STA 207 Project 3: Analysis of US Traffic Fatalities Data

Team ID: 12

Name (responsibilities): Joseph Gonzalez (Proofread, Introduction, Background)

Name (responsibilities): Yanhao Jin (Main analysis, Causal statement, Conclusion and discussion)

Name (responsibilities): Ruichen Xu (Descriptive analysis, Plots and tables, Data processing)

Name (responsibilities): Bohao Zou (Main analysis, Model diagnostics)

1. Introduction

This study's purpose is to estimate the effects of drunk driving deterrents and other possible related factors on the number of vehicle fatalities. In this project, we use an annual time-series data set that contains state cross-sections for the 48 contiguous states of the U.S. from 1982 to 1988. For the data set and experiment, (a) This dataset includes an extensive set of categorical variables that describe the various state laws related to driving or alcohol. These laws represent all important DUI legislation. (b) During the 1980s, numerous states authorized police to administer roadside breath tests for alcohol, enacted administrative per se laws that require license suspension or revocation if a driver's blood alcohol content (BAC) exceeds a prespecified level and mandated minimum jail sentences or community service for driving under the influence.

In particular, we are interested in the following issues: 1. find the important factors that can potentially affect the fatality rate in the state; 2. determine if a mandatory jail sentence is associated with reduced traffic fatalities; 3. Based on the analysis, make suggestions for policymakers to reduce the fatality rate.

2. Statistical Analysis

Disparities in difficult to observe characteristics such as road conditions, driving patterns, and social attitudes towards drinking may influence interstate differences in vehicle mortality. Studies that choose to ignore this heterogeneity may generate biased estimates, which are a result of the correlation between the unobserved factors and cross-state variations in alcohol policies. To balance this heterogeneity, it is appropriate to use the linear fixed-effect model. That is

$$Y_{it} = \boldsymbol{\beta}_1^T \mathbf{X}_{it} + \boldsymbol{\beta}_2^T \mathbf{Z}_i + \varepsilon_{it}$$

where Y_{it} is the fatality rate for the i-th state and time t. $\boldsymbol{\beta}_1$ is the coefficient vector for the selected explanatory variables in the model, $\mathbf{X}_{it}$ is the selected explanatory variables for i-th state and year t. $\boldsymbol{\beta}_2$ is the coefficient vector for the time-invariant explanatory variables. $\mathbf{Z}_i$ is the time-invariant explanatory variables which measures the unobserved time-invariant heterogeneities across the state. ε_{it} is the individual-specific random effect, which measures the deviation of the fatality rate for t-th year from the average fatality rate over all years the i-th state.

There are four assumptions in our model: 1. The error terms are normally distributed; 2.The error terms are in equal variance. 3.No Serial correlation: Serial correlation (also called Autocorrelation) is where error terms in a time series transfer from one period to another. In other words, the error for one time period a is correlated with the error for a subsequent time period b; 4. No cross-sectional dependence: the problem of cross-sectional dependence arises if the n individuals in our sample are no longer independently drawn observations but affect each other's outcomes.

3. Results

3.1 Descriptive Analysis

In our project, we use the fatality ratio(Number of fatalities in one state divided by the population of this state) as the response variable rather than the number of vehicle fatalities. We use this as a response variable because the raise of fatalities may be influenced by an increase in this state's population.

In this section, we analyze the changes in the fatality rate due to traffic over time in various US states. We also examine some factors that may cause changes in fatality rate.

Table 3.3.1: Cumulative population density distribution table in 1985

State	ca	ny	tx	pa	il	fl	oh	...
Population as a percentage of total population (unit: %)	11.1	7.5	6.9	5.0	4.8	4.8	4.5	...
Cumulative population as a percentage of total population (unit: %)	11.1	18.6	25.5	30.5	35.4	40.2	44.8	...

("ca" means "California", "fl" means "Florida", "il" means "Illinois", "ny" means "new york", "pa" means "Pennsylvania" and "tx" means "Texas")

First, we reduce the number of states to consider in this project. According to the Table 3.3.1: Cumulative population density distribution table in 1985 , selecting the six most populous states in the United States will allow us to consider forty percent of the US population. This approach maximizes the reliability of the analysis without having to examine 48 states in the descriptive statistical analysis. For each state, we considered the time period from 1982 to 1988. The reason why the 1985 population was chosen is because the population has grown uniformly during these seven years, so I think it is more representative to choose the median value.

