Diff-on-Diff implementation in R

Introduction

This document walks through the Difference-in-Difference analysis of thr miniimum legal drinking age example as it is presented in Angrist and Pischke's Mastering Metrics book. The code here is R code which is based on the excellent work by Jeffrey Arnold.

This document assumes that you have access to Chapter 5 of Angist and Pischke's book. In here we will replicate most of the results presented in that chapter relating to the investigation of whether changes in the minimum legal driving age (MLDA) impact numbers of traffic fatalities in young people.

Look at this wikipedia page to get an overview of the (useful) patchwork of MLDA legislation in the US.

MLDA Difference-in-Difference

Load necessary libraries.

```
library(tidyverse)
library(stargazer)
library(readxl)
library(ggplot2)
library(clubSandwich)
```

Setup your working directory and load the data which are saved in deaths.Rdata.

```
setwd("YOUR_WORKING_DIRECTORY")

load("deaths.Rdata")
```

The data contain the number of deaths due to various reasons, reported by US State and Year.

```
deaths %>% filter(state == "5", year == "1980", agegr == "18-20 yrs") %>%
  select(state,year,agegr, count,dtype, pop) %>%  print()
```

```
## # A tibble: 6 x 6
##   state  year agegr     count dtype            pop
##   <fct> <dbl> <fct>     <dbl> <fct>          <dbl>
## 1 5      1980 18-20 yrs   173 all           125804
## 2 5      1980 18-20 yrs    73 MVA           125804
## 3 5      1980 18-20 yrs    16 suicide       125804
## 4 5      1980 18-20 yrs    17 homicide      125804
## 5 5      1980 18-20 yrs    40 other external 125804
## 6 5      1980 18-20 yrs    27 internal      125804
```

So, for instance, in State 5 (and we don't know which state this is) we can see that in 1980 altogether 173 18-20 year olds passed away, 73 of which due to a Motor Vehicle accident (MVA).

A few important variables are

- `count`, this is the count of Deaths for the indicated year, state and reason for death

- `mrate`, this is the mortatlity rate (per 100,000) for the indicated year, state and reason for death
- `year`, gives the year for which the data are
- `state`, a numeric variable which indicates which US state the data are from
- `dtype`, stands for death type
- `agegr`, this is the age group
- `pop`, this is the population, in the respective age group.
- `legal`, described in detail below

Identifying States

In the dataset states are represented by numbers. This is a little annoying. So let's add the state names. But which number belongs to which state. It is very typical that countries and states have particular numbers associated to them to assist in identifying states. For instance Utah should be represented by the number 49. But how do you know that it is these number are called FIPS (Federal Information Processing Standard) codes?

If you look at a [https://www.census.gov/geographies/reference-files/2017/demo/popest/2017-fips.html] listing of these codes (and how the match to state names) you will find a couple of peclarities. In general the numbers go from 1 (Alabama) to 56 (Wyoming), but a few numbers are missing, in particular 3, 7, 12, 43 and 52 are not allocated to states. So if we were to look at our state numbers and these numbers are also missing then we can be pretty certain that we are looking at FIPS codes and then we can use the information in the above link to match our numbers to state names.

```
unique(deaths$state)
```

```
##  [1] 1  2  4  5  6  8  9  10 11 12 13 15 16 17 18 19 20 21 22 23 24 25 26 27 28
## [26] 29 30 31 32 33 34 35 36 37 38 39 40 41 42 44 45 46 47 48 49 50 51 53 54 55
## [51] 56
## 51 Levels: 1 2 4 5 6 8 9 10 11 12 13 15 16 17 18 19 20 21 22 23 24 25 26 ... 56
```

Success!!! We have numbers from 1 to 56 and none of the above listed numbers appears. We have FIPS codes. In some sense this is not 100% proof and you could check whether population numbers which we have match to reported population numbers in 1980, but we can trust that the authors have done the sensible thing and used FIPS codes.

Now we need to import the spreadsheet which lists these codes and the associated state abbreviations (2 letter codes) and state names. As it is a little tedious to find an easily accessible file which includes all these information I have prepared `states.xlsx` for your convenience.

```
state<-read_excel("states.xlsx")
state$FIPS <- as.factor(state$FIPS)
names(state)
```

```
## [1] "Cstate" "NAME"   "FIPS"
```

This leaves us with a Table which contains state number (`FIPS`), state name (`NAME`) and a two letter code (`Cstate`).

Now we use `state` in `deaths` and `FIPS` in `state` to merge the State appreviation (`Cstate`) and their names (`NAME`) into our dataset.

```
deaths <- merge(deaths,state,by.x = "state", by.y = "FIPS",x.all = TRUE)
```

This has added the variables `Cstate` and `NAME` to our dataframe.

Data Summaries

Let's see what the data look like

```
str(deaths)
```

```
## 'data.frame':    24786 obs. of  17 variables:
##  $ state       : Factor w/ 51 levels "1","2","4","5",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ year        : num  1970 1971 1972 1973 1974 ...
##  $ legal1820   : num  0 0 0 0 0 ...
##  $ dtype       : Factor w/ 6 levels "all","MVA","suicide",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ agegr       : Factor w/ 3 levels "15-17 yrs","18-20 yrs",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ count       : num  224 241 270 258 224 207 231 219 234 176 ...
##  $ pop         : num  213574 220026 224877 227256 229025 ...
##  $ age         : num  16 16 16 16 16 ...
##  $ legal       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ beertaxa    : num  1.37 1.32 1.28 1.2 1.08 ...
##  $ beerpercap  : num  0.6 0.66 0.74 0.79 0.83 ...
##  $ winepercap  : num  0.09 0.09 0.09 0.1 0.16 ...
##  $ spiritpercap: num  0.7 0.76 0.78 0.79 0.81 ...
##  $ totpercap   : num  1.38 1.52 1.61 1.69 1.8 ...
##  $ mrate       : num  104.9 109.5 120.1 113.5 97.8 ...
##  $ Cstate      : chr  "AL" "AL" "AL" "AL" ...
##  $ NAME        : chr  "Alabama" "Alabama" "Alabama" "Alabama" ...
```

