       -4.713e-04  1.627e-02  -0.029 0.976891
## drinkage     -2.129e-01  2.116e-01  -1.006 0.314282
## dry          -1.336e-01  4.003e-02  -3.337 0.000847 ***
## youngdrivers  4.268e+00  8.879e+00   0.481 0.630760
## miles        -2.889e-04  2.253e-04  -1.282 0.199806
## breathyes    -2.860e+00  4.638e-01  -6.166 7.00e-10 ***
## pop          -3.545e-07  9.608e-08  -3.690 0.000224 ***
## gsp           9.105e+00  5.670e+00   1.606 0.108326
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 398.31  on 335  degrees of freedom
## Residual deviance: 243.89  on 316  degrees of freedom
## AIC: 283.89
##
## Number of Fisher Scoring iterations: 6
```

```
prs_df = data.frame(pr_score = predict(m_ps, type = "response"),
                    jail = m_ps$model$jail)
#head(prs_df)
#dim(prs_df)
```

0.1.1.1 Examining the region of common support

After estimating the propensity score, it is useful to plot histograms of the estimated propensity scores by treatment status:

```
labs = paste("Actual jail law implemented:", c("Yes", "No"))
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("Probability of implementing mandatory jail sentence") +
  theme_bw()
```



0.2 Executing a matching algorithm

A simple method for estimating the treatment effect of Catholic schooling is to restrict the sample to observations within the region of common support, and then to divide the sample within the region of common support into 5 quintiles, based on the estimated propensity score. Within each of these 5 quintiles, we can then estimate the mean difference in student achievement by treatment status. Rubin and others have argued that this is sufficient to eliminate 95% of the bias due to confounding of treatment status with a covariate.

However, most matching algorithms adopt slightly more complex methods. The method we use below is to find pairs of observations that have very similar propensity scores, but that differ in their treatment status. We use the package MatchIt for this. This package estimates the propensity score in the background and then matches observations based on the method of choice ("nearest" in this case).

```
data_cov = c("year", "spirits" , "unemp" , "income" , "beertax" , "baptist" , "mormon" , "drinkage" , "c

mod_match = matchit(jail ~ year + spirits + unemp + income + beertax + baptist + mormon + drinkage + dr

dta_m = match.data(mod_match)
#dim(dta_m)
```

0.3 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")

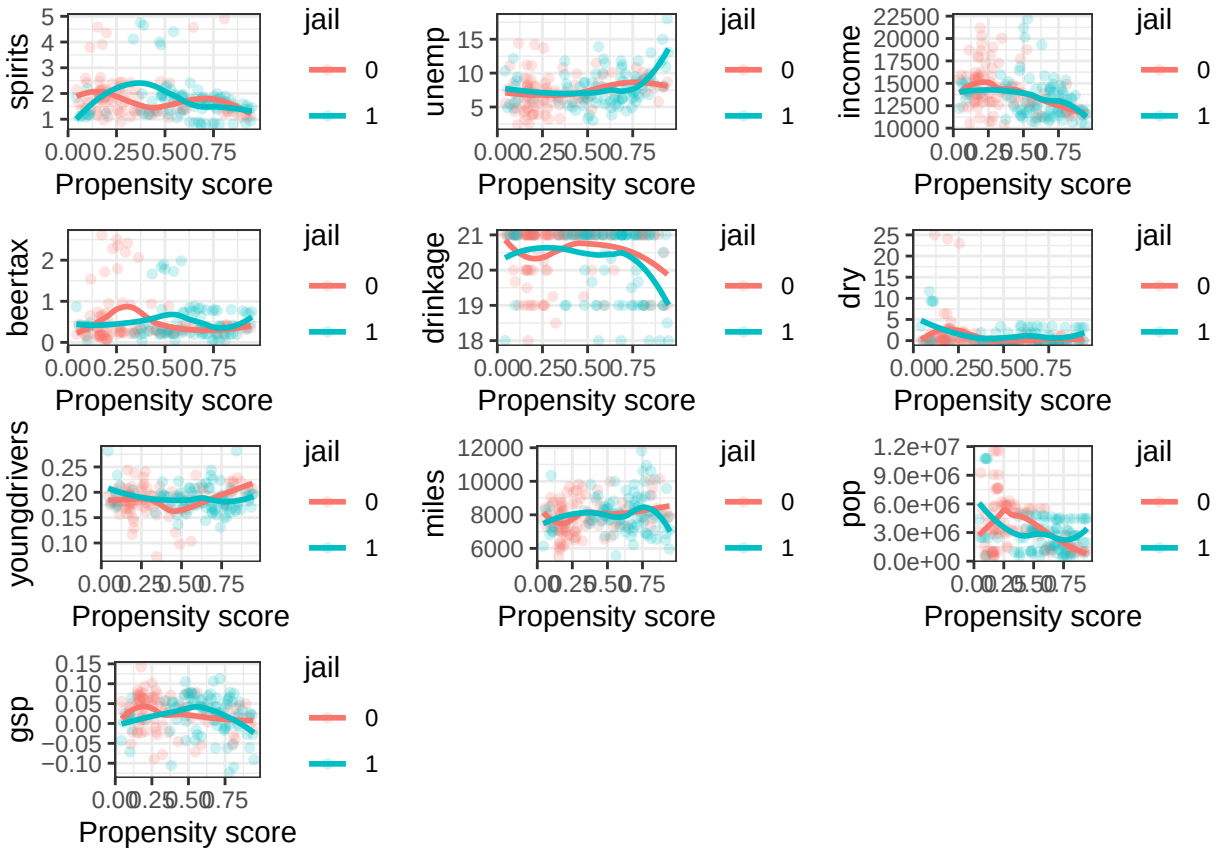
```
fn_bal <- function(dta, variable) {
  dta$variable <- dta[, variable]
  dta$jail <- as.factor(dta$jail)
  support <- c(min(dta$variable), max(dta$variable))
  ggplot(dta, aes(x = distance, y = variable, color = jail)) +
    geom_point(alpha = 0.2, size = 1.3) +
    geom_smooth(method = "loess", se = F) +
    xlab("Propensity score") +
    ylab(variable) +
    theme_bw() +
    ylim(support)
}

