Seatbelt Dataset

Econometric Project Group 17

UNIVERSITÀ DEGLI STUDI DI BERGAMO

STATA®

Index

Step 1
Introduction

Step 2
The Dataset

Step 3
Statistical and Econometric Methods

Step 4
Results

Step 5
Conclusion

Step 6
Appendix

1. Introduction

Research questions

01

Does the usage of seat belts reduce the fatality rate?

02

How the first and secondary enforcement influence the behaviours of the drivers like seat belts usage.

03

How the alcohol street regulations influence the speed limit and the fatality rate?

Summary of the main result

From our analysis, we can see that an increase in the use of seat belts involves a decrease in fatalities.

We also found that first and secondary enforcement increase the seat belts usage and in particular primary enforcement influences the use of seat belts more than the secondary enforcement.

Another important result concerns the fact that alcohol street regulation can't affect the speed limit which is instead related to the number of fatalities.

Plan of the Paper

- In the first part of the project, we are focusing on the variables and their description, showing by graphs their flow and their main properties.
- Secondly, we focused on the descriptions of the statistical models useful for the resolution of our analysis. Describing the theory behind the study and all the analysis we did later on in the paper.
- In the third part, we analyze the result of the statistical model we used to answer our question.
- In the end, we can easily calculate what we wanted to prove by answering our questions. So, computing the regressions, and showing numerically how and what influenced the dependent variables.
- In the Appendix, the reader can have the possibility to look and find the path where he/she can find the.do File

2. The Dataset

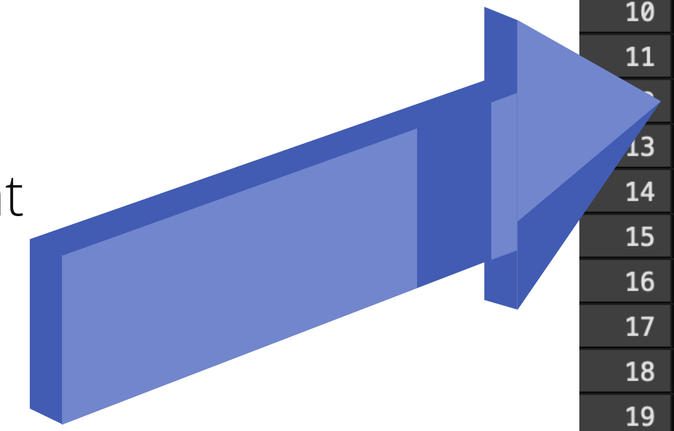
Sample

Seatbelts is a balanced panel from 50 U.S. States, plus the District of Columbia, for the years 1983-1997. These data were provided by Professor Liran Einav of Stanford University and were used in his paper with Alma Cohen “The Effects of Mandatory Seat Belt Laws on Driving Behavior and Traffic Fatalities,” The Review of Economics and Statistics, 2003, Vol. 85, No. 4, pp 828-843

Panel Data

The Seatbelts dataset is a Panel data dataset (also known as a longitudinal or cross-sectional time-series data). It contain observations about different cross sections across time.

As we can see our dataset is a Panel Data, due to different observations across time across different states.



	state	year	fips	vmt	fatalityrate
1	IL	1983	17	67370	.022651
2	MD	1983	24	30618	.0214253
3	GA	1983	13	48837	.0265373
4	WI	1983	55	34106	.0212573
5	NH	1983	33	7181	.026598
6	AR	1983	5	16684	.0333853
7	MT	1983	30	7181	.0398273
8	PA	1983	42	72302	.0238029
9	MS	1983	28	17802	.040164
10	WA	1983	53	36144	.0193116
11	WY	1983	56	5059	.0341965
[illegible]	KY	1983	21	26719	.0291179
[illegible]3	MN	1983	27	31063	.0178669
14	FL	1983	12	81776	.0328458
15	ME	1983	23	7924	.0282686
16	VT	1983	50	4151	.0226451
17	IN	1983	18	39837	.0255039
18	NJ	1983	34	52217	.0178486
19	WV	1983	54	11696	.0363372
20	CO	1983	8	24109	.026795

Description

This is the command in order to decribe in short our Panel Data.
One important thing we can easily see from this dataset is the fact that it has 51 states and for each state 15 periods of time.

```
. xtdescribe

    fips:  1, 2, ..., 56                                    n =        51
    year:  1983, 1984, ..., 1997                            T =        15
           Delta(year) = 1 unit
           Span(year)  = 15 periods
           (fips*year uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%      max
                        15      15      15        15        15      15       15

     Freq.  Percent    Cum.  |  Pattern
-----------------------------+---------------
        51    100.00  100.00  |  111111111111111
-----------------------------+---------------
        51    100.00          |  XXXXXXXXXXXXXXX
```

Brief description

The command: *description* can show different basic values of our dataset.
For instance, the storage type is the type of variable that is loaded into memory. To be more clear, the type *float* store a 4byte space. These values can be very relevant to the calculus of efficiency of a problem resolution.

```
. des

Contains data from /Users/ettorefalde/Documents/Ettore Falde/Magistrale/Corsi/Industrial Economics/Santos/Econometric Project/SeatBelts.dta
  obs:           765
 vars:            14                          21 Nov 2005 18:34
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
state           str2    %9s                 * State Initials
year            int     %8.0g                 Year of data collection
fips            byte    %9.0g                 State code
vmt             float   %9.0g                 Millions of traffic miles per year.
fatalityrate    float   %9.0g                 Number of fatalities per million of traffic miles
sb_useage       float   %9.0g                 Seat belt useage rate
speed65         float   %9.0g                 Dummy variable for 65 mile per hour speed limit
speed70         float   %9.0g                 Dummy variable for 70 or higher mile per hour speed limit
drinkage21      float   %9.0g                 Dummy variable for age 21 drinking age
ba08            float   %9.0g                 Binary variable for blood alcohol limit ≤ .08%
income          float   %9.0g                 Per capita income
age             float   %9.0g                 Mean age
primary         float   %9.0g                 Binary variable for primary enforcement of seat belt laws
secondary       float   %9.0g                 Binary variable for secondary enforcement of seat belt laws
                                            * indicated variables have notes
-------------------------------------------------------------------------------
```

Moreover, the *display format* just shows that we are justified on the right a general description of our float, byte and int values. For the *state* var, we right-justified it and then: "s" stays for: string.

The command: *des* is just a shortcut, in order to make more efficient code writing. To make a more efficient code we can use also the command: "*d*"

Stata View of our Dataset

- In this table, we can view all the variables and their principal data.
- The command used in Stata is: "su()" but this is just an abbreviation of the complete command: "summarize()"
- One main and encouraging value that we can see is the Mean of the variable sb_usage.

```
. su()

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       state |          0
        year |        765        1990     4.32332        1983       1997
        fips |        765    28.96078    15.68709           1         56
         vmt |        765    41447.73    43961.99        3099     285612
fatalityrate |        765    .0214895    .0061713    .0083273   .0454701
-------------+---------------------------------------------------------
   sb_useage |        556    .5288518    .1701859         .06        .87
     speed65 |        765    .6457516    .4785978           0          1
     speed70 |        765    .0705882    .2563034           0          1
   drinkage21 |        765    .8849673    .3192701           0          1
        ba08 |        765    .1163399    .3208418           0          1
-------------+---------------------------------------------------------
      income |        765    17992.59    4811.459        8372      35863
         age |        765    35.13719    1.698131    28.23497   39.16958
     primary |        765    .1215686    .3270007           0          1
   secondary |        765    .4954248    .5003062           0          1
```

We are aware that with the panel data is preferable to use the command *xtsum* but in our presentation *su()* allows us to have a general and simple view of the dataset.

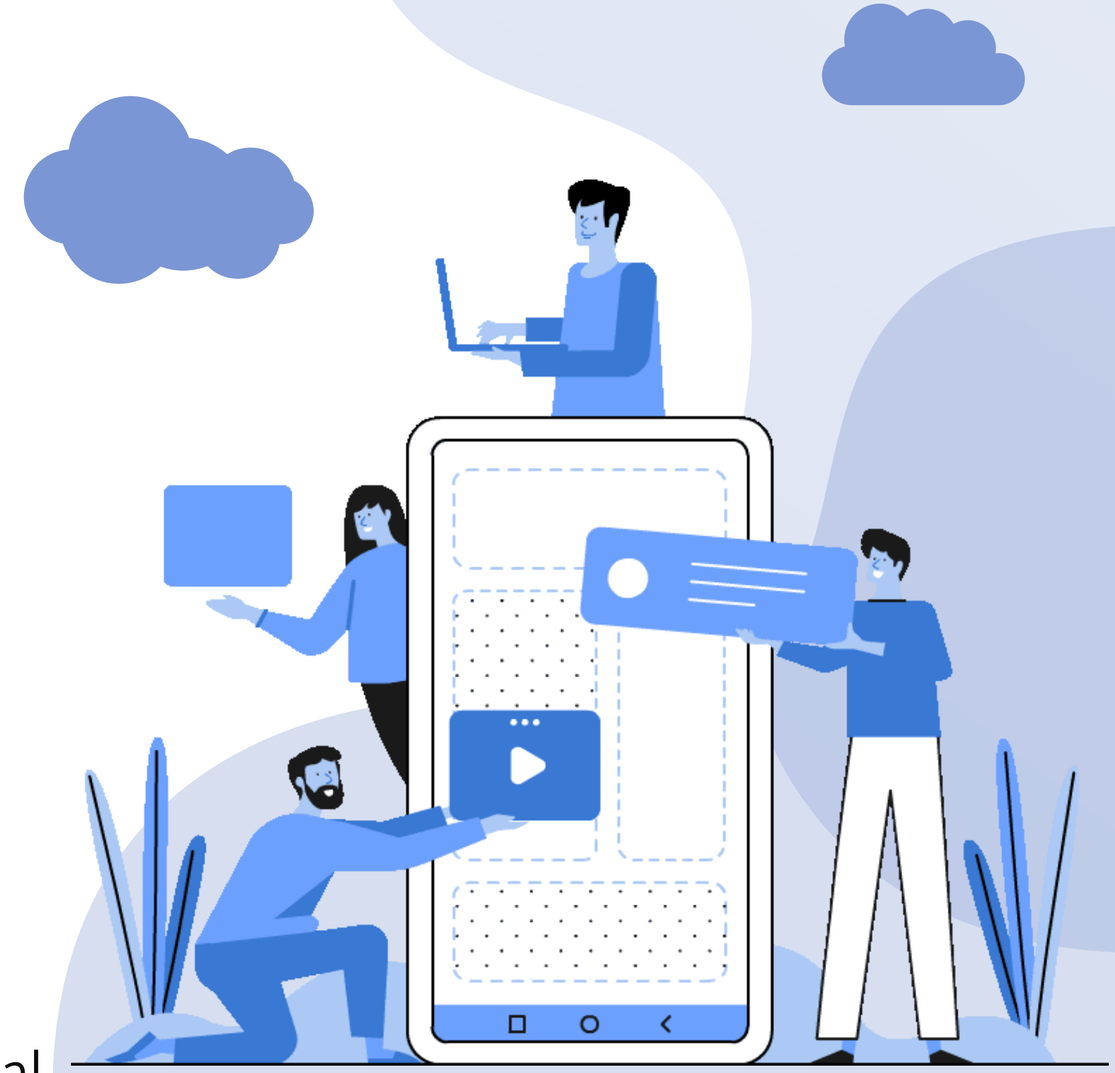
Variables

Variable	Definition
fatalityrate	Number of fatalities per million of traffic miles
sb_useage	Seat belt useage rate
speed65	Binary variable for 65 mile per hour speed limit
speed70	Binary variable for 70 or higher mile per hour speed limit
ba08	Binary variable for blood alcohol limit ≤ .08%
drinkage21	Binary variable for age 21 drinking age
income	Per capita income
age	Mean age
primary	Binary variable for primary enforcement of seat belt laws
secondary	Binary variable for secondary enforcement of seat belt laws
vmt	Millions of traffic miles per year.
state	State
year	Year
fips	State ID Code