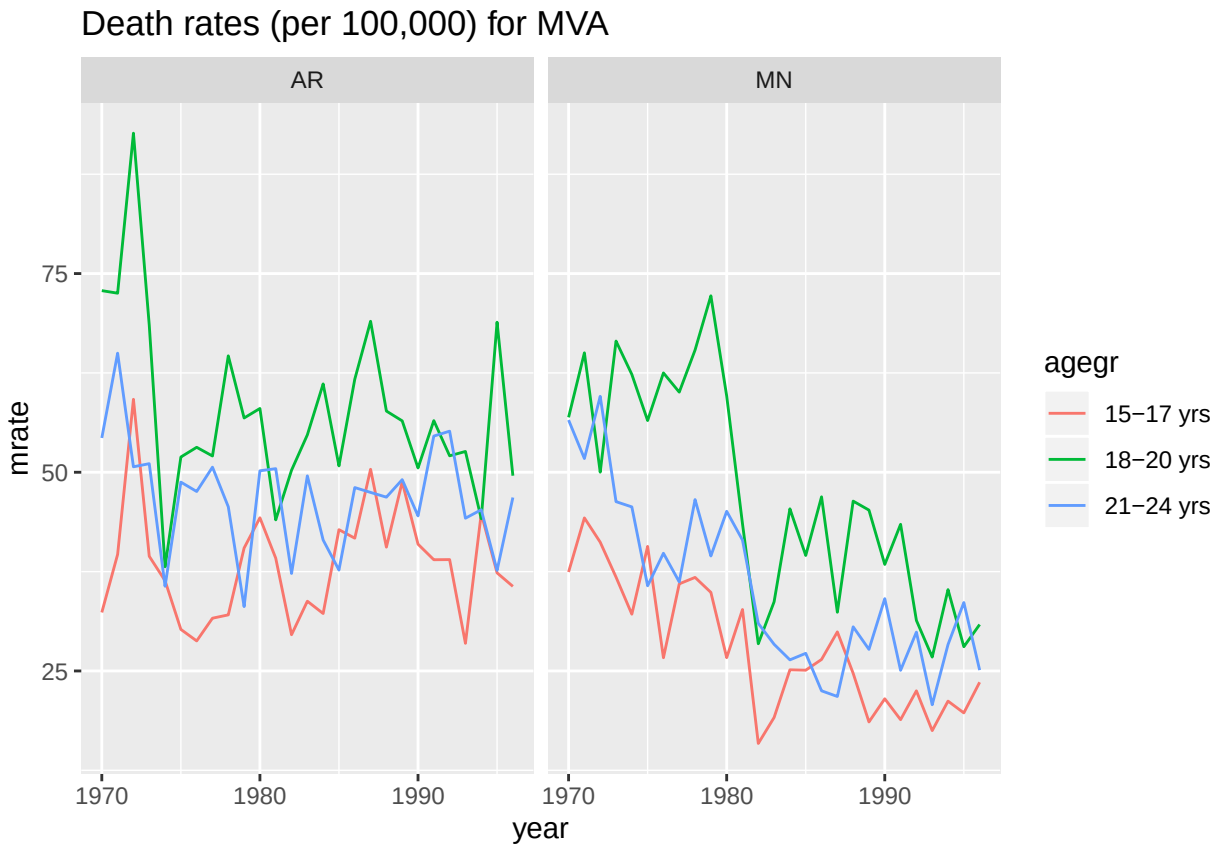
```
summary(deaths)
```

```
##      state            year        legal1820                 dtype
##  1      :  486   Min.   :1970   Min.   :0.0000   all           :4131
##  2      :  486   1st Qu.:1976   1st Qu.:0.0000   MVA           :4131
##  4      :  486   Median :1983   Median :0.0000   suicide       :4131
##  5      :  486   Mean   :1983   Mean   :0.3399   homicide      :4131
##  6      :  486   3rd Qu.:1990   3rd Qu.:0.7001   other external:4131
##  8      :  486   Max.   :1996   Max.   :1.0000   internal      :4131
##  (Other):21870
##        agegr          count              pop               age
##  15-17 yrs:8262   Min.   :   0.00   Min.   :  15023   Min.   :15.95
##  18-20 yrs:8262   1st Qu.:  14.00   1st Qu.:  65106   1st Qu.:16.02
##  21-24 yrs:8262   Median :  38.00   Median : 174899   Median :19.01
##                   Mean   :  91.42   Mean   : 255117   Mean   :19.17
##                   3rd Qu.:  96.00   3rd Qu.: 309315   3rd Qu.:22.47
##                   Max.   :2639.00   Max.   :2023616   Max.   :22.79
##
##      legal           beertaxa         beerpercap      winepercap
##  Min.   :0.0000   Min.   :0.01275   Min.   :0.600   Min.   :0.0600
##  1st Qu.:0.0000   1st Qu.:0.09102   1st Qu.:1.150   1st Qu.:0.1700
##  Median :0.0000   Median :0.15524   Median :1.310   Median :0.2700
##  Mean   :0.4466   Mean   :0.23138   Mean   :1.319   Mean   :0.3046
##  3rd Qu.:1.0000   3rd Qu.:0.28169   3rd Qu.:1.440   3rd Qu.:0.4000
##  Max.   :1.0000   Max.   :1.98454   Max.   :2.280   Max.   :1.1100
##                   NA's   :288
##   spiritpercap      totpercap        mrate           Cstate
##  Min.   :0.3700   Min.   :1.21   Min.   :  0.00   Length:24786
##  1st Qu.:0.7000   1st Qu.:2.12   1st Qu.: 12.07   Class :character
##  Median :0.8700   Median :2.49   Median : 21.35   Mode  :character
```

```
##  Mean   :0.9964   Mean   :2.62   Mean   : 37.33
##  3rd Qu.:1.1200   3rd Qu.:2.89   3rd Qu.: 43.73
##  Max.   :4.4500   Max.   :6.92   Max.   :623.02
##
##      NAME
##  Length:24786
##  Class :character
##  Mode  :character
##
##
##
##
```

Let's look at the data for the state of Minnesota (MN) and Arkansas (AR) and compare the time-series of death rates from motor vehicle accidents (`dtype == "MVA"`) for the three age groups.