library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
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).
```



0.3.0.1 Difference-in-means

The means below indicate that we have attained a high degree of balance on the five covariates included in the model.

```
dta_m_mean = dta_m %>%
  group_by(jail) %>%
  summarise_all(funs(mean))
```

```
## Warning: funs() is soft deprecated as of dplyr 0.8.0
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## This warning is displayed once per session.
```

```
## Warning in mean.default(state): argument is not numeric or logical: returning NA
```

```
## Warning in mean.default(state): argument is not numeric or logical: returning NA
```

```
## Warning in mean.default(year): argument is not numeric or logical: returning NA
```

```
## Warning in mean.default(year): argument is not numeric or logical: returning NA
```

```
## Warning in mean.default(breath): argument is not numeric or logical: returning
## NA
```

```
## Warning in mean.default(breath): argument is not numeric or logical: returning
## NA
```

```
## Warning in mean.default(service): argument is not numeric or logical: returning
## NA
```

```
## Warning in mean.default(service): argument is not numeric or logical: returning
## NA
```

```
dta_m_mean = dta_m_mean[,c('jail', data_cov)]
dta_m_mean
```

```
## # A tibble: 2 x 15
##    jail  year spirits unemp income beertax baptist mormon drinkage   dry
##   <dbl> <dbl>   <dbl> <dbl>  <dbl>   <dbl>   <dbl>  <dbl>    <dbl> <dbl>
## 1     0    NA    1.85  7.18 14282.   0.534    6.52   3.18     20.5  1.49
## 2     1    NA    1.70  7.94 13327.   0.485    5.90   6.18     20.3  1.11
## # ... with 5 more variables: youngdrivers <dbl>, miles <dbl>, breath <dbl>,
## #   pop <dbl>, gsp <dbl>
```

You can test this more formally using t-tests. Ideally, we should not be able to reject the null hypothesis of no mean difference for each covariate:

Not working:

```
lapply(data_cov, function(v) {
    t.test(dta_m[,v] ~ dta_m$jail)
})
```

0.4 Estimating treatment effects

Estimating the treatment effect is simple once we have a matched sample that we are happy with. We can use a t-test:

```
with(dta_m, t.test(fr ~ jail))
```

```
##
##  Welch Two Sample t-test
##
## data:  fr by jail
## t = -3.6346, df = 179.54, p-value = 0.0003633
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.4983973 -0.1476523
## sample estimates:
## mean in group 0 mean in group 1
##        1.971571        2.294596
```

```
summary(data$fr)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.8212  1.6237  1.9560  2.0404  2.4179  4.2178
```

Or OLS:

```
lm_treat1 <- lm(fr ~ jail, data = dta_m)
summary(lm_treat1)
```

```
##
## Call:
## lm(formula = fr ~ jail, data = dta_m)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.15036 -0.46664 -0.08058  0.36787  2.24627
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.97157    0.06284  31.373  < 2e-16 ***
## jail         0.32302    0.08887   3.635  0.00036 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6093 on 186 degrees of freedom
## Multiple R-squared:  0.06631,    Adjusted R-squared:  0.06129
## F-statistic: 13.21 on 1 and 186 DF,  p-value: 0.0003601
```

```
confint(lm_treat1)
```

```
##                 2.5 %   97.5 %
## (Intercept) 1.8475933 2.095549
## jail        0.1476936 0.498356
```

```
lm_orig = lm(fr ~ jail, data = data)
summary(lm_orig)
```

```
##
## Call:
## lm(formula = fr ~ jail, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.12051 -0.41380 -0.07913  0.33630  2.27612
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.94172    0.03526  55.070  < 2e-16 ***
## jail         0.35287    0.06666   5.293 2.18e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5485 on 334 degrees of freedom
## Multiple R-squared:  0.0774, Adjusted R-squared:  0.07464
## F-statistic: 28.02 on 1 and 334 DF,  p-value: 2.179e-07
```

```
confint(lm_orig)
```

```
##                 2.5 %   97.5 %
## (Intercept) 1.8723658 2.011081
## jail        0.2217429 0.484002
```

```r
sum(dta_m$jail == 1) #94
```

```
## [1] 94
```

```r
sum(dta_m$jail == 0) #94
```

```
## [1] 94
```

```r
sum(data$jail == 1)
```

```
## [1] 94
```

```r
sum(data$jail == 0)
```

```
## [1] 242
```

```r
mean(dta_m$fr[dta_m$jail == 1]) - mean(dta_m$fr[dta_m$jail == 0])
```

```
## [1] 0.3230248
```

```r
mean(dta_m$fr[dta_m$jail == 0])
```

```
## [1] 1.971571
```

```r
mean(dta_m$fr[dta_m$jail == 1])
```

```
## [1] 2.294596
```

```r
mean(data$fr[data$jail == 1]) - mean(data$fr[data$jail == 0])
```

```
## [1] 0.3528725
```

```r
plot(dta_m$year)
```


```r
plot(data$year)
```


```
#What about if we inlcude state in the linear regression model

#Fixed
fit1 = lm(fr ~ jail + state, data = dta_m)
summary(fit1)
```

```
##
## Call:
## lm(formula = fr ~ jail + state, data = dta_m)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.58488 -0.08507  0.00216  0.08470  0.72365
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.27109    0.11527  19.702  < 2e-16 ***
## jail         0.04644    0.08352   0.556 0.579040
## stateaz      0.38837    0.16112   2.411 0.017124 *
## stateco     -0.42601    0.14581  -2.922 0.004013 **
## statect     -0.83412    0.14581  -5.721 5.47e-08 ***
## statede     -0.27209    0.16302  -1.669 0.097166 .
## stategа      0.16827    0.14118   1.192 0.235169
## stateid      0.30058    0.13778   2.182 0.030673 *
## stateil     -0.91968    0.16302  -5.642 8.01e-08 ***
## statein     -0.42328    0.14118  -2.998 0.003174 **
## stateia     -0.69887    0.15249  -4.583 9.50e-06 ***
## stateks     -0.34786    0.16112  -2.159 0.032412 *
## statela     -0.19670    0.16112  -1.221 0.224032
## stateme     -0.44740    0.16112  -2.777 0.006180 **
## statema     -1.07170    0.13778  -7.779 1.04e-12 ***
## statemi     -0.60920    0.16302  -3.737 0.000263 ***
## statemo     -0.29364    0.13778  -2.131 0.034675 *
## statemt      0.58549    0.16112   3.634 0.000381 ***
## statene     -0.79537    0.23054  -3.450 0.000726 ***
```