Variables

There are two types of variables in SeatBelts dataset

- Nominal variables: variables that take on discrete values over time.
- Dummy variables: binary variable that allows you to incorporate qualitative information.
- Ordinal variables: It's a variable identified by groups that are ranked in a specific order. For instance, age that can be both ordinal a nominal.



Hypothesis of correlations

```
. correlate fatalityrate sb_useage speed65 speed70 drinkage21 ba08 income primary secondary age vmt year
(obs=556)
```

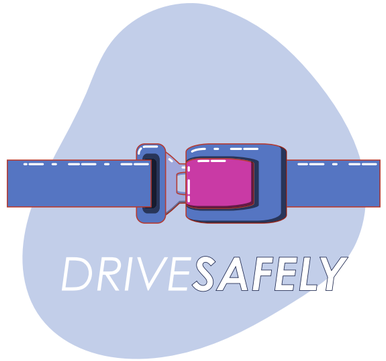
	fatali~e	sb_use~e	speed65	speed70	drink~21	ba08	income	primary	second~y	age	vmt	year
fatalityrate	1.0000											
sb_useage	-0.4027	1.0000										
speed65	-0.0802	0.3289	1.0000									
speed70	0.0006	0.1982	0.1684	1.0000								
drinkage21	-0.2391	0.3738	0.3859	0.0650	1.0000							
ba08	-0.1789	0.2286	0.1898	0.2261	0.0782	1.0000						
income	-0.6929	0.6223	0.1121	0.1534	0.3089	0.1193	1.0000					
primary	-0.0434	0.4209	-0.1769	-0.0321	0.0818	0.1056	0.1254	1.0000				
secondary	-0.1311	0.2651	0.3968	0.1709	0.2240	-0.0755	0.2400	-0.5488	1.0000			
age	-0.2730	0.1590	0.0159	-0.0351	0.1654	-0.0302	0.3622	0.0601	0.0103	1.0000		
vmt	-0.1113	0.2065	0.0028	0.1014	0.0534	0.1203	0.1581	0.0920	0.1185	0.0162	1.0000	
year	-0.4752	0.6792	0.4977	0.3861	0.3814	0.2774	0.6966	0.0512	0.3678	0.2798	0.0230	1.0000

Through the correlation command, we can easily see that the *fatalityrate* variable has a negative impact on all the other variables, except the *speed70* var.

Moreover, we can also observe that *sb_useage* has a positive correlation between all the other variables except for *fatalityrate.* In addition, all the other significant positive correlations are highlighted by the green square, like in the columns *income* and *speed65.*

Dependent variables

fatalityrate



sb_useage

The definition of a dependent variable is:

- The dependent variable is the variable being tested and measured in an experiment and is 'dependent' on the independent variable. For instance, *fatalityrate* is depending on the useage of seatbelts. Another example is the *sb_useage* which is depending on the government rules.

Fatalityrate

Number of fatalities per million traffic miles

1. We had to rename the label in order to organize better our work.
2. Using the command: "describe" (in short just *"d") we can have a description of what kind of variable was allocated in memory. In this case is a float with 9 decimal values.*

```
. summarize fatalityrate
```

Variable	Obs	Mean	Std. Dev.	Min	Max
fatalityrate	765	.0214895	.0061713	.0083273	.0454701

```
. d fatalityrate
```

variable name	storage type	display format	value label	variable label
fatalityrate	float	%9.0g		

We used the command summarize fatalityrate to show:

- Obs, which is the number of observation;
- Mean which is the average value;
- St.dev. which is the standard deviation;
- The maximum and minimum values assumed by the variables

To get a better understanding of the fatality rat we decided also to show its progression during time and its values for each state.

How we did that?

1. Set Stata in order to be sure that it is able to handle a panel data for our work.

```
.             xtset fips year
       panel variable:  fips (strongly balanced)
        time variable:  year, 1983 to 1997
                delta:  1 unit
```

NOTE:
Strongly balanced means that all the States (taken by their codes) have data for all the years of the examinations.
Moreover, if don't have the code a the states, you can easily correct this, by doing: *encode state, gen(fips)*

Fatalityrate across time and states

Clear view of the evolutions of fatality rate across time, across each State of the USA.

How we did that?

1. Command: *xtline fatalityrate*

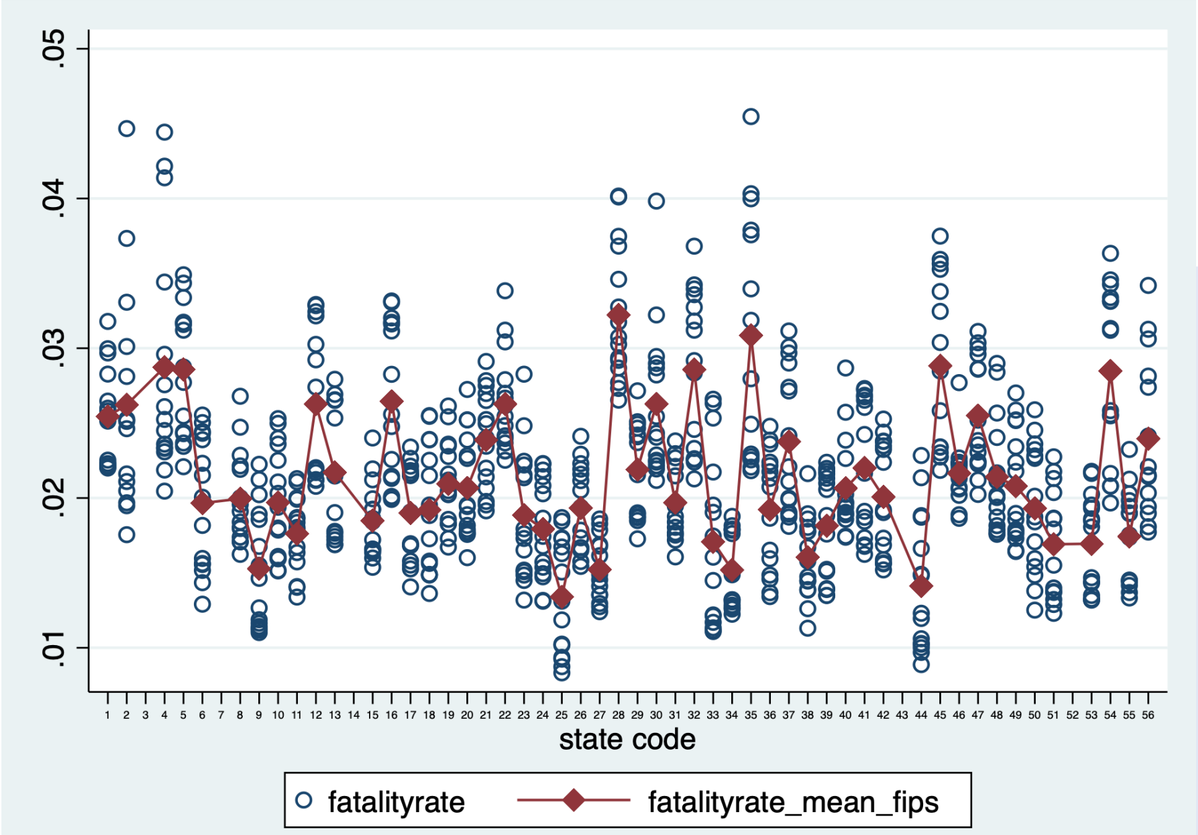
Fatality rate...a mean across states

In this graph, we can see the mean of *fatalityrate* across all the *state*s.
And it's pretty clear that in the graph the bigger values of *fatalityrate* are given by the state code 28, which is Mississippi.