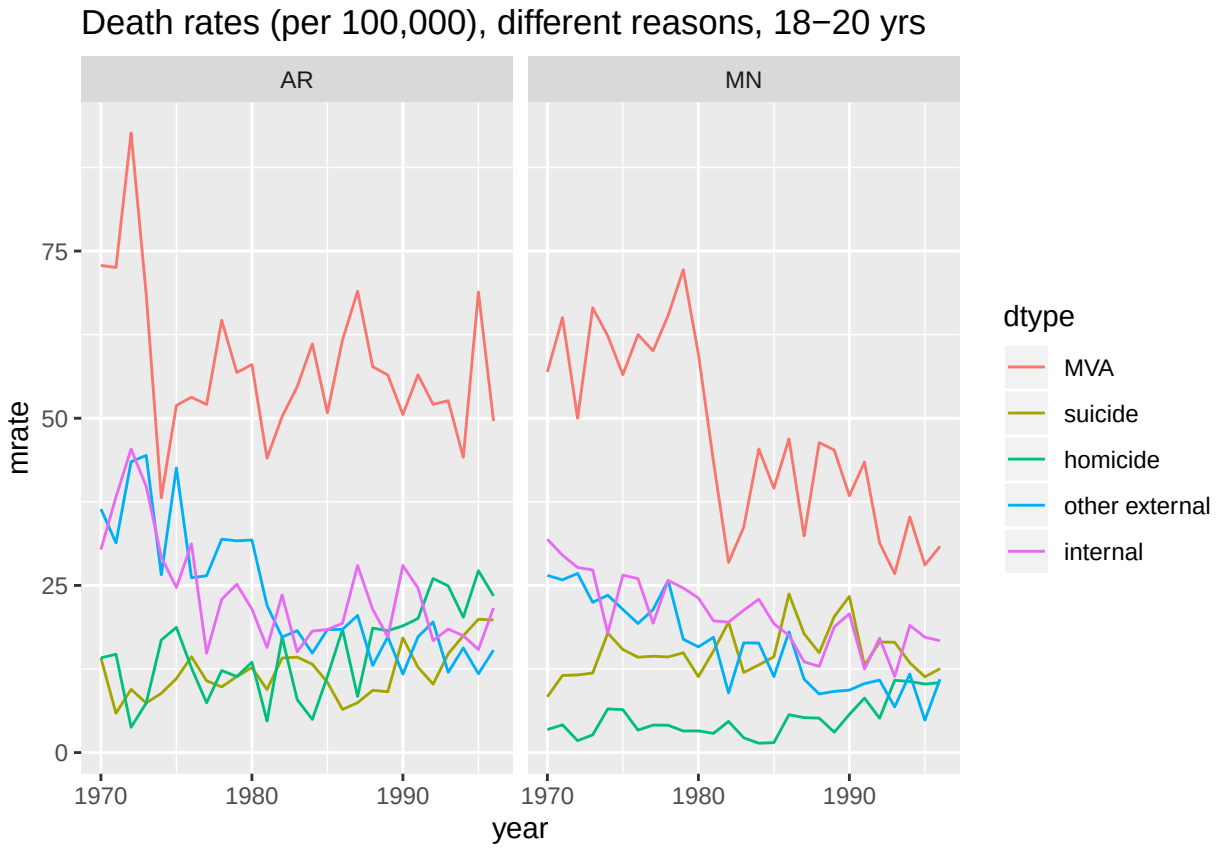
```
deaths_sel <- deaths %>%  filter(Cstate %in% c("MN","AR"), dtype == "MVA")

g1 <- ggplot(deaths_sel, aes(y=mrate, x=year, color = agegr)) +
  geom_line() +
  ggtitle("Death rates (per 100,000) for MVA") +
  facet_grid(.~Cstate)
g1
```



Further look at the death rates for the 18-20 year group for different causes of death (excluding "all").

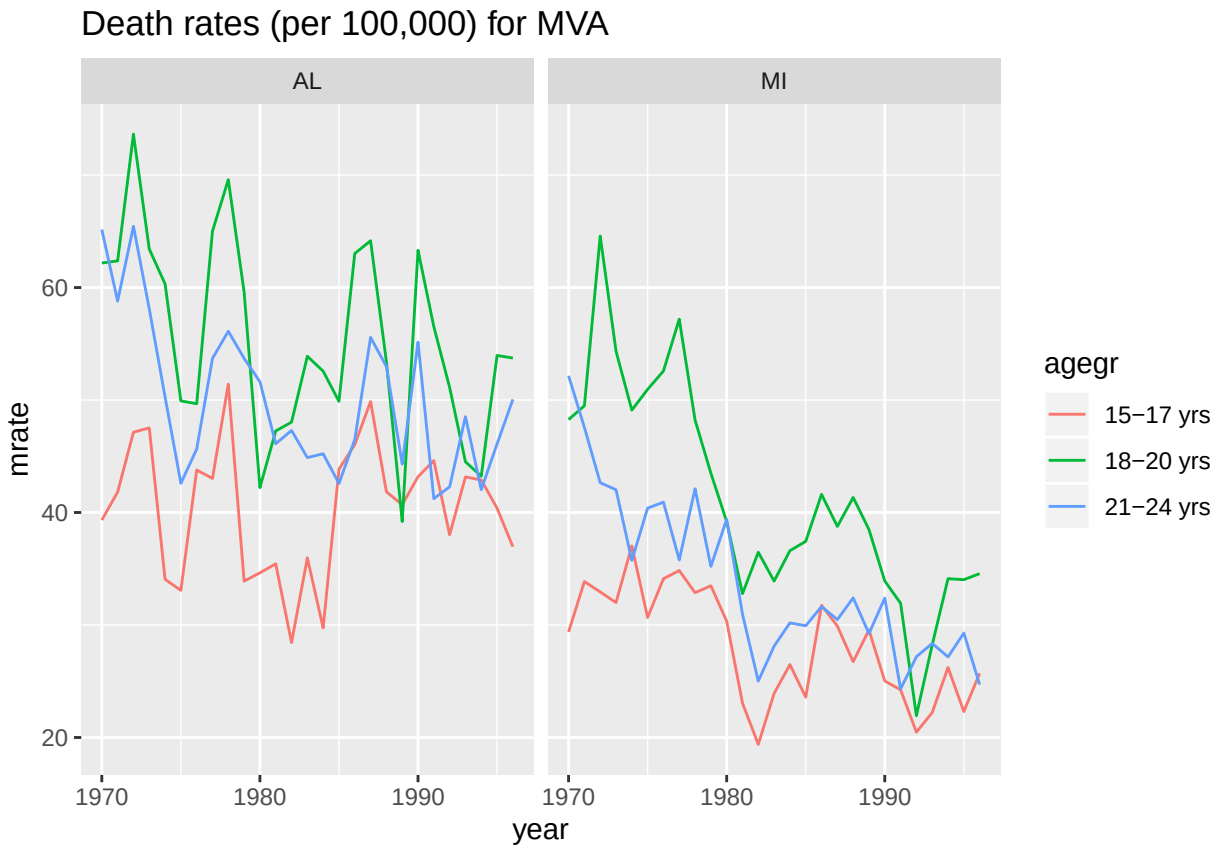
```
deaths_sel <- deaths %>%
  filter(Cstate %in% c("AR","MN"), agegr == "18-20 yrs") %>%
```

```
  filter(!(dtype == "all"))

g2 <- ggplot(deaths_sel, aes(y=mrate, x=year, color = dtype)) +
  geom_line() +
  ggtitle("Death rates (per 100,000), different reasons, 18-20 yrs") +
  facet_grid(.~Cstate)
g2
```