After basically screening, we think beer tax, income, spirits consumption, unemployment rate, and drink age may have an impact on traffic mortality.

The first state we examined was Texas. In the picture above, the relevant data for Texas is represented by the purple lines. From Figure 3.3.1 (a), we observe that the traffic death rate in Texas decreases year to year for the entire time period. In Figure 3.3.1 (b), we see that the Texas beer tax has not significantly changed in 7 years, and we can determine that the traffic mortality has no obvious relationship with beer tax. For Figure 3.3.1 (c), This income level change is small and flattening. We can also speculate that the decline in traffic mortality in Texas may be associated with a non-increasing population income. From Figure 3.3.1 (d) we can see that the rate of decline is much faster than the other states. As a result, we can form the conjecture: The decline of spirits cosumption may be a likely cause of the decline in Texas' traffic fatalities. In Figure 3.3.1 (e), Texas' unemployment rate changed drastically compared to the other states. So, we believe that this fluctuation in unemployment may lead to a decline in Texas' traffic fatalities. For Figure 3.3.1 (f), we can see that Texas' first legal drinking age was 19 years old and, in 1984, the age began to increase until it reached 21 in 1987. During this time period, the Texas traffic death rate decreased significantly with the increase in the legal drinking age(strong correlation).

Since their trends and data are similar, we analyze New York, Pennsylvania, Illinois, and California at the same time. New York is represented by the light blue line, Pennsylvania is represented by the blue line, Illinois is represented by the green line, and California is represented by the red line. In Figure 3.3.1 (a), the four states have roughly the same trend, but it is worth noting that California's traffic fatality rate is higher than the other states. Illinois' traffic fatality rate increased from 1987 to 1988. California 's traffic death rate increased in 1984. In Figure 3.3.1 (b), it can be seen that there was little to no change in the beer taxes in these four states. In Figure 3.3.1 (c), the per capita income of these four states increased at a similar rate. For Figure 3.3.1 (d), the spirit consumption of these four states decreased at a similar rate, but it is important to note that New York's baseline of spirits consumption has the slowest rate of decline. Compared with

other states, this difference in spirits consumption may be the reason why New York State 's traffic mortality has increased the most compared to the other three states. In Figure 3.3.1 (e), Pennsylvania and Illinois' unemployment rate decreased in the seven years. During this same time period, Pennsylvania's traffic death rates increased, which means there may be a strong correlation between these occurences. It is worth noting that, in both states, the unemployment rate increased in 1983 and their corresponding traffic death rates decreased, which would indicate a negative correlation between traffic mortality and unemployment. Between 1982 and 1988, New York's unemployment rate declined. It should be mentioned that when the decline in the unemployment rate slowed down(e.g. 1983), the traffic death rate in New York began to increase. We found that this trend also occurred in California. In Figure 3.3.1 (f), New York started raising its legal drinking age in 1985 and, in 1986, it reached 21. Between 1985 and 1986, the traffic death rate in New York increased. We also found that the remaining three states did not change their legal drinking age during these years.