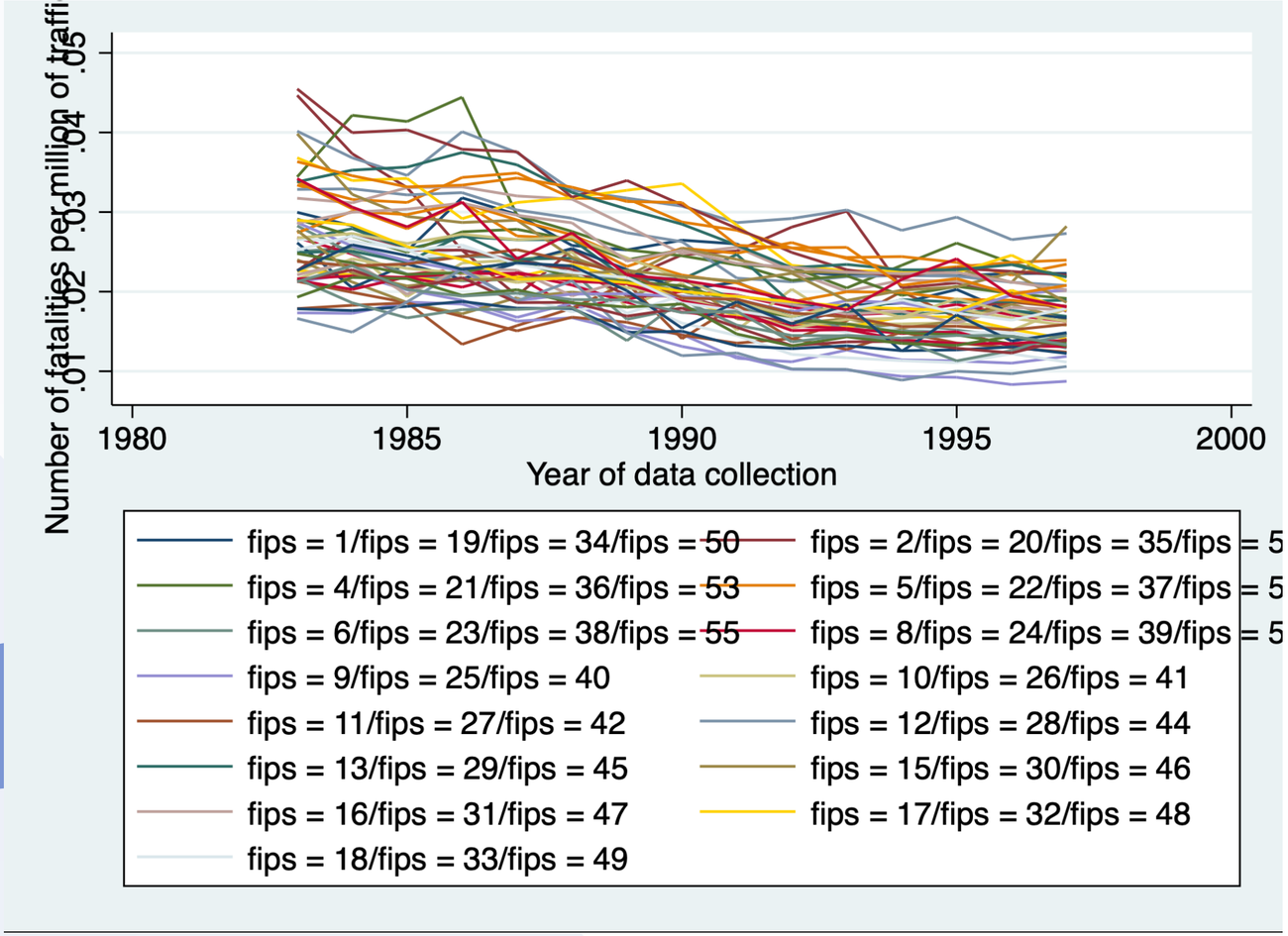
To obtain this we did the following commands:

1. *bysort fips: egen fatalityrate_mean_fips=mean(fatalityrate)*
2. *twoway scatter fatalityrate fips, msymbol(circle_hollow) || connected fatalityrate_mean_fips fips, msymbol(diamond)*

The 1. command creates a new var: *fatalityrate_mean_fips* that can allow us to show in the 2. command the mean values of that var, represented by the red diamond dots.



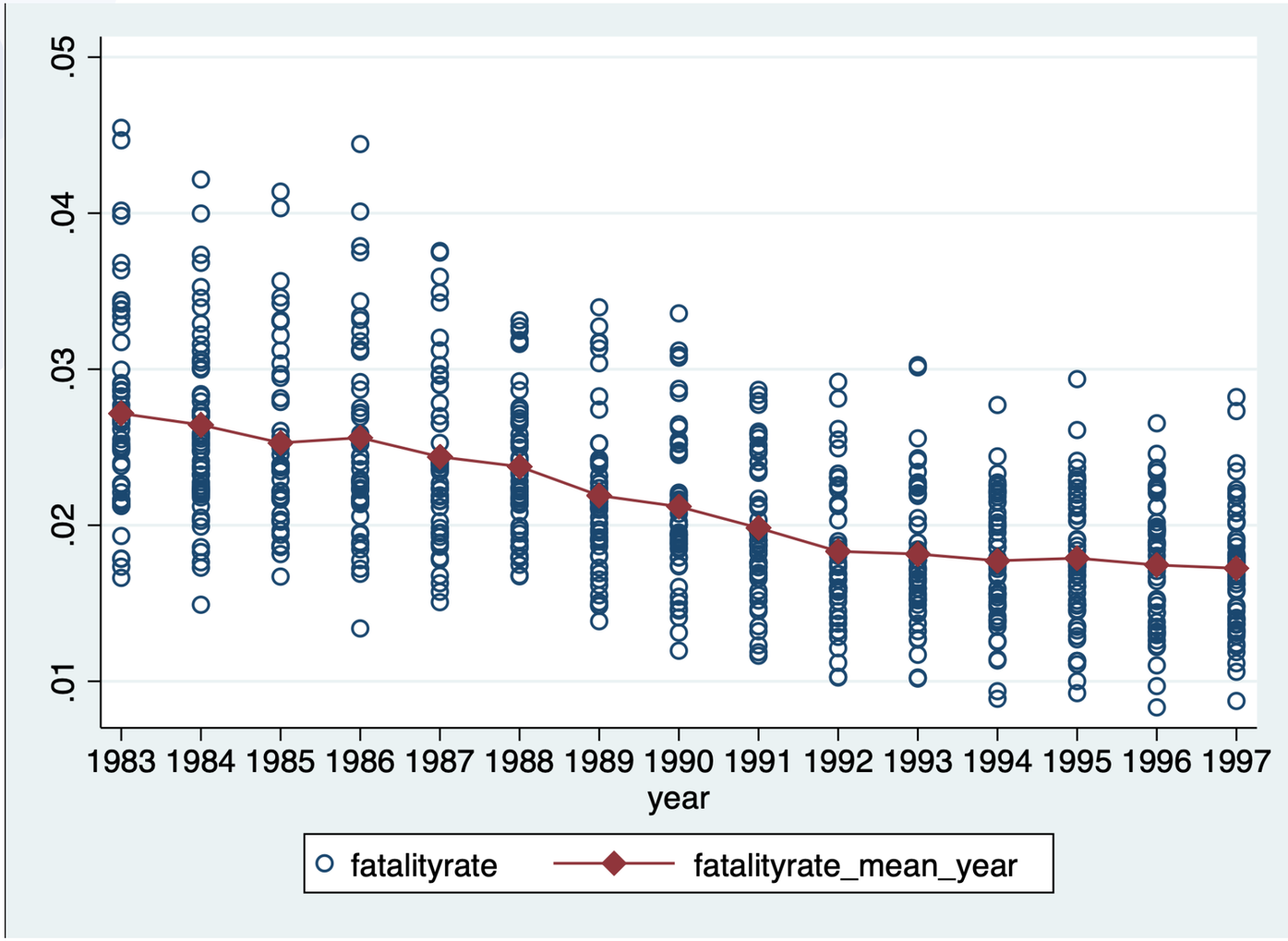
Fatality rate a clear, but messy view



A clear view of the evolutions of fatality rate across time, across each State of the USA. All in just one graph...but we have too many states and it came out very confused.

Even so, we can easily view that 1982 we had a bigger variation between *fatalityrate* and higher values per million of traffic miles. So, during the years these values seem to lower their values and their dispersion too.

Fatality rate...a mean across years

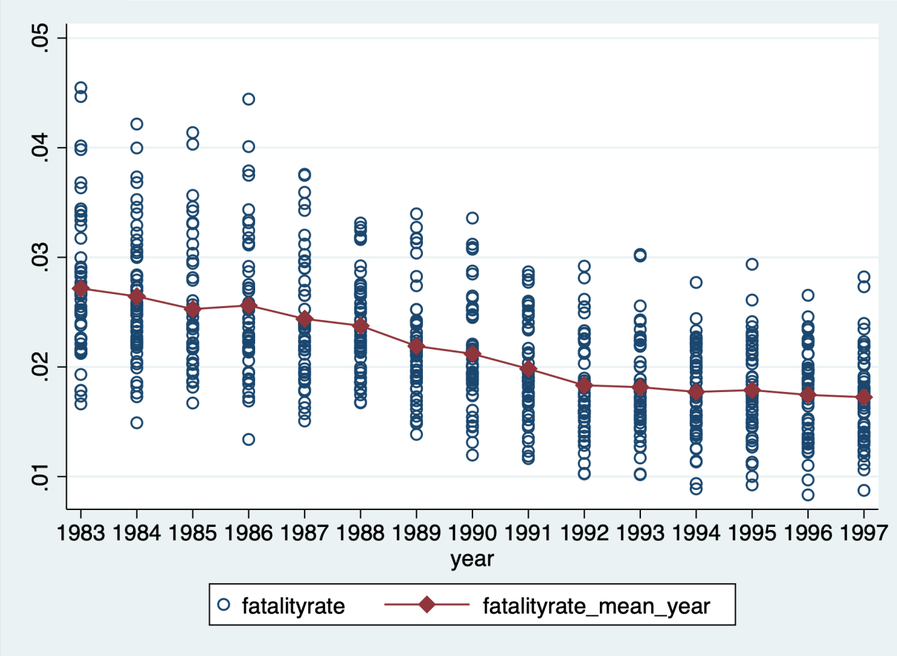


In this graph, we can show, pretty clearly that the*fatalityrate* variable decreased across time.
Registering a lower value in 1997.

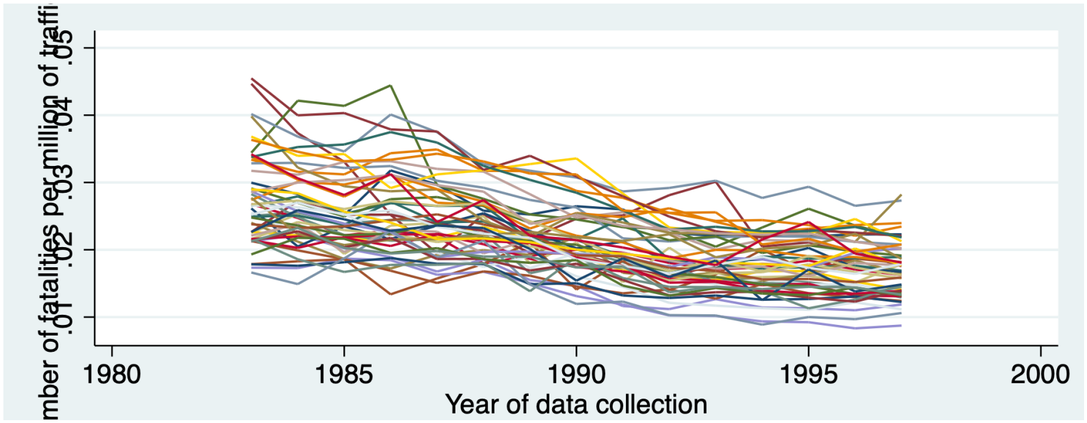
The command used is almost identical to the previous slide.

1. *bysort year: egen fatalityrate_mean_year=mean(fatalityrate)*
2. twoway scatter fatalityrate year, msymbol(circle_hollow) || connected fatalityrate_mean_year year, msymbol(diamond) || , xlabel(1983(1)1997)

Fatality rate...a mean across years



A clear view of the evolutions of fatality rate across time, across each State of the USA. All in just one graph...but we have too many states and it came out very confused. Even so, we can easily view that at in 1982 we had a bigger variation between *fatalityrate* and higher values per million of traffic miles. So, during the years these values seem to lower their values and their dispersion too.