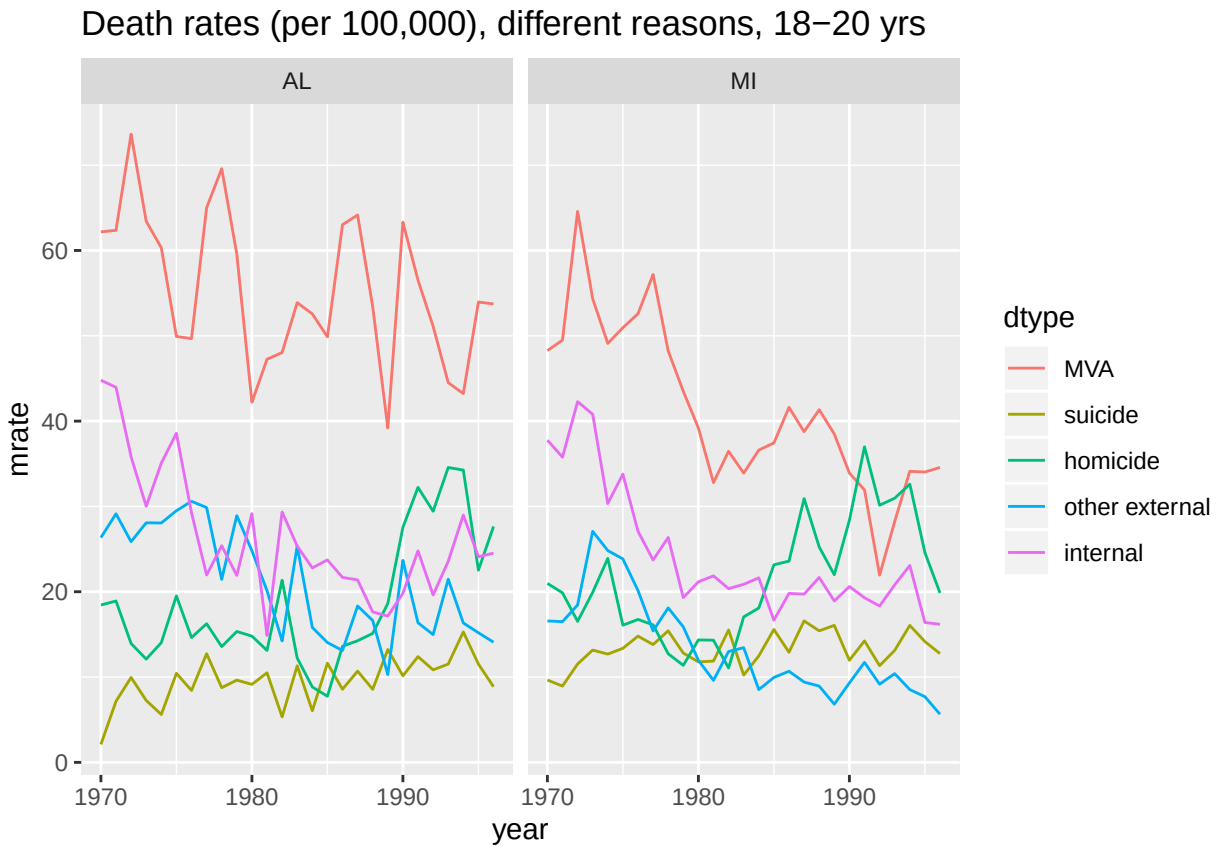
The same graphs for Alabama (AL) and Michigan (MI) are here:

```
deaths_sel <- deaths %>%  filter(Cstate %in% c("AL","MI"), dtype == "MVA")

g1 <- ggplot(deaths_sel, aes(y=mrate, x=year, color = agegr)) +
  geom_line() +
  ggtitle("Death rates (per 100,000) for MVA") +
  facet_grid(.~Cstate)
g1
```



Further look at the death rates for the 18-20 year group for different causes of death (excluding "all").

```
deaths_sel <- deaths %>%
  filter(Cstate %in% c("AL","MI"), agegr == "18-20 yrs") %>%
  filter(!(dtype == "all"))

g2 <- ggplot(deaths_sel, aes(y=mrate, x=year, color = dtype)) +
  geom_line() +
  ggtitle("Death rates (per 100,000), different reasons, 18-20 yrs") +
  facet_grid(.~Cstate)
g2
```