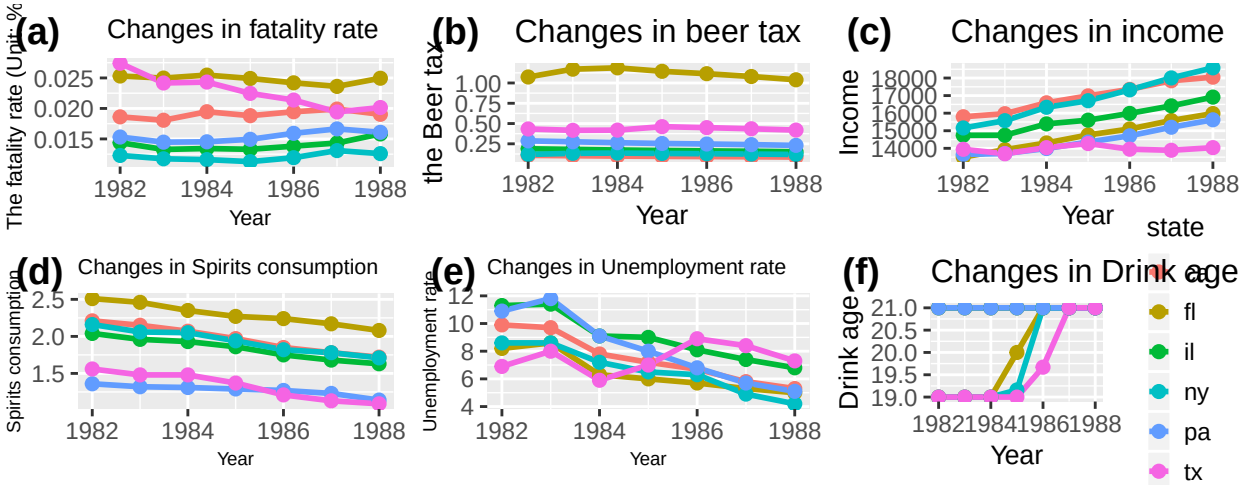

Figure 3.3.1: Changes in variables with different years.("ca" means "California", "fl" means "Florida", "il" means "Illinois", "ny" means "new york", "pa" means "Pennsylvania" and "tx" means "Texas")

According to the Figure 3.3.1 (a) and (f), Florida is marked by earthy yellow, and it can be seen that after the legal drinking age increased from 1984 to 1986, Florida's traffic death rate fell rapidly. So there is a strong negative correlation between legal drinking age and traffic mortality in Florida.

3.2 Main Results

Just like the reasons stated in the descriptive statistical analysis, we use the fatality ratio as the response variable in the main analysis.

First, we use the standardization transformation to eliminate the impact of different units and to make all variables comparable in magnitude.

`Table 3.2.1 Twoways effects Within Model within jailyes`

	Estimate	Std.Error	t-value	Pr(>|t|)	
spirits	0.967	0.136	7.129	8.857e-12	***
unemp	-0.244	0.046	-5.304	2.325e-07	***
income	0.342	0.080	4.303	0.004	***
beertax	-0.376	0.130	-2.89	0.004	**
jailyes	0.077	0.099	0.777	0.438	

In our project, we use the additive model because the interactions between the factors are not meaningful. Furthermore, the interaction term is difficult to interpret and explain to the policymakers. The additive model is concise and makes it easier to provide suggestions to policymakers. We will use AIC as the criterion

to select a model by randomly adding a variable or randomly deleting a variable. The upper variables in this model contain all variables and the lower variables in this model must contain state and year because we need to control the influence of time and the entities. With no effect of time and entities, we can determine which of the other variables are truly related to the response variable. It is reasonable to use this method to find which variables are correlated with the response variable because the fixed-effect model, with time fixed, can be transformed into a general linear regression model. Using this method, the selected model contains unemployment rate, per capita personal income, spirits consumption, tax on a case of beer, dry, miles, state and year as variables. Because the coefficients of dry and miles are not significant at all, we drop these two variables from the model. Since we need to discuss whether having a mandatory jail sentence is associated with reduced traffic fatalities, we add the jail variable into our model. The results are displayed on the table 3.2.1.

The jail coefficient is not significant, which indicates that having a mandatory jail sentence will not increase or decrease the traffic fatalities. However, in our casual inference, we concluded that having a mandatory jail sentence is associated with reduced traffic fatalities and the existence of a confounder makes the jail coefficient not significant.

3.3 Model Diagnostics

3.3.1 Diagnostics for Mixed Effect Models

Table 3.3.1.1: The results of diagnostics for mixed effect models

Diagnostics for fixed effect model			df2	p-value
F test for time fixed effect	F = 8.529	df1 = 6	276	1.646e-08
Huasman test with time fixed effect	chisq = 13.174	df = 5		0.022
Hausman test without time fixed effect	chisq = 241.63	df = 5		< 2.2e-16
F test for basic OLS model	F = 57.184	df1 = 53	276	< 2.2e-16

First, we need to test if the regression model needs the fixed time effect. The result of the F test (null hypothesis is that no time fixed effects are needed) given in the Table 3.3.1.1 (line 1) indicates we need to add time effects in our model. Next, we need to determine whether the random-effects should be involved in the model. First, we construct two random-effects models. The result of Hausman test (null hypothesis is that the preferred model is a random-effects model) given in the Table 3.3.1.1 (line 2 and 3) shows that the fixed effects model is preferred. Finally, we need to determine whether basic OLS model is preferred than fixed effect model. The result of F test for Individual and/or Time effects (the null hypothesis is no significant fixed effect) given by the Table 3.3.1.1 (line 4) shows that fixed effects model is preferred.