In this graph, we can show, pretty clearly that the *fatalityrate* variable decreased across time. Registering a lower value in the 1997.

The command used are almost identical to the previous slide.

1. *bysort year: egen fatalityrate_mean_year=mean(fatalityrate)*
2. twoway scatter fatalityrate year, msymbol(circle_hollow) || connected fatalityrate_mean_year year, msymbol(diamond) || , xlabel(1983(1)1997)

Sb_useage

```
. summarize sb_useage
```

Variable	Obs	Mean	Std. Dev.	Min	Max
sb_useage	556	.5288518	.1701859	.06	.87

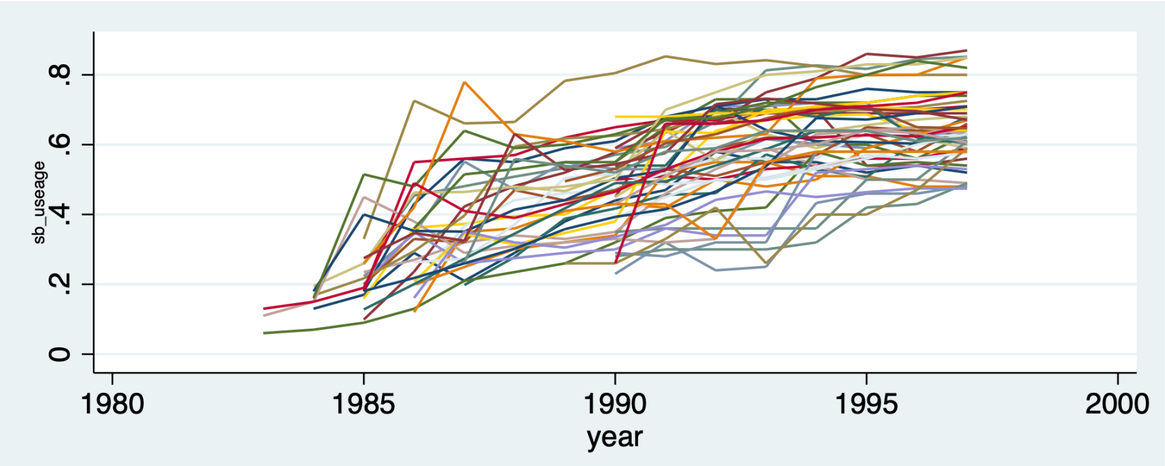
```
. describe sb_useage
```

variable name	storage type	display format	value label	variable label
sb_useage	float	%9.0g		

We used the command sb_useage to show:

- Obs, which is the number of observation;
- Mean which is the average value;
- St.dev. which is the standard deviation;
- The maximum and minimum values assumed by the variables

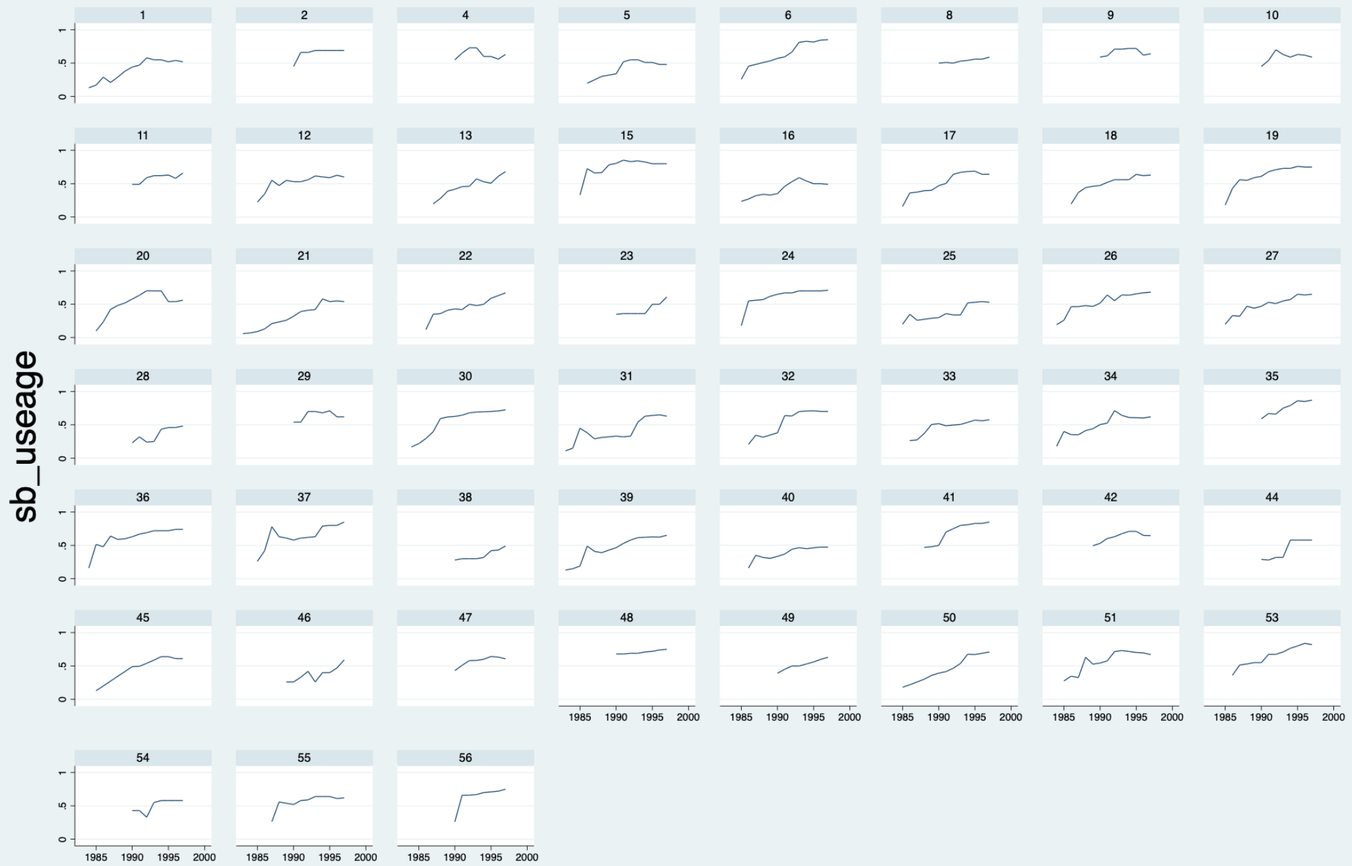
With the command describe we can have a description of what kind of variable was allocated in memory. Even in this case is a float with 9 decimal values.



In this graph, we can easily see that the trend of *sb_useage* is increasing over time.

So, to do this we execute the command:
xtline sb_useage, overlay

Sb_useage across time and states



As we can see, the *sb_usage* across time has an increase for almost every state in the US.
Note, that sometimes we have shorter graphs because the dataset does not contain all the *sb_useage* variables for each year.

How we did that?

1. Command: *xtline sb_useage*

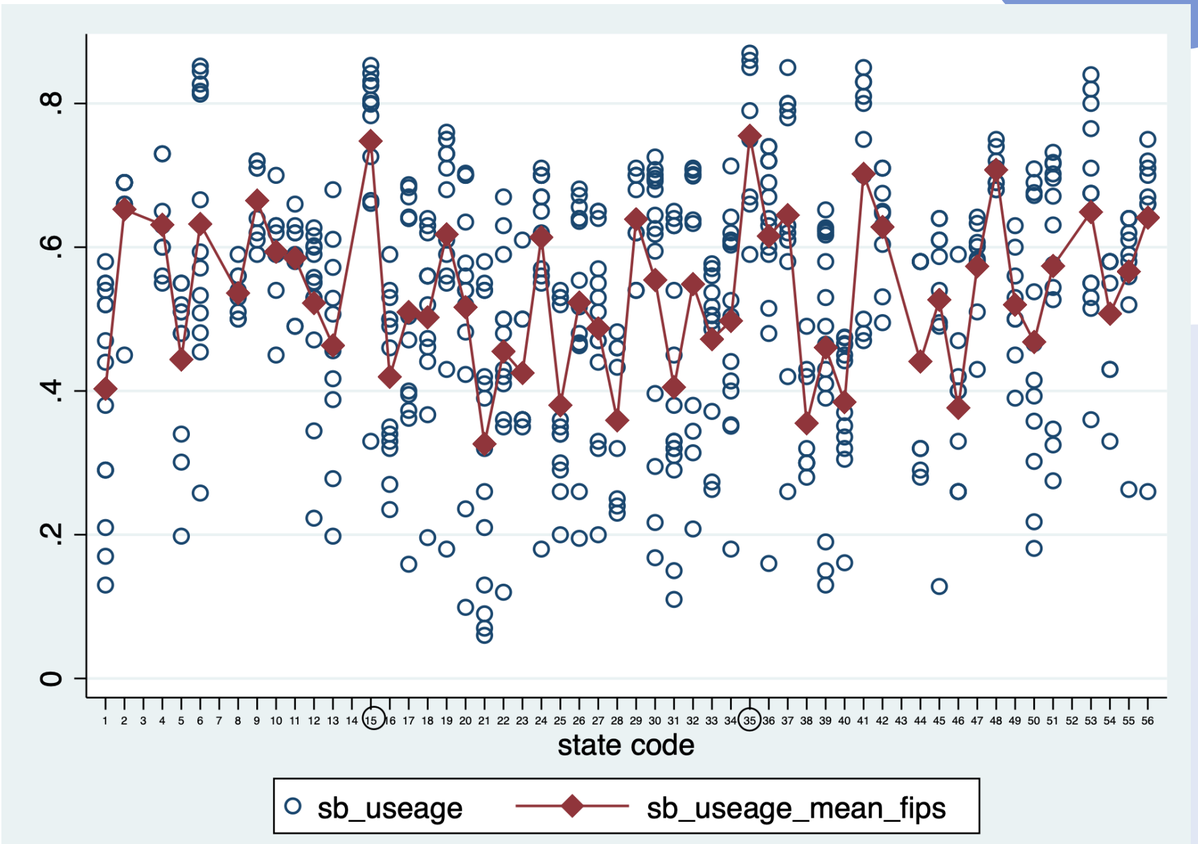
The mean of sb_useage across states

After creating a mean variable for the *sb_usage* across states, we display its values, represented by the red diamonds.

Showing that the states that have the higher *sb_useage* in the US are: 15 and 35 which respectively are: Hawaii and New Mexico.