Diff-in-Diff, Two States

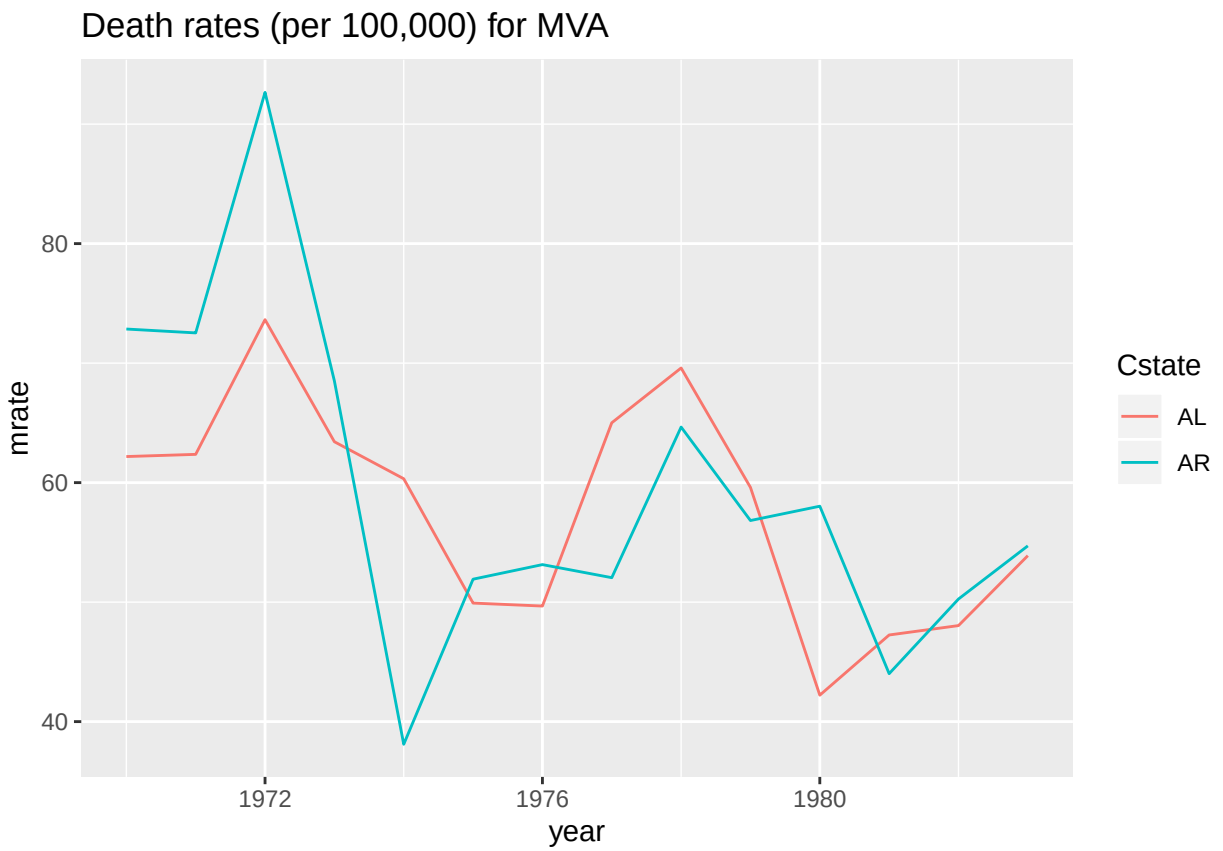
Let us look at two states only for now. Alabama (AL) and Arkansas (AR). Throughout the sample period Arkansas had a minimum legel drinking age (MLDA) of 21. Alabama, however, reduced the MLDA, in 1975, to 19 (from 21) and only in 1985 increased it back to 21.

Hence we could look at the impact of the reduction of the MLDA in Alabama, using Arkansas as a control state. Let us select the data for these two states, up to 1983 and for the motor vehicle accidents (`dtype = "MVA"`).

```
AP_data_1 <- deaths %>% filter(Cstate %in% c("AL","AR")) %>%
  filter(year <= 1983, agegr == "18-20 yrs", dtype == "MVA") %>%
  arrange(Cstate,year)
```

This leaves us with 14 observation for each of the two states (1970 to 1983). Let's illustrate thes data graphically:

```
gdd <- ggplot(AP_data_1, aes(y=mrate, x=year, color = Cstate)) +
  geom_line() +
  ggtitle("Death rates (per 100,000) for MVA")
gdd
```



Looking at this picture it is not so obvious whether there has been a (causal) effect of the reduction of the MLDA in Alabama on death rates through motor vehicle accidents for 18-20 year olds. But just graphical analysis does not always unveil the effects, which is why we will often look at the problem through a regression lense.

The way in which we would use this dataset to estimate a simple Diff-in-Diff is as follows.

The general specification is:

$Y_{st} = \alpha + \beta TREAT_s + \gamma POST_t + \delta(TREAT_s \times POST_t) + e_{st}$

we need to specify the variables

- Y_{st},
- $TREAT_S$, and
- $POST_t$

Y_{st} is the `mrate` in state s at time t. $TREAT_s$ is a dummy variable which takes a value 1 (or `TRUE`) if we have an observation from Alabama, the state in which the policy change (lowering of MLDA) happened, and 0 (or `FALSE`) otherwise. $POST_t$ is another dummy variable which is 1 (or `TRUE`) for periods affected by the policy change, i.e. 1975 and onwards, and 0 (or `FALSE`) otherwise.

```
AP_data_1 <- AP_data_1 %>%
              mutate(treat = (Cstate == "AL"),
                     post = (year >= 1975),
                     treatpost = treat*post)
```

Have a look at the variables to confirm that `treatpost` only takes the value 1 for AL from 1975 onwards.

It is the estimator to δ which is indicative of any impact the policy has on teh outcome variable (assuming

the parallel trend assumption can be justified).

```
dd_2state <- lm(mrate~treat+post+treatpost, data = AP_data_1)
stargazer(dd_2state, type = "text")
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                  mrate
## ------------------------------------------------
## treat                           -4.546
##                                 (6.540)
##
## post                           -14.978**
##                                 (5.767)
##
## treatpost                       4.500
##                                 (8.156)
##
## Constant                       68.932***
##                                 (4.624)
##
## ------------------------------------------------
## Observations                      28
## R2                               0.299
## Adjusted R2                      0.211
## Residual Std. Error       10.340 (df = 24)
## F Statistic             3.408** (df = 3; 24)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

While $\hat{\delta}$ is positive, it is not statiystically significant and, on the basis of these data only, we would not attribute any adverse consequences to the lowering of the MLDA.

However, we only looked at two states, and disregarding information coming from other states is unwise.

Diff-in-Diff, Many States, variable timing

The question now arises how we would cater for multiple states when implementing a Diff-in-Diff approach using a regression.

Two complications will need to be taken into account.

1) Different states may implement policies at different times. For instance, Wyoming lowered the MLDA to 19 as did Alabama, but did so two years earlier, in 1973. This raises the question how the $POST_t$ variable shoudl be defined.
2) Some states may have implemented the policy in a slightly different way. As an example, Michigan lowered the MLDA in 1972 and it lowered the MLDA to 18 (rather than 19), therefore potentially affecting a larger proportion of 18-20 Year olds. This raises a question about how the $TREAT_s$ variable should be defined.

The solution here is to define a new variable, here $LEGAL_{st}$ which takes the role of $(TREAT_s \times POST_t)$. This variable takes the value 0 if 0 % of a specific age group (we are interested in the 18-20 year old group) in state s is allowed to drink legally in a particular year t, say when the MLDA is 21. It takes the value 1 if 100% of the respective age group is allowed to legally drink alcohol, say when the MLDA is 18. It will take

values between 0 and 1 to reflect that only a portion of the 18-20 year olds may be legally allowed to drink alcohol (say, if the MLDA is 19) and/or if they were only allowed to drink alcohol legally for a part of the respective year, say if the law changed half way through the year.

This variable will be at the core and the coefficient estimate relating to this variable will deliver an estimate of the causal effect (assuming the parallel trends assumption can be justified).

What happens to the $TREAT_s$ and $POST_t$ variables? They will typically be replaced by state- and time fixed-effects respectively. These two variables controled for differences between the states unrelated to the policy implementation ($TREAT_s$) and the differences across time which applied universally to all states ($POST_t$). The fixed effects will do the same thing, merely allowing for more than two states and more than two time periods.

In our example we have data for 51 states ($K = 51$) and for 14 years ($T = 14$). We would include 50 state dummy variables (treating one state as the base case, here Alaska, AK) and 13 year dummy variables (typically treating the first year as the base case, here 1970). In order to do this we require the state variable (say the two letter appreviation variable `Cstate`) and a year variable as factor variables as it is very easy to add such dummy variables for variables of the `factor` type.