3.3.2 Diagnostics for assumptions

Table 3.3.2.1: The resulfs of diagnostics for assumptions

Diagnostics for assumption			
Lilliefors (Kolmogorov-Smirnov) normality test	D = 0.046		p-value = 0.091
Breusch-Pagan test	BP = 125.26	df = 58	p-value = 7.487e-07
Breusch-Godfrey/Wooldridge test	chisq = 23.398	df = 6	p-value = 0.000674
Pesaran CD test	z = -1.69		p-value = 0.0919

(a) Normality test: The Lilliefors (Kolmogorov-Smirnov) normality test given in the Table 3.3.2.1 (line 1) shows that the residuals in the fixed effect model is normal. (b) Equal variances: The result of Breusch-Pagan test (null hypothesis is there is no heteroskedasticity in model) given in Table 3.3.2.1 (line 2) indicates that there is heteroskedasticity in our model. (c) No Serial correlation: We use Breusch-Godfrey test for the panel model to test if there is serial correlation in our model. According to

Table 3.3.2.1: line 3, this indicates that there is serial correlation in our model. (d) No cross-sectional dependence: We will use Pesaran's CD test to test if there is cross-sectional dependence in our model. According to Table 3.3.2.1: line 4, the result shows that there is no cross-sectional dependence in our model. In order to check the multicolinearity between the model variables, we calculated the VIF factors between the various variables, [1.511, 1.278, 2.133, 1.259]. Their VIFs are all less than 10, which basically determines that there is less multicolinearity between the variables. Since there is serial correlation and heteroskedasticity in our model, we should use an adjustment method to correct our coefficients in our model. This method is Heteroskedasticity-consistent estimation of the covariance matrix for the coefficient estimates in regression models. This method can correct our standard error of each coefficient. The coefficients of the original model and after correction model are shown in Table 3.3.2.2. We can see that the standard errors of each coefficient are bigger than the standard errors from the original model. This indicates that some variables may become non-significant. For example, the beertax variable is significant under 0.1 level but not significant under 0.05 level.

`Table 3.3.2.2: The result of the original model and after correction model`

	Estimate	Std. Error	t value	Pr(>|t|)	Estimate	Std. Error	t value	Pr(>|t|)
spirits	0.967	0.136	7.129	8.857e-12	0.967	0.138	7.020	1.719e-11
unemp	-0.244	0.046	-5.304	2.325e-07	-0.244	0.052	-4.729	3.610e-06
income	0.343	0.080	4.303	2.338e-05	0.343	0.125	2.746	0.006
beertax	-0.376	0.130	-2.89	0.004	-0.376	0.200	-1.876	0.004
jailyes	0.077	0.099	0.777	0.438	0.077	0.163	0.471	0.638

(The left is original model and the right is after correction model)

3.3.3 Possibility of Making Causal Statements

To make causal statement about whether having a mandatory jail sentence is associated with reduced traffic fatalities, we use propensity score methods to check the causality of having a mandatory jail sentence. The propensity for panel data requires following assumptions:

- The Stable Unit Treatment Value Assumption: This requires that the outcome of one subject is unaffected by the particular assignment of all other subjects. In our project, fatal rate for each state is independent with others. The relationship between all variables and youth border crossing can be roughly ignored as mentioned before. Besides, the law legislation in each state is independent with other states. Therefore, the fata rate of one state to jail is independent with the fatal rate of other state.
- The numbering of units is done at random, so that the index i contains no information. In this project, there is no specific criteria for determining the order of observations for all states. The observed information is mostly involved in other observed variables, for example the spirits consumption, beertax and other variables.
- The treatment is binary: In our project, the treatment jail only has two levels (yes and no).
- The inclusion of all significant covariates. When we use propensity score methods, we need to include all covariates in the final model that are related to both the treatment assignment and potential outcomes. In our project, the final model contains four covariate(spirits, unemployment, income and beertax). Only unemployment and income are significant to both fatal rate and jail in our model. Therefore, we choose only unemployment and income as our covariates in propensity score methods.
- Positivity Assumption: all subjects in the analysis have some probability of receiving the treatment. In our project, we can compare the distribution of propensity scores after executing a matching alogrithm in the group receiving mandatory jail sentence to the distribution of scores in the control group. These distributions shown in Figure 3.3.3.1 are similar. Therefore, it is appropriate to use propensity score method.