So, to do this we execute the commands:

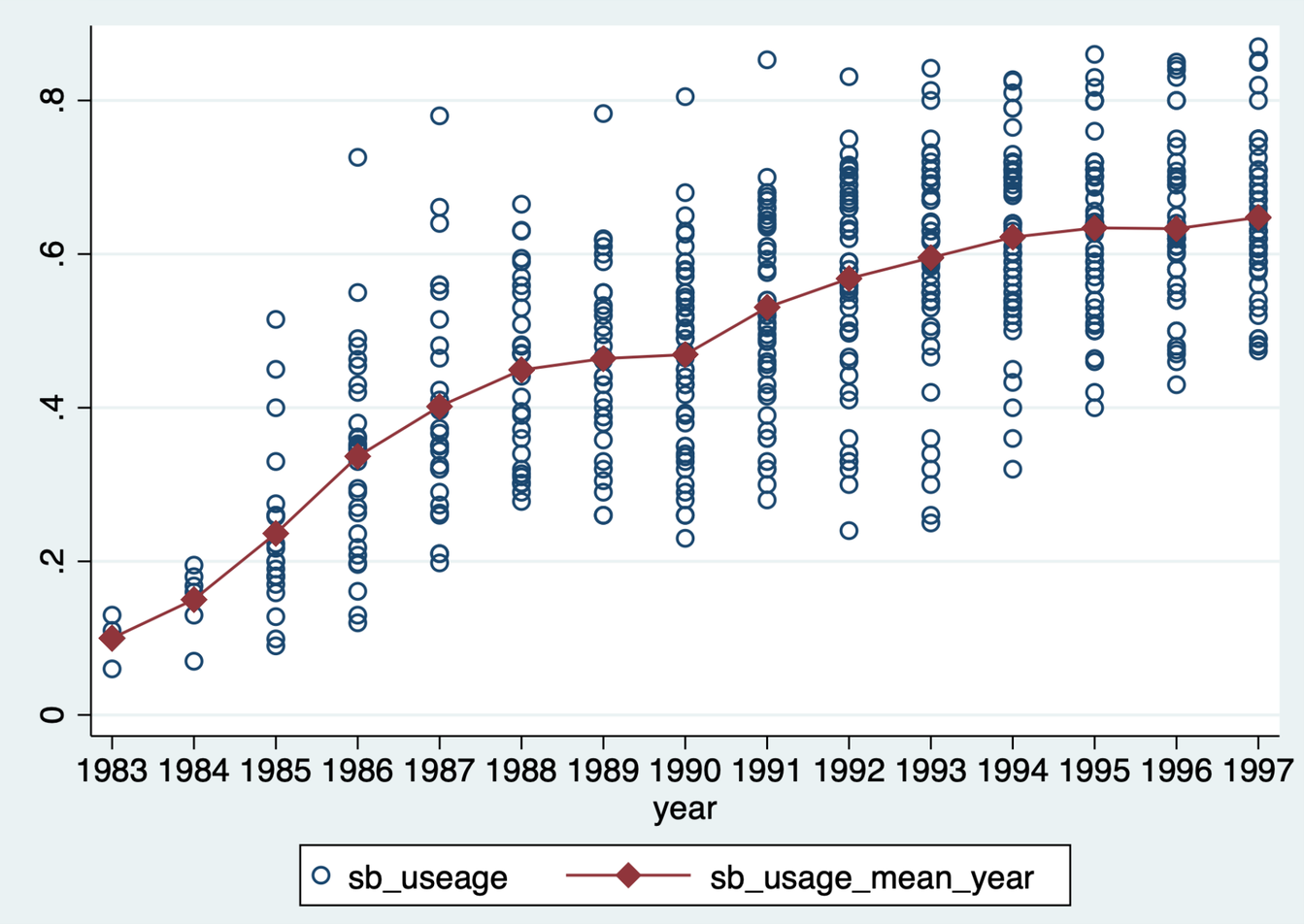
1. *bysort fips: egen sb_useage_mean_fips=mean(sb_useage)*
2. *twoway scatter sb_useage fips, msymbol(circle_hollow) || connected sb_useage_mean_fips fips, msymbol(diamond)*



To remove the blue dots do the following command:

twoway connected sb_useage_mean_fips fips, msymbol(diamond)

Mean of sb_useage across years

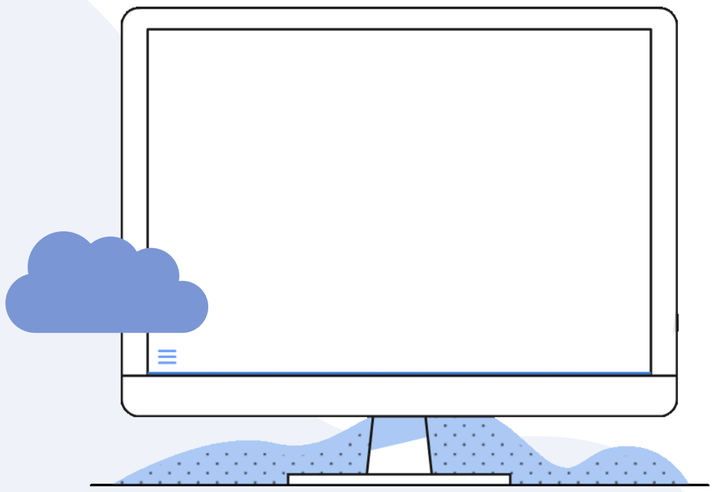


The usage of seatbealts experienced growth over time, reaching its peak in 1997.

The command used is almost identical to the previous slide.

1. b*ysort year: egen sb_usage_mean_year=mean(sb_useage)*
2. *twoway scatter sb_useage year, msymbol(circle_hollow) || connected sb_usage_mean_year year, msymbol(diamond) || , xlabel(1983(1)1997)*

Independent Variables



The definition of an independent variable is:

- The independent variable is the variable the experimenter changes or controls and is assumed to have a direct effect on the dependent variable. Two examples are the speed limit and the enforcement (identified in our db as primary and secondary).

state	year
fips	vmt
speed65	speed70
drinkage21	ba08
income	age
primary	secondary

Independent Variables

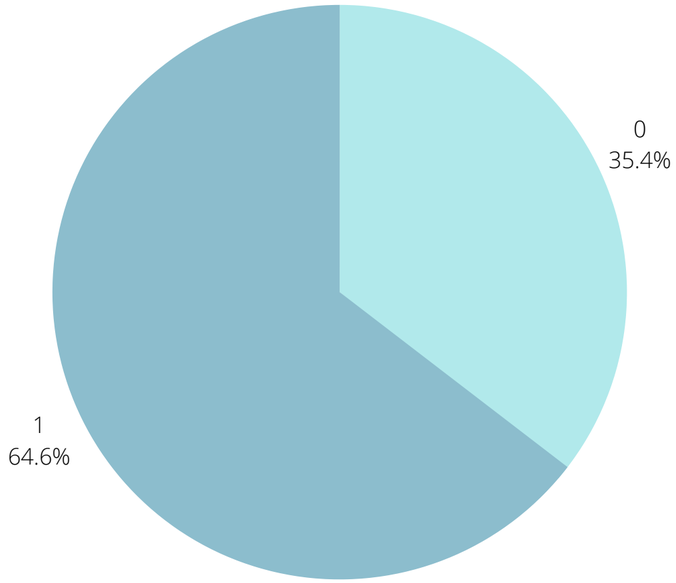
Definition of Independent Variable: An independent variable (exogenous) is that variable whose values are used to explain the values of one or more other dependent variables.

For all the independent variables we used the command tab, which is an abbreviation of tabulate to show the frequency and the percentage of the variables.
We represented the dummy variables with a pie chart and the continuous variables with a histogram.

Speed 65

```
. xttab speed65
```

speed65	Overall Freq.	Overall Percent	Between Freq.	Between Percent	Within Percent
0	271	35.42	51	100.00	35.42
1	494	64.58	49	96.08	67.21
Total	765	100.00	100	196.08	51.00
			(n = 51)		




```
. tab speed65 year, column
```

Key
frequency *column percentage*

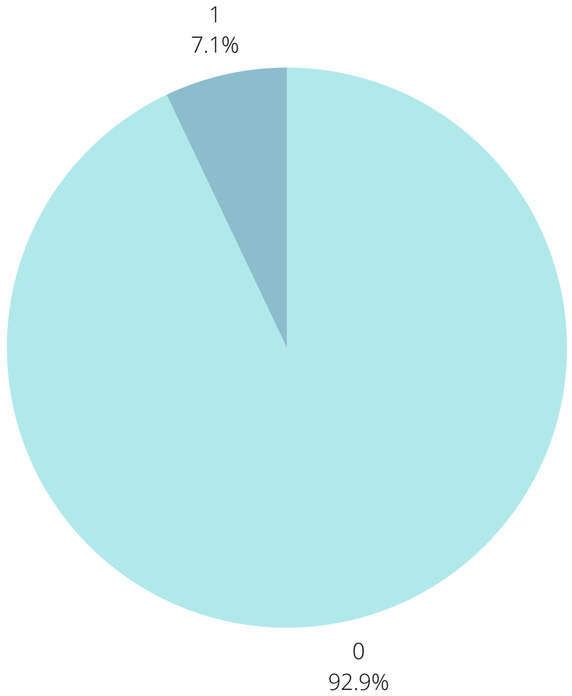
speed65	year 1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	Total
0	51 100.00	51 100.00	51 100.00	51 100.00	10 19.61	8 15.69	8 15.69	8 15.69	6 11.76	6 11.76	6 11.76	6 11.76	3 5.88	3 5.88	3 5.88	271 35.42
1	0 0.00	0 0.00	0 0.00	0 0.00	41 80.39	43 84.31	43 84.31	43 84.31	45 88.24	45 88.24	45 88.24	45 88.24	48 94.12	48 94.12	48 94.12	494 64.58
Total	51 100.00	51 100.00	51 100.00	51 100.00	51 100.00	51 100.00	51 100.00	51 100.00	51 100.00	51 100.00	51 100.00	51 100.00	51 100.00	51 100.00	51 100.00	765 100.00

Once we have done the following command, we can easily see that in 1983 there was no state that imposed the *speedlimit65.* On the contrary in 1997 the 94.12% of states decided to adopt this speed limit.