```
deaths <- deaths %>% mutate(year_fct = factor(year),
                            Cstate = factor(Cstate))
```

We now subset the data as in Angrist and Pischke. Use all 51 states, only look at the "18-20 yrs" age group, for years up to and including 1983 and for starters look at (`dtype == "all"`).

```
AP_data <- deaths %>% filter(year <= 1983, agegr == "18-20 yrs", dtype == "all") %>%
  arrange(Cstate,year)
```

The model is now represented as:

$Y = \alpha + \Sigma_{k=AL}^{WY}\beta_k STATE^k + \Sigma_{\tau=1971}^{1983}\gamma_\tau YEAR^\tau + \delta LEGAL + e$

The index for the $STATE^K$ variable runs alphabetically from "AL" (Alabama) to "WY" (Wyoming). The state first in the alphabet , "AK" (Alaska), is the base case Observations are state-years normally indexed by *st* but to avoid a too laden notation they are not showing. To illustrate what the dummy variables do let's look at the following table which shows a few observations and their respective values for selected variables.

CState	Year	mrate	`LEGAL`	...	$Year^{1975}$	$Year^{1976}$	...	$STATE^{AL}$	$STATE^{AR}$	...
AL	1974	147.9	0.0	...	0	0	...	1	0	...
AL	1975	147.9	0.294	...	1	0	...	1	0	...
AL	1976	132.6	0.665	...	0	1	...	1	0	...
...	...	...	...	...	...	...	...	...	...	...
AR	1976	137.5	0.0	...	0	1	...	0	1	...
AR	1977	111.5	0.0	...	0	0	...	0	1	...

Fortunately we do not have to create all the year and state dummy variables. R will do this internally.

The crucial variabke which indicates the "strength" of the policy (reduction of minimum legal drinking age) is the variable `LEGAL`. It expresses the proportion of 18-20 year olfd which are allowed to legally dring alcohol. YOu can see in the table that this propotion was 0 in Arkansas in 1976 and 1977 (as the MLDA was 21). In Alabama the story is different. In 1974 it was also 0 as the MLDA was still 21. In 1976 the proportion was 0.665 or about 67% as the MLDA had been reduced to 19 (2/3s of 18-20 year olds were allowed to legally drink). The value for 1975 (LEGAL = 0.294) is a combination of the fact that 2/3s of 18-20 year olds were allowed to legally drink, but that legislation was in effect for only about 5 months (0.294 is approximately 0.665 * (5/12)).

The above model is then implemented in R as follows:

```
mod1<- lm(mrate ~ legal + state + year_fct, data = AP_data)
```

Yes that is it! `state` (we also could have used `Cstate`) and `year_fct` are factor variables. R therefore recognises that these are categorical variables and the appropriate way to include these is by internally creating the $STATE^k$ and $YEAR^\tau$ dummy variables.

With this under our belt we can tackle recreating the values in Angrist and Pischke's tables 5.2 and 5.3.

Table 5.2

This first regression will estimate the above Diff-in-Diff specification for all causes of death in 18-20 year olds. We use the `lm` function to estimate the regression.

```
mod1 <- lm(mrate ~ legal + state + year_fct, data = AP_data)
stargazer(mod1, keep = "legal", type="text") # keep = "legal" only reports legal
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                mrate
## ------------------------------------------------
## legal                        10.804***
##                               (3.138)
##
## ------------------------------------------------
## Observations                   714
## R2                            0.821
## Adjusted R2                   0.804
## Residual Std. Error      17.339 (df = 649)
## F Statistic          46.561*** (df = 64; 649)
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

This is the first entry in Table 5.2 (column (1) "All deaths"). By using the `keep = "legal"` option in the call to the `stargazer` function we ensure that only the coefficient we are really interested in is reported.

Let's implement the additional state time-trends (as for column (2)). We obtain these by adding `state:year` as "a" regressor. Here we use the `year` variable (not the `year_fct` variable). This is important as, when we want a time trend, we need R to recognise that `year` is a numeric variable (as it is). Adding `state:year` actually adds 50 (51 minus one base state) time trends.

```
mod2 <- lm(mrate ~ legal + year_fct + state + state:year, data = AP_data)
stargazer(mod1, mod2, dep.var.caption = "All Deaths", keep = "legal", type="text")
```

```
##
## ===================================================================
##                                    All Deaths
##                     ----------------------------------------------
##                                      mrate
##                            (1)                     (2)
## -------------------------------------------------------------------
## legal                   10.804***               8.467**
##                          (3.138)                 (3.724)
##
## -------------------------------------------------------------------
```

```
## Observations                  714                       714
## R2                           0.821                     0.848
## Adjusted R2                  0.804                     0.820
## Residual Std. Error    17.339 (df = 649)         16.616 (df = 599)
## F Statistic         46.561*** (df = 64; 649) 29.412*** (df = 114; 599)
## ============================================================================
## Note:                                          *p<0.1; **p<0.05; ***p<0.01
```

The standard errors which are calculated in this routine are the usual standard errors which are not corrected for heteroskedaticity of any kind. And in fact, when you compare these to those in Angrist and Pischke, you can see that they are notably smaller. In situations like this it is commonly appropriate to calculate cluster-robust standard errors, i.e. standard errors which recognise that there may be clusters of data which exhibit different error variances. The natural cluster variable here is the state variable. In other words we want to allow for error variances which differ across states. To achieve this we use the `vcovCL` function from the `sandwich` package. This will be fed into the function `coeftest` from the `lmtest` package.

```
library(sandwich)
library(lmtest)
mod1_cr <- coeftest(mod1,vcovCL(mod1, cluster = ~ state))

# Create vectors with cluster robust standard errors
# for use in stargazer below
mod1_cr_se <- sqrt(diag(vcovCL(mod1, cluster = ~ state)))
mod2_cr_se <- sqrt(diag(vcovCL(mod2, cluster = ~ state)))
```