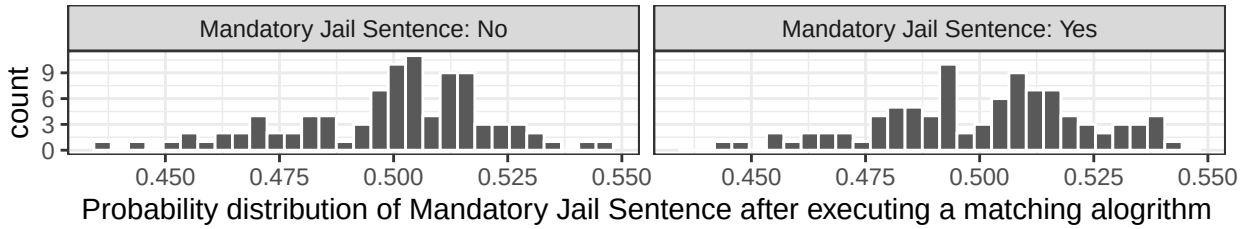

Figure 3.3.3.1 The distribution of propensity scores after executing a matching alogrithm in the group receiving mandatory jail sentence and control group.

Based on the previous analysis, we can make causal statement about whether having a mandatory jail sentence is associated with reduced traffic fatalities. The result given in Table 3.3.3.1 shows that the mandatory jail sentence can cause the increase in fatal rate.

Welch two Sample t-test	t = -5.2853	df=170.03	p-value = 3.809e-07
95 percent confidence interval	-0.850		-0.388
sample estimates:	mean in group 0 = -0.173		mean in group 1 = 0.446

Table 3.3.3.1 The result of t-test for true difference of fatal rate means in jail sentence group and non jail sentence group is not equal to 0.

4. Conclusion and Discussion

Based on the previous analysis, following conclusion can be drawn: at significance level = 0.1, four variables (spirit consumption, unemployment, income and beertax) in our model are all significant to fatal rate. (a) In particular, given other variables fixed, the fatal rate will increase if the spirit consumption increases. This makes sense because if people spend more money on alcohol, then the possibility of drunk driving will increase. (b) Given other variables fixed, the fatal rate will decrease if the unemployment increases. It is because if the unemployment is relatively high, then the purchasing ability of the resident will be at relative low level. More people can not buy personal vehicles and so the fatal rate would be relatively lower. (c) Given other variables fixed, the fatal rate will increase if the income increases. This also makes sense because if the income for residents increases, more people can buy the personal vehicles and the possibility of using personal vehicles would increase. More vehicles in the road, fatal rate would be potentially higher. (d) Given other variables fixed, the fatal rate will decrease if the beertax increases. The increasement of beertax would decrease the possibility of alcohol consuming, which further decrease the possibility of drunk driving and then decrease the fatal rate.

In general, following suggestions could be make to the policymakers. (a) Make appropriate legislation to limit the spirits consumption and alcohol using. This can reduce the possibility of drunk driving in the state. (b) Increase the beertax appropriately. This method can also decrease the alcohol consumption. (c) Another feasible way to control the fatal rate is to improve the awareness of the residents who have high income. (d) Besides, further study conducted by other researchers shows that increase the legal drinkage could also reduce the vehicle fatal rate of resident in particular age group. [1][2]