Speed70

```
. xttab speed70
```

speed70	Overall Freq.	Overall Percent	Between Freq.	Between Percent	Within Percent
0	711	92.94	51	100.00	92.94
1	54	7.06	18	35.29	20.00
Total	765	100.00	69	135.29	73.91
			(n = 51)		




```
. tab speed70 year, column
```

Key

frequency
column percentage

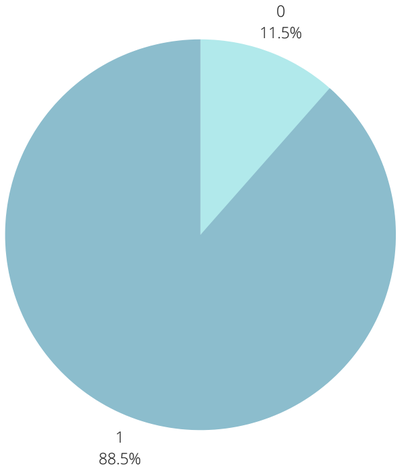
Dummy variable for 70 or higher mile per hour speed limit	year 1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	Total
0	51	51	51	51	51	51	51	51	51	51	51	51	33	33	33	711
	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	64.71	64.71	64.71	92.94
1	0	0	0	0	0	0	0	0	0	0	0	0	18	18	18	54
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	35.29	35.29	35.29	7.06
Total	51	51	51	51	51	51	51	51	51	51	51	51	51	51	51	765
	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Once we have done the following command, we can easily see that in 1983 there was no state that imposed the *speedlimit70*. On the contrary in 1997 the 35.9% of states decided to adopt this speed limit.

Drinkage21

```
. xttab drinkage21
```

drinka~21	Overall Freq.	Overall Percent	Between Freq.	Between Percent	Within Percent
0	88	11.50	32	62.75	18.33
1	677	88.50	51	100.00	88.50
Total	765	100.00	83	162.75	61.45
			(n = 51)		




```
. tab drinkage21 year, column
```

Key
frequency *column percentage*

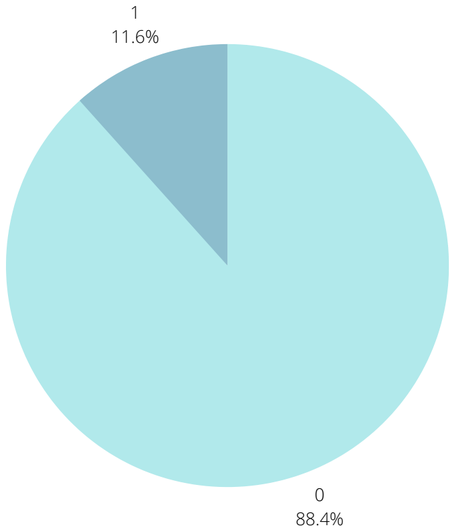
drinkage21	year 1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	Total
0	32	29	19	7	1	0	0	0	0	0	0	0	0	0	0	88
	62.75	56.86	37.25	13.73	1.96	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	11.50
1	19	22	32	44	50	51	51	51	51	51	51	51	51	51	51	677
	37.25	43.14	62.75	86.27	98.04	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	88.50
Total	51	51	51	51	51	51	51	51	51	51	51	51	51	51	51	765
	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

As we can see in the table, in 1983 the regulation involving *drinkage21* was adopted by only 37.25% of the states. Then in 1997 (fortunately), all the US states decided to adopt these regulations, forbidding the use of alcohol for people under 21 years old.

Ba08

```
. xttab ba08
```

ba08	Overall Freq.	Overall Percent	Between Freq.	Between Percent	Within Percent
0	676	88.37	49	96.08	91.97
1	89	11.63	16	31.37	37.08
Total	765	100.00	65	127.45	78.46
			(n = 51)		



As we can see in the table, in 1983 the regulation forbidding people to drive when their alcohol blood limit was over 0.8% was adopted by only 3.92% of the states. Then, in 1997 all the states decided to adopt this regulation in order to prevent further risk.

```
. tab ba08 year, column
```

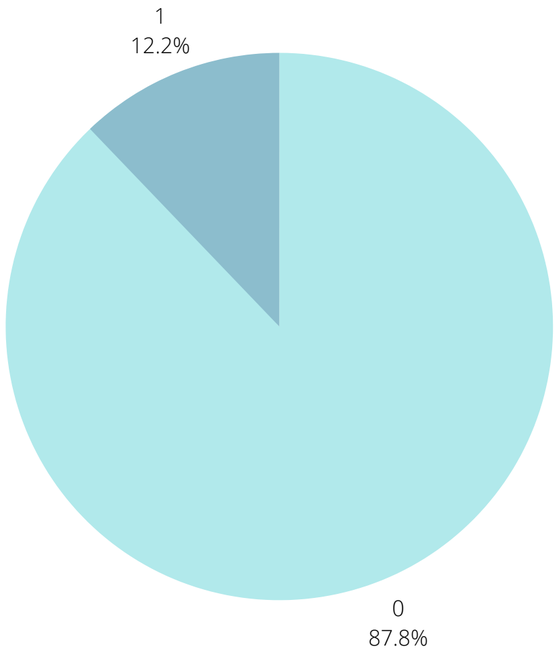
Key
frequency *column percentage*

ba08	year 1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	Total
0	49	49	49	49	49	48	48	46	46	46	44	40	39	39	35	676
	96.08	96.08	96.08	96.08	96.08	94.12	94.12	90.20	90.20	90.20	86.27	78.43	76.47	76.47	68.63	88.37
1	2	2	2	2	2	3	3	5	5	5	7	11	12	12	16	89
	3.92	3.92	3.92	3.92	3.92	5.88	5.88	9.80	9.80	9.80	13.73	21.57	23.53	23.53	31.37	11.63
Total	51	51	51	51	51	51	51	51	51	51	51	51	51	51	51	765
	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Primary

```
. xttab primary
```

primary	Overall Freq.	Overall Percent	Between Freq.	Between Percent	Within Percent
0	672	87.84	51	100.00	87.84
1	93	12.16	8	15.69	77.50
Total	765	100.00	59	115.69	86.44
			(n = 51)		




```
. tab primary year, column
```

Key
frequency *column percentage*

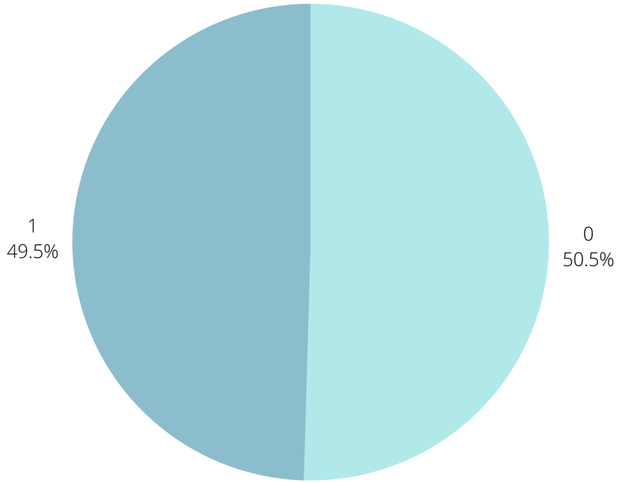
primary	year 1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	Total
0	51	51	50	44	44	43	44	44	43	43	43	43	43	43	43	672
	100.00	100.00	98.04	86.27	86.27	84.31	86.27	86.27	84.31	84.31	84.31	84.31	84.31	84.31	84.31	87.84
1	0	0	1	7	7	8	7	7	8	8	8	8	8	8	8	93
	0.00	0.00	1.96	13.73	13.73	15.69	13.73	13.73	15.69	15.69	15.69	15.69	15.69	15.69	15.69	12.16
Total	51	51	51	51	51	51	51	51	51	51	51	51	51	51	51	765
	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Doing this command, we are able to see that in 1983 the primary enforcement wasn't adopted in any State. So, we suppose that there used to be some other type of enforcement. Anyway, the percentage level of primary enforcement across the states reached a peak in 1991 and maintained its constant value of 15.69% since 1997.

Secondary

```
. xttab secondary
```

secondary	Overall Freq.	Overall Percent	Between Freq.	Between Percent	Within Percent
0	386	50.46	51	100.00	50.46
1	379	49.54	42	82.35	60.16
Total	765	100.00	93	182.35	54.84
			(n = 51)		




```
. tab secondary year, column
```

Key
frequency *column percentage*

secondary	year 1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	Total
0	51	51	50	38	31	26	24	23	20	17	15	12	10	9	9	386
	100.00	100.00	98.04	74.51	60.78	50.98	47.06	45.10	39.22	33.33	29.41	23.53	19.61	17.65	17.65	50.46
1	0	0	1	13	20	25	27	28	31	34	36	39	41	42	42	379
	0.00	0.00	1.96	25.49	39.22	49.02	52.94	54.90	60.78	66.67	70.59	76.47	80.39	82.35	82.35	49.54
Total	51	51	51	51	51	51	51	51	51	51	51	51	51	51	51	765
	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

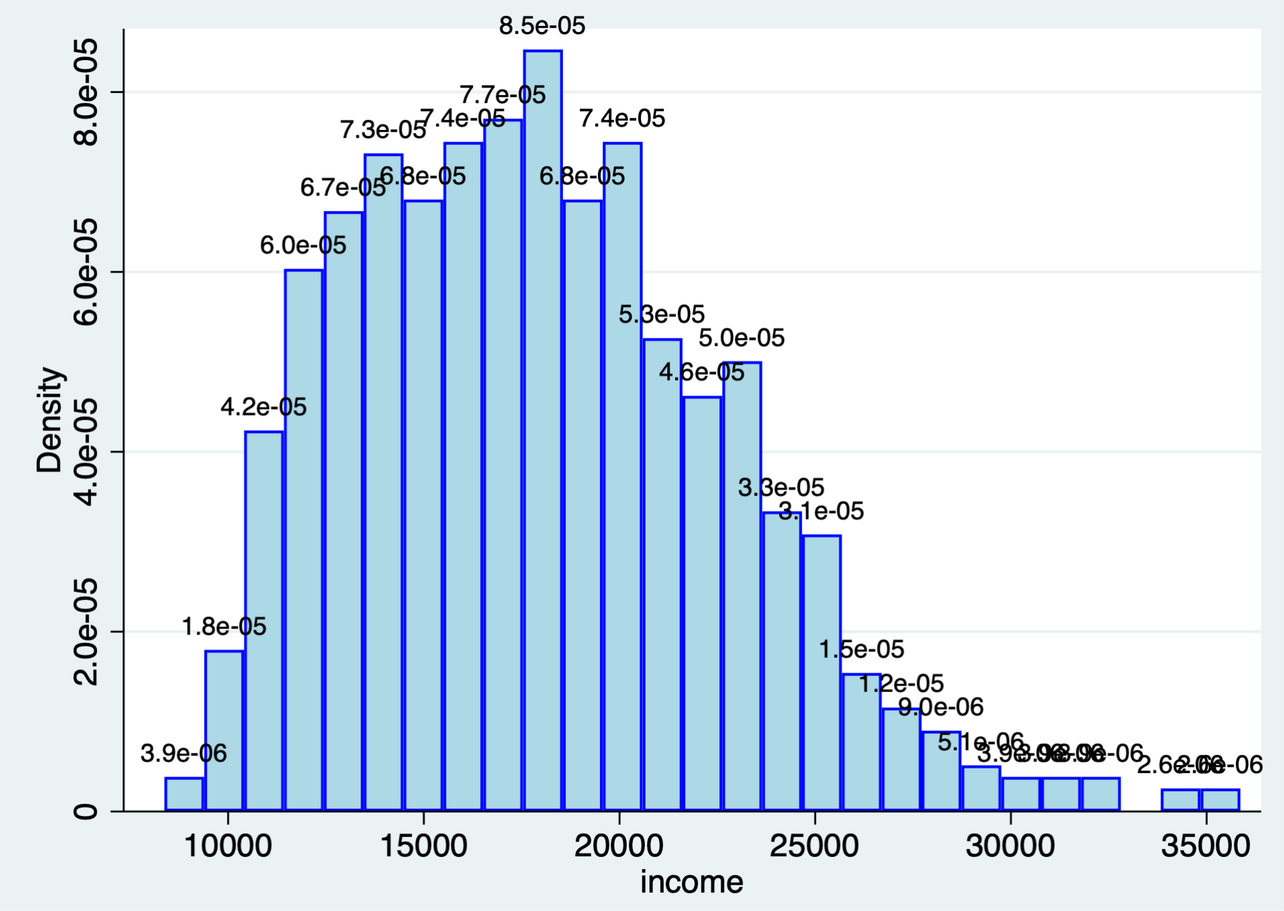
Doing this command, we are able to see that in 1983 the secondary enforcement wasn't adopted in any State. So, we suppose that there used to be some other type of enforcement. Anyway, the percentage level of secondary enforcement across the states reached a peak in 1996 and maintained its constant value of 82.35% since 1997.