```
## statenv      0.43437    0.15527   2.798 0.005816 **
## statenh     -0.48602    0.14581  -3.333 0.001078 **
## statenj     -0.95937    0.14581  -6.580 7.18e-10 ***
## statenm      1.38211    0.13778  10.032  < 2e-16 ***
## statend     -0.75511    0.18226  -4.143 5.67e-05 ***
## stateoh     -0.79391    0.17387  -4.566 1.02e-05 ***
## stateok     -0.15985    0.14581  -1.096 0.274681
## stateor     -0.12711    0.15014  -0.847 0.398530
## stateri     -1.20307    0.15249  -7.889 5.54e-13 ***
## statesc      0.51078    0.15527   3.290 0.001247 **
## statetn      0.08554    0.16112   0.531 0.596253
## stateut     -0.47506    0.15527  -3.060 0.002620 **
## statevt     -0.18893    0.18226  -1.037 0.301572
## statewa     -0.64032    0.16112  -3.974 0.000109 ***
## statewv     -0.01690    0.16112  -0.105 0.916585
## statewi     -0.70931    0.23054  -3.077 0.002483 **
## statewy      0.90001    0.16112   5.586 1.04e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1997 on 152 degrees of freedom
## Multiple R-squared:  0.9181, Adjusted R-squared:  0.8992
## F-statistic: 48.66 on 35 and 152 DF,  p-value: < 2.2e-16
```

```r
#Random
fit2 <- lmer(fr ~ jail + (1 | state), data = dta_m)
summary(fit2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: fr ~ jail + (1 | state)
##    Data: dta_m
##
## REML criterion at convergence: 61.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.8315 -0.4840  0.0085  0.3928  3.7203
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  state    (Intercept) 0.32466  0.5698
##  Residual             0.03989  0.1997
## Number of obs: 188, groups:  state, 35
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  1.99462    0.10258  19.444
## jail         0.09928    0.07764   1.279
##
## Correlation of Fixed Effects:
##      (Intr)
## jail -0.303
```

0.4.1 Team ID: Team 6

0.4.1.1 NAME: Connor Rosenberg

0.4.1.2 NAME: Rongkui Han

0.4.1.3 NAME: Yuqing Yang

0.4.1.4 NAME: Nassim Ali-Chaouche

0.5 1.0 Introduction

0.5.0.1 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.

0.5.0.2 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?

0.6 2.0 Analysis Plan

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

0.6.0.1 2.1 Population and study design

0.6.0.2 2.2 Descriptive Analysis

spaghetti plot

0.6.0.3 2.3 Propensity Score Analysis

0.6.0.3.1 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; the 15 independent variables of the logistic regression model are demographic variables that could potentially influence whether a state mandates such a jail sentence. These variables include year, population, 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 our set of covariates. The logistic regression model we used to estimate the propensity score is as follows:

$$log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik},$$

Where $\pi_i = P(Y_i = 1|\overrightarrow{X_i} = \overrightarrow{x_i})$, Y_i is the indicator variable for mandatory jail sentence upon initial drunk driving conviction. $Y_i = 1$ when the state has mandatory jail sentence, and $Y_i = 0$ other wise. $\overrightarrow{X_i}$ is a vector of length 15, indicating the realized value of the 15 independent variables of the i-th subject in the logistic regression model. $k = 1, ..., 15, i = 1, ..., 336$

0.6.0.3.2 2.3.2 Matching

To match the samples with mandatory jail sentences to samples without, we will divide the data contained within the region of common support into 5 quintiles, based on the estimated propensity score. Within each of these 5 quintiles, we can then estimate the mean difference in fatality rate by jail status. Rubin and others have argued that this is sufficient to eliminate 95% of the bias due to the confounding of treatment status with a covariate.

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

0.6.0.3.3 2.3.3 Examining covariate balance in the matched sample

We must assess the covariate balance in our matched sample to ensure our assumptions for a propensity score are met. We will perform a visual inspection through covariate balance plots and perform several t-tests for difference-in-means.

0.6.0.3.4 2.3.4 Estimate Treatment Effect

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

0.6.0.4 2.4 Model Diagnostics

0.7 3.0 Results

0.7.0.1 3.1 Descriptive Analysis

0.7.0.2 3.2 Propensity Score Analysis

0.7.0.2.1 3.2.1 Propensity Score Estimation

0.7.0.2.2 3.2.2 Matching

0.7.0.2.3 3.2.3 Examining covariate balance in the matched sample

0.7.0.2.4 3.2.4 Estimate Treatment Effect

0.7.0.3 3.3 Model Diagnostics

0.7.0.4 3.4 Causal Effects

0.8 4.0 Discussion

0.8.1 Propensity Score Matching

0.8.2 Causal Inference

0.8.3 Caveats of the study

0.9 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