5. Appendix

5.1 Session Information

```
print(sessionInfo(), local = FALSE)
```

```
## R version 3.6.2 (2019-12-12)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 17763)
##
## Matrix products: default
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] ggdag_0.2.2    dagitty_0.2-2  dplyr_0.8.3    MatchIt_3.0.2  ggpubr_0.2.5
##  [6] magrittr_1.5   ggplot2_3.2.1  nortest_1.0-4  aTSA_3.1.2     MASS_7.3-51.5
## [11] plm_2.2-0      AER_1.2-9      survival_3.1-8 sandwich_2.5-1 lmtest_0.9-37
## [16] zoo_1.8-7      car_3.0-6      carData_3.0-3
##
## loaded via a namespace (and not attached):
##  [1] viridis_0.5.1      tidyr_1.0.2        viridisLite_0.3.0  tidygraph_1.1.2
##  [5] jsonlite_1.6.1     splines_3.6.2      ggraph_2.0.1       Formula_1.2-3
##  [9] Rdpack_0.11-1      assertthat_0.2.1   cellranger_1.1.0   ggrepel_0.8.1
## [13] yaml_2.2.1         pillar_1.4.3       lattice_0.20-38    glue_1.3.1
## [17] digest_0.6.23      polyclip_1.10-0    ggsignif_0.6.0     colorspace_1.4-1
## [21] cowplot_1.0.0      htmltools_0.4.0    Matrix_1.2-18      pkgconfig_2.0.3
## [25] bibtex_0.4.2.2     haven_2.2.0        purrr_0.3.3        scales_1.1.0
## [29] tweenr_1.0.1       openxlsx_4.1.4     rio_0.5.16         ggforce_0.3.1
## [33] tibble_2.1.3       farver_2.0.3       withr_2.1.2        maxLik_1.3-8
## [37] lazyeval_0.2.2     cli_2.0.1          crayon_1.3.4       readxl_1.3.1
## [41] evaluate_0.14      fansi_0.4.1        nlme_3.1-142       forcats_0.4.0
## [45] foreign_0.8-75     tools_3.6.2        data.table_1.12.8  hms_0.5.3
## [49] gbRd_0.4-11        lifecycle_0.1.0    stringr_1.4.0      V8_3.0.1
## [53] munsell_0.5.0      zip_2.0.4          compiler_3.6.2     rlang_0.4.4
## [57] grid_3.6.2         miscTools_0.6-26   igraph_1.2.4.2     labeling_0.3
## [61] rmarkdown_2.1      boot_1.3-24        gtable_0.3.0       abind_1.4-5
## [65] curl_4.2           graphlayouts_0.5.0 R6_2.4.1           gridExtra_2.3
## [69] knitr_1.28         bdsmatrix_1.3-4    utf8_1.1.4         stringi_1.4.4
## [73] Rcpp_1.0.3         vctrs_0.2.2        tidyselect_1.0.0   xfun_0.12
```

5.2 Reference

[1.]Alcohol Policies and Highway Vehicle Fatalities By: Christopher J. Ruhm Ruhm, C. (1996). Alcohol Policies and Highway Vehicle Fatalities. Journal of Health Economics 15(4): 435-454.

[2.]Alcohol policies and highway vehicle fatalities a,b, Christopher J. Ruhm a Department of Economies, University of North Carolina Greensboro, Greensboro, NC 27412-500l, USA b National Bureau of Economic Research, Cambridge. MA, USA Received 1 July 1995; revised 1 January 1996

[3.]Stock J.H., Watson M.W. Introduction to Econometrics (2ed., AW, 2006) (text book)

[4.]The central role of the propensity score in observational studies for causal effects BY PAUL R. ROSENBAUM Departments of Statistics and Human Oncology, University of Wisconsin, Madison, Wisconsin, U.S.A. AND DONALD B. RUBIN University of Chicago, Chicago, Illinois, U.S.A

[5.]An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies

[6.]The Journal of Infectious Diseases EDITORIAL COMMENTARY • JID 2019:219 (1 January) • 1 The Journal of Infectious Diseases® 2019;219:1–2 Causal Inference for Observational Studies David Kaplan Case Western Reserve University, Cleveland, Ohio

[7.]The Stata Journal (2007) 7, Number 4, pp. 507–541 Causal inference with observational data Austin Nichols Urban Institute Washington, DC austinnichols@gmail.com

5.3 Resources

[1.] https://rstudio-pubs-static.s3.amazonaws.com/372492_3e05f38dd3f248e89cdedd317d603b9a.html#43_random_effects_model. (Getting Started in Fixed/Random Effects Models using R)

[2.] https://sejdemyr.github.io/r-tutorials/statistics/tutorial8.html#estimating-treatment-effects. (R Tutorial 8: Propensity Score Matching)

[3.] https://www.schmidheiny.name/teaching/panel2up.pdf (Short Guides to Microeconometrics, Panel Data: Fixed and Random Effects)

5.4 Github information

https://github.com/BillXu999/STA207_project03