To be more clear about the meaning of the variables primary and secondary:

Primary enforcement is a law allowing the police to stop and fine violators even if they do not engage in other offenses, rather than secondary enforcement is a law allowing the police to fine violators only when they are stopped for some other offense. While observers and policymakers have noticed that states with primary enforcement have on average higher usage rates, we are able to identify and estimate the effects of primary enforcement in a statistically more reliable way.

Income

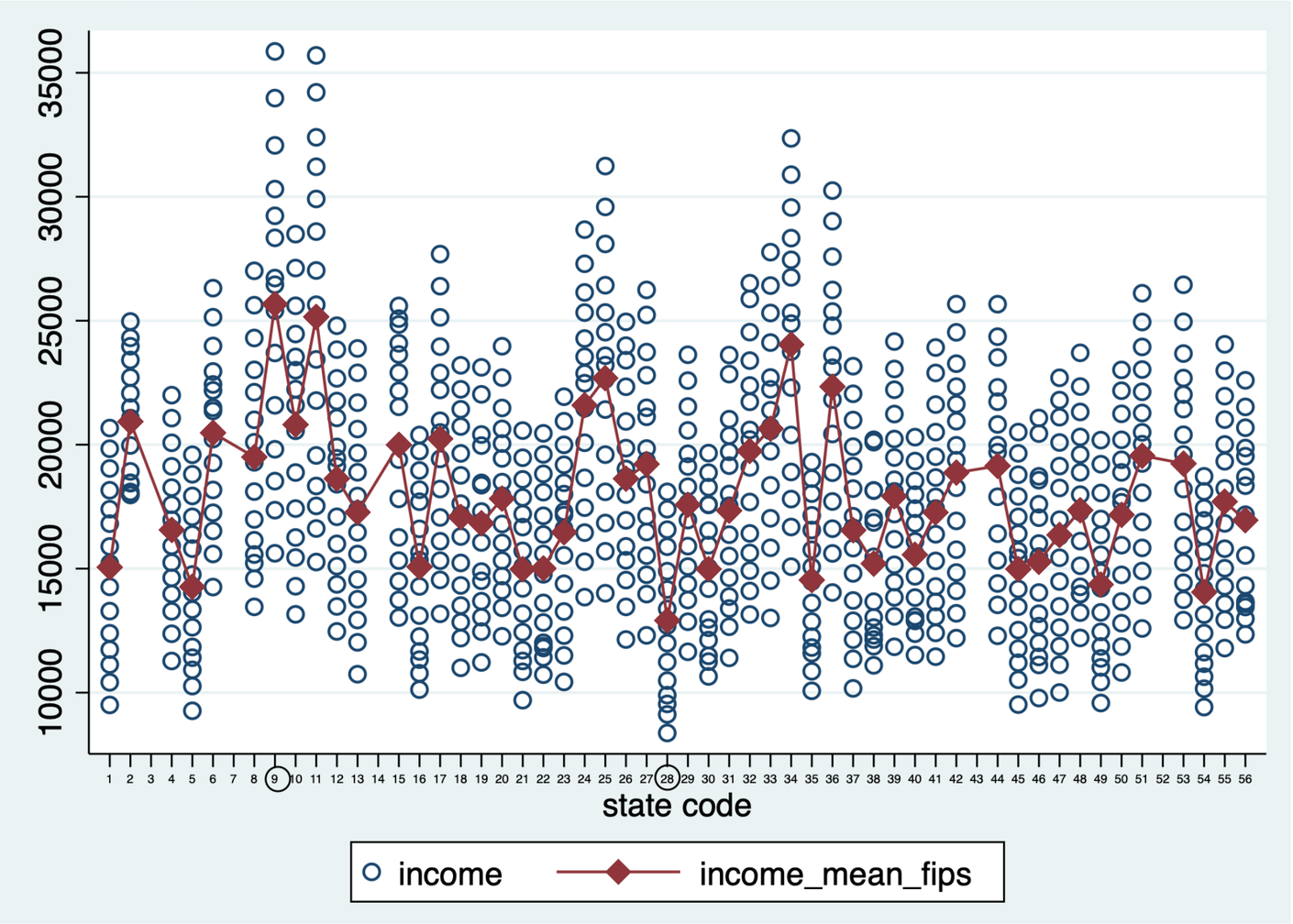
```
. histogram income, fcolor(ltblue) lcolor(blue) addlabel
(bin=27, start=8372, width=1018.1852)
```



By doing this histogram, we can compare the income and its density, which it seem clear that reach higher peaks between 1500 and 2000. Across all the dataset given by the study case.

Note, that in this case doing a tab command is pretty useless due to we will obtain a large table with all its values f

Income across states



Command:

1. *bysort fips: egen income_mean_fips=mean(income)*
2. *twoway scatter income fips, msymbol(circle_hollow) || connected income_mean_fips fips, msymbol(diamond)*

The higher income was registered on average by the state with fips=9, which is Connecticut. On the contrary, the lower average income value was registered by the 28, which is Mississippi.

Income across years

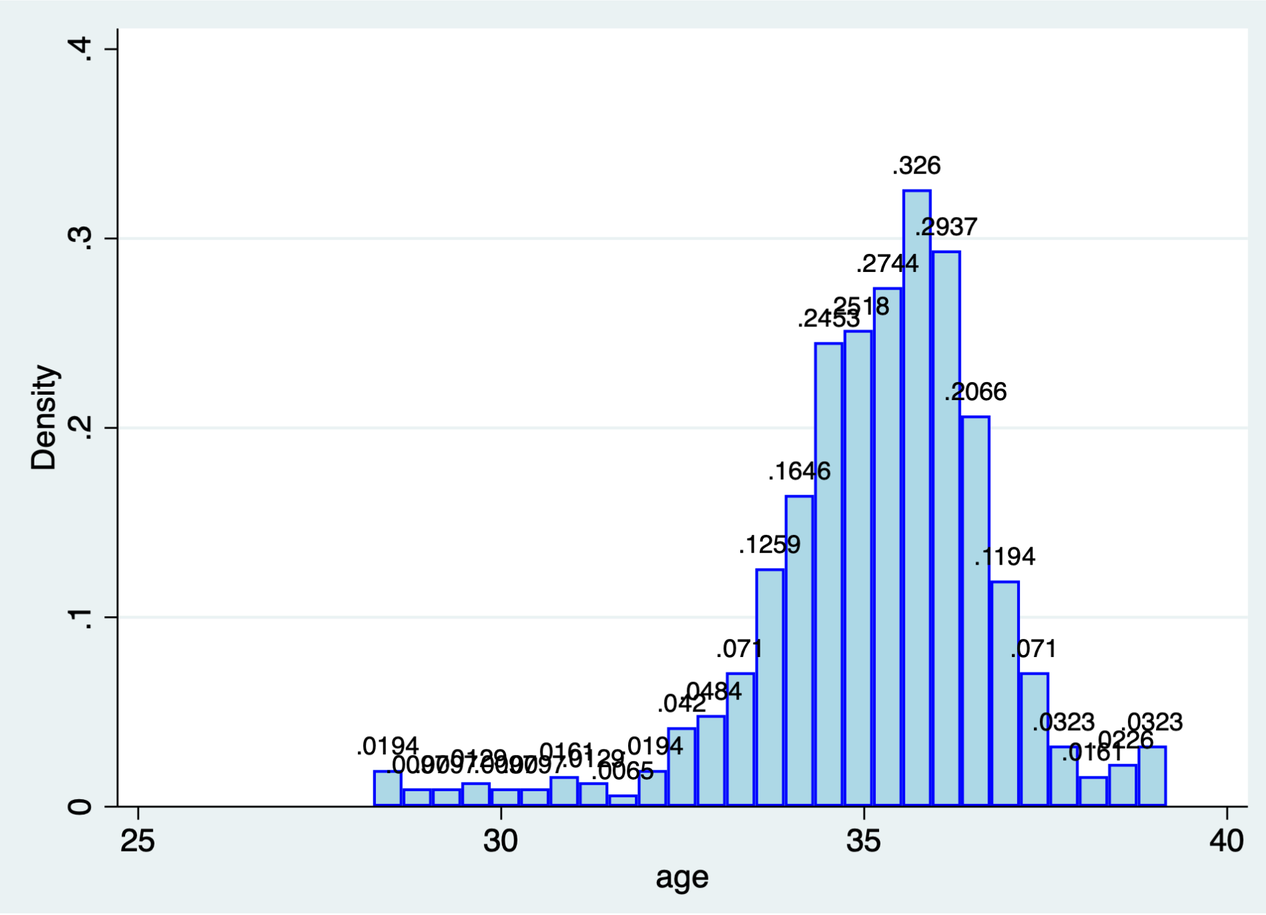
Command:

1. *bysort year: egen income_mean_year=mean(income)*
2. *twoway scatter income year, msymbol(circle_hollow) || connected income_mean_year year, msymbol(diamond) || , xlabel(1983(1)1997)*

So, as we can see, the income variable was subjected to a constant increase since 1983.