In `mod1_ct` you now find all the (unchanged) coefficient estimates, the new standard errors and as a result new t-statistics and p-values for the null hypothesis that the respective population coefficients are 0, for `mod1`. Let's just see the first few rows which also contain the `legal` variable.

```
mod1_cr[1:4,]
```

```
##              Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) 153.66915  1.9813614 77.557356  0.000000e+00
## legal        10.80414  4.5922045  2.352713  1.893458e-02
## state2      107.88660  1.1828382 91.209937  0.000000e+00
## state4       37.80343  0.6717837 56.273219 7.753281e-252
```

You can see that the new standard error to `legal` is 4.59 as in Table 5.2 in Angrist and Pischke. And in fact you can work this new standard error into the 'stargazer display

```
# keep = "legal" only reports legal
stargazer(mod1, keep = "legal", type="text", se=list(mod1_cr_se))
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                mrate
## ------------------------------------------------
## legal                         10.804**
##                               (4.592)
##
## ------------------------------------------------
## Observations                    714
## R2                             0.821
## Adjusted R2                    0.804
## Residual Std. Error      17.339 (df = 649)
```

```
## F Statistic          46.561*** (df = 64; 649)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

or for models `mod1` and `mod2`

```
stargazer(mod1,mod2, keep = "legal", type="text",
          se=list(mod1_cr_se,mod2_cr_se)) # keep = "legal" only reports legal
```

```
##
## ===============================================================================
##                                      Dependent variable:
##                     ------------------------------------------------------
##                                             mrate
##                                (1)                          (2)
## -------------------------------------------------------------------------------
## legal                        10.804**                     8.467*
##                              (4.592)                      (5.098)
##
## -------------------------------------------------------------------------------
## Observations                   714                          714
## R2                            0.821                        0.848
## Adjusted R2                   0.804                        0.820
## Residual Std. Error    17.339 (df = 649)          16.616 (df = 599)
## F Statistic         46.561*** (df = 64; 649) 29.412*** (df = 114; 599)
## ===============================================================================
## Note:                                          *p<0.1; **p<0.05; ***p<0.01
```

Now we repeat this analysis for the different reasons of Death, always reporting the cluster robust standard errors:

```
# MVAs
AP_data <- filter(deaths, year <= 1983, agegr == "18-20 yrs", dtype == "MVA")
mod1 <- lm(mrate ~ legal + state + year_fct, data = AP_data)
mod1_cr_se <- sqrt(diag(vcovCL(mod1, cluster = ~ state)))

mod2 <- lm(mrate ~ legal + year_fct + state + state:year, data = AP_data)
mod2_cr_se <- sqrt(diag(vcovCL(mod2, cluster = ~ state)))

stargazer(mod1, mod2, dep.var.caption = "MVA", keep = "legal",
          type="text",se=list(mod1_cr_se,mod2_cr_se))
```

```
##
## ===============================================================================
##                                            MVA
##                     ------------------------------------------------------
##                                             mrate
##                                (1)                          (2)
## -------------------------------------------------------------------------------
## legal                        7.592***                     6.644**
##                              (2.496)                      (2.656)
##
## -------------------------------------------------------------------------------
## Observations                   714                          714
## R2                            0.798                        0.836
## Adjusted R2                   0.778                        0.805
```

```
## Residual Std. Error    10.960 (df = 649)         10.269 (df = 599)
## F Statistic         39.976*** (df = 64; 649) 26.797*** (df = 114; 599)
## ==============================================================================
## Note:                                          *p<0.1; **p<0.05; ***p<0.01
```

```
# Suicides
AP_data <- filter(deaths, year <= 1983, agegr == "18-20 yrs", dtype == "suicide")
mod1 <- lm(mrate ~ legal + state + year_fct, data = AP_data)
mod1_cr_se <- sqrt(diag(vcovCL(mod1, cluster = ~ state)))

mod2 <- lm(mrate ~ legal + year_fct + state + state:year, data = AP_data)
mod2_cr_se <- sqrt(diag(vcovCL(mod2, cluster = ~ state)))

stargazer(mod1, mod2, dep.var.caption = "Suicide", keep = "legal",
          type="text",se=list(mod1_cr_se,mod2_cr_se))
```

```
##
## ==============================================================================
##                                          Suicide
##                      ---------------------------------------------
##                                           mrate
##                                (1)                      (2)
## ------------------------------------------------------------------------------
## legal                         0.591                    0.474
##                              (0.590)                  (0.795)
##
## ------------------------------------------------------------------------------
## Observations                   714                      714
## R2                            0.608                    0.637
## Adjusted R2                   0.570                    0.568
## Residual Std. Error     4.372 (df = 649)         4.381 (df = 599)
## F Statistic         15.747*** (df = 64; 649) 9.218*** (df = 114; 599)
## ==============================================================================
## Note:                                          *p<0.1; **p<0.05; ***p<0.01
```

```
# Internal
AP_data <- filter(deaths, year <= 1983, agegr == "18-20 yrs", dtype == "internal")
mod1 <- lm(mrate ~ legal + state + year_fct, data = AP_data)
mod1_cr_se <- sqrt(diag(vcovCL(mod1, cluster = ~ state)))

mod2 <- lm(mrate ~ legal + year_fct + state + state:year, data = AP_data)
mod2_cr_se <- sqrt(diag(vcovCL(mod2, cluster = ~ state)))

stargazer(mod1, mod2, dep.var.caption = "Internal", keep = "legal",
          type="text",se=list(mod1_cr_se,mod2_cr_se))
```

```
##
## ==============================================================================
##                                          Internal
##                      ---------------------------------------------
##                                           mrate
##                                (1)                      (2)
## ------------------------------------------------------------------------------
## legal                         1.333                    0.079
##                              (1.586)                  (1.933)
##
```

```
## ------------------------------------------------------------------------
## Observations                  714                         714
## R2                           0.622                       0.681
## Adjusted R2                  0.585                       0.620
## Residual Std. Error     6.365 (df = 649)         6.088 (df = 599)
## F Statistic         16.680*** (df = 64; 649) 11.206*** (df = 114; 599)
## ========================================================================
## Note:                                        *p<0.1; **p<0.05; ***p<0.01
```

Table 5.3

For the results in this table we add an additional control variable, namely the amount of beer taxes.

```
# All deaths
AP_data <- filter(deaths, year <= 1983, agegr == "18-20 yrs", dtype == "all")
mod1 <- lm(mrate ~ legal + beertaxa + state + year_fct, data = AP_data)
mod1_cr_se <- sqrt(diag(vcovCL(mod1, cluster = ~ state)))

mod2 <- lm(mrate ~ legal + beertaxa + year_fct + state + state:year, data = AP_data)
mod2_cr_se <- sqrt(diag(vcovCL(mod2, cluster = ~ state)))

stargazer(mod1, mod2, dep.var.caption = "MVA - incl. beertax",
          keep = c("legal","beertaxa"), type="text",
          se=list(mod1_cr_se,mod2_cr_se))
```

```
##
## ========================================================================
##                                    MVA - incl. beertax
##                        ---------------------------------------------
##                                           mrate
##                              (1)                        (2)
## ------------------------------------------------------------------------
## legal                      10.983**                   10.029**
##                            (4.691)                    (4.915)
##
## beertaxa                    1.505                     -5.525
##                            (9.071)                    (32.238)
##
## ------------------------------------------------------------------------
## Observations                  700                         700
## R2                           0.825                       0.850
## Adjusted R2                  0.807                       0.821
## Residual Std. Error     17.230 (df = 635)         16.570 (df = 586)
## F Statistic         46.645*** (df = 64; 635) 29.452*** (df = 113; 586)
## ========================================================================
## Note:                                        *p<0.1; **p<0.05; ***p<0.01
```

```
# MVA
AP_data <- filter(deaths, year <= 1983, agegr == "18-20 yrs", dtype == "MVA")
mod1 <- lm(mrate ~ legal + beertaxa + state + year_fct, data = AP_data)
mod1_cr_se <- sqrt(diag(vcovCL(mod1, cluster = ~ state)))

mod2 <- lm(mrate ~ legal + beertaxa + year_fct + state + state:year, data = AP_data)
mod2_cr_se <- sqrt(diag(vcovCL(mod2, cluster = ~ state)))
```

```
stargazer(mod1, mod2, dep.var.caption = "MVA - incl. beertax",
          keep = c("legal","beertaxa"), type="text",
          se=list(mod1_cr_se,mod2_cr_se))
```

```
##
## ================================================================
##                                 MVA - incl. beertax
##                       ---------------------------------------
##                                        mrate
##                                (1)                    (2)
## ----------------------------------------------------------------
## legal                        7.588***               6.888***
##                              (2.561)                (2.664)
##
## beertaxa                      3.819                  26.882
##                              (5.394)                (20.114)
##
## ----------------------------------------------------------------
## Observations                  700                    700
## R2                           0.797                  0.836
## Adjusted R2                  0.777                  0.804
## Residual Std. Error    10.953 (df = 635)      10.254 (df = 586)
## F Statistic         38.994*** (df = 64; 635) 26.425*** (df = 113; 586)
## ================================================================
## Note:                                        *p<0.1; **p<0.05; ***p<0.01
```

```
# Suicide
AP_data <- filter(deaths, year <= 1983, agegr == "18-20 yrs", dtype == "suicide")
mod1 <- lm(mrate ~ legal + beertaxa + state + year_fct, data = AP_data)
mod1_cr_se <- sqrt(diag(vcovCL(mod1, cluster = ~ state)))

mod2 <- lm(mrate ~ legal + beertaxa + year_fct + state + state:year, data = AP_data)
mod2_cr_se <- sqrt(diag(vcovCL(mod2, cluster = ~ state)))

stargazer(mod1, mod2, dep.var.caption = "Suicide",
          keep = c("legal","beertaxa"), type="text",
          se=list(mod1_cr_se,mod2_cr_se))
```

```
##
## ================================================================
##                                    Suicide
##                       ---------------------------------------
##                                        mrate
##                                (1)                    (2)
## ----------------------------------------------------------------
## legal                         0.448                  0.378
##                              (0.595)                (0.768)
##
## beertaxa                     -3.052*                -12.127
##                              (1.633)                (8.818)
##
## ----------------------------------------------------------------
## Observations                  700                    700
## R2                           0.619                  0.648
## Adjusted R2                  0.581                  0.580
```

```
## Residual Std. Error     4.314 (df = 635)         4.317 (df = 586)
## F Statistic         16.146*** (df = 64; 635) 9.559*** (df = 113; 586)
## ========================================================================
## Note:                                        *p<0.1; **p<0.05; ***p<0.01
```

```
# Internal
AP_data <- filter(deaths, year <= 1983, agegr == "18-20 yrs", dtype == "internal")
mod1 <- lm(mrate ~ legal + beertaxa + state + year_fct, data = AP_data)
mod1_cr_se <- sqrt(diag(vcovCL(mod1, cluster = ~ state)))

mod2 <- lm(mrate ~ legal + beertaxa + year_fct + state + state:year, data = AP_data)
mod2_cr_se <- sqrt(diag(vcovCL(mod2, cluster = ~ state)))

stargazer(mod1, mod2, dep.var.caption = "Internal",
          keep = c("legal","beertaxa"), type="text",
          se=list(mod1_cr_se,mod2_cr_se))
```

```
##
## ========================================================================
##                                        Internal
##                        ------------------------------------------------
##                                         mrate
##                                (1)                      (2)
## ------------------------------------------------------------------------
## legal                         1.465                    0.877
##                              (1.609)                  (1.808)
##
## beertaxa                      -1.357                   -10.309
##                              (3.069)                  (11.636)
##
## ------------------------------------------------------------------------
## Observations                   700                      700
## R2                            0.637                    0.692
## Adjusted R2                   0.600                    0.632
## Residual Std. Error     6.251 (df = 635)         5.995 (df = 586)
## F Statistic         17.382*** (df = 64; 635) 11.628*** (df = 113; 586)
## ========================================================================
## Note:                                        *p<0.1; **p<0.05; ***p<0.01
```

Summary

This document illustrated how to implement a Diff-in-Diff estimation when you are dealing with multiple groups which implement policies at different times. In this particular case we also allowed for implementations which differ, not only in timing, but also in terms of intensity.

As usual, the actual estimation was rather straightforward. The real work lies in the collection and handling of the data as well as in figuring out how to apply the straightword techniques.