Age

```
.                    histogram age, fcolor(ltblue) lcolor(blue)addlabel
(bin=27, start=28.234966, width=.40498578)
```



By doing this histogram, we can compare the age and its density, which it seem clear that reach higher peaks around 37.

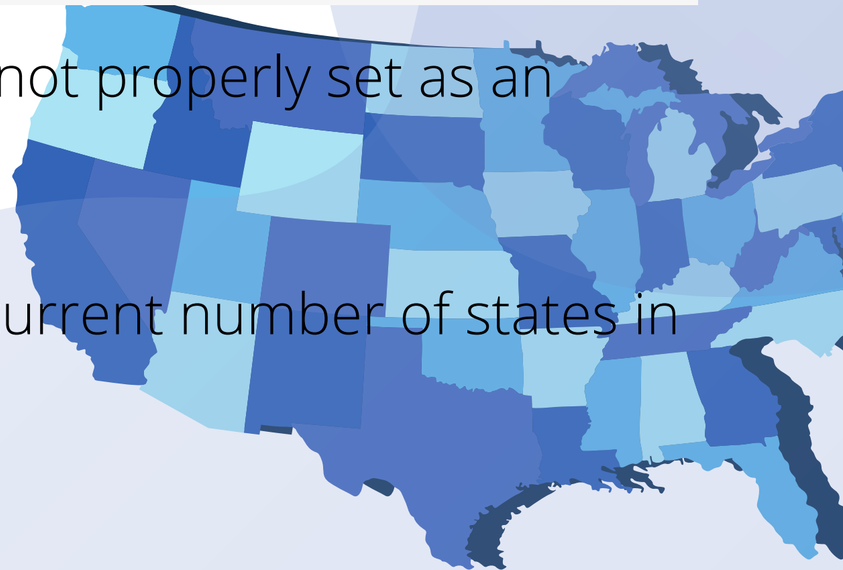
State & Fips

```
. levelsof state
`"AK"' `"AL"' `"AR"' `"AZ"' `"CA"' `"CO"' `"CT"' `"DC"' `"DE"' `"FL"' `"GA"' `"HI"' `"IA"' `"ID"' `"IL"' `"IN"' `"KS"' `"KY"' `"LA"' `"MA"' `"MD"' `"ME"' `"MI"' `"MN"' `"MO"' `"MS"' `"MT"'
> "NM"' `"NV"' `"NY"' `"OH"' `"OK"' `"OR"' `"PA"' `"RI"' `"SC"' `"SD"' `"TN"' `"TX"' `"UT"' `"VA"' `"VT"' `"WA"' `"WI"' `"WV"' `"WY"'
```

By doing the command: *levelsof state* we can clearly see that Stata can show us all the states presented in our dataset without duplicates values. This is very important in this case in order to avoid to represent an enormous table of states, with the aim just to show all the 56 states of the US.

```
. levels fips
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 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 56
```

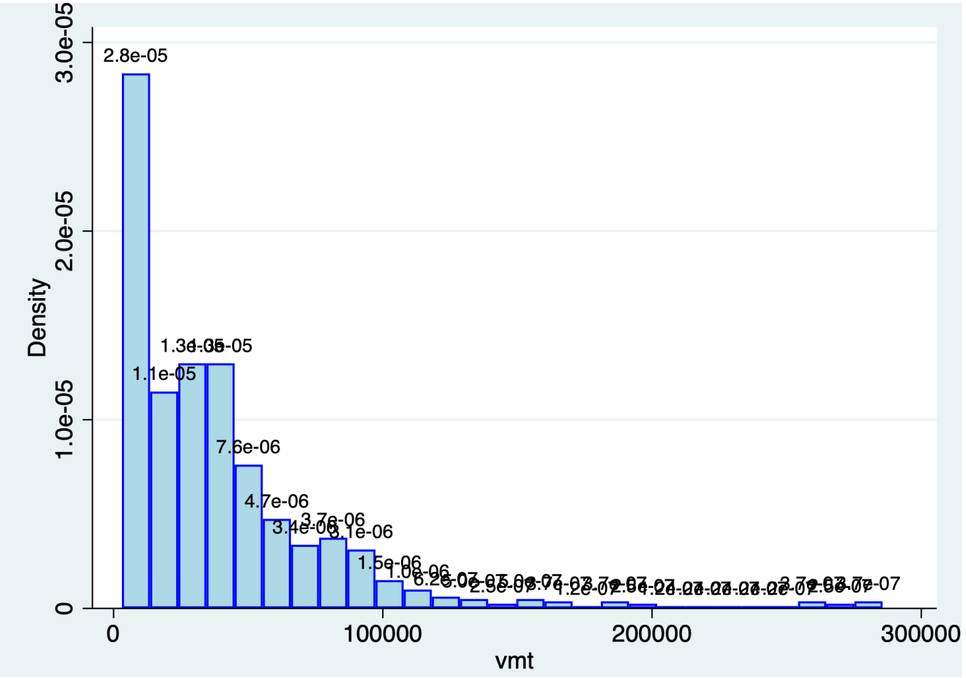
With this command, we can see that the code of each state is not properly set as an incremental counter.
For instance, states number 3 and 7 doesn't exist.
In fact, the totality of those values is equal to 51, which is the current number of states in the US.



Year & VMT

```
. levels year
1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997
```

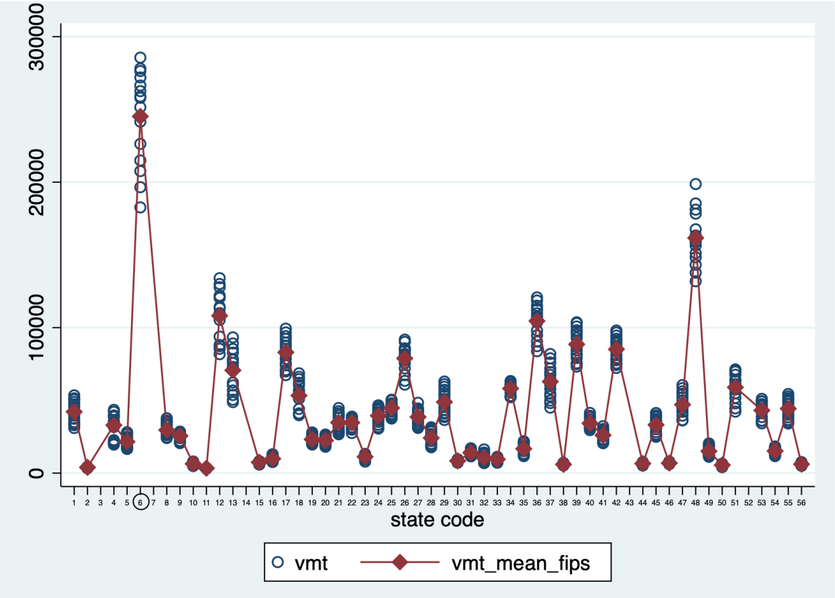
By doing the command: *levelsof year* we can display all the years we are working with. Moreover, knowing that we have a strongly balanced panel from the previous analysis. We also know that all these years are showed for each state of the US.




```
. histogram vmt, fcolor(ltblue) lcolor(blue) addlabel
(bin=27, start=3099, width=10463.444)
```

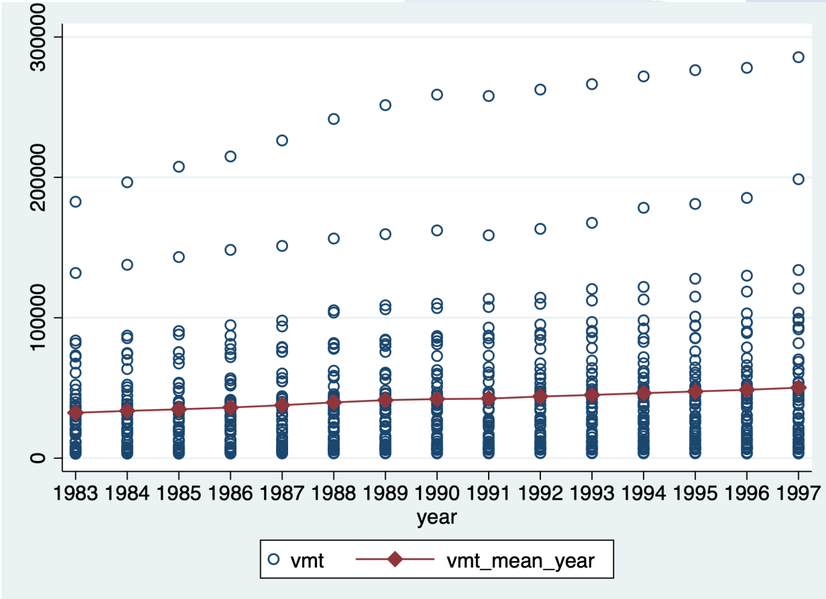
By doing this histogram, we can compare the *vmt* **millions of traffic miles per year** and its density, which it seems clear that the majority of the drivers do less than 100,000vmt. They usually drive for short routes.

Mean of vmt across states and years



Commands:

1. *bysort fips: egen vmt_mean_fips=mean(vmt)*
2. *twoway scatter vmt fips, msymbol(circle_hollow) || connected vmt_mean_fips fips, msymbol(diamond)*



Commands:

1. *bysort year: egen vmt_mean_year=mean(vmt)*
2. *twoway scatter vmt year, msymbol(circle_hollow) || connected vmt_mean_year year, msymbol(diamond) || , xlabel(1983(1)1997)*

3. Statistical and Econometric Methods

Regression model

There are three main types of panel data models:

1. Pooled OLS (Ordinary Least Square): model treats a dataset like any other cross-sectional data and ignores that the data has a time and individual dimensions. That is why the assumptions are similar to that of ordinary linear regression.
2. Fixed effects models go a step further by taking into account the differences between individual entities.
3. Random effects model: In fixed effects model we have controlled for differences between individual countries. But what about variables that are constant across individuals but change over time? A random effects model takes into consideration these individual variations as well as time dependent-variations. The model eliminates biases from variables that are unobserved and change over time.

Pooled OLS

A pooled OLS analysis is just an OLS analysis that can be done for Panel Data.
So, its regression model is:

$$Y_i = \beta_0 + \sum_{i=1}^{n} B_i * X_i + u_i$$

Y_i = is the dependent variable
B_0 = intercept
∑B_i*X_i = regression function
B_i = angolar coefficient
X_i = independent variable
u_i = statistical error

```
. reg fatalityrate sb_useage vmt speed65 speed70 drinkage21 ba08 income age primary secondary year
```

Source	SS	df	MS	Number of obs	=	556
				F(11, 544)	=	52.75
Model	.007246243	11	.000658749	Prob > F	=	0.0000
Residual	.006793163	544	.000012487	R-squared	=	0.5161
				Adj R-squared	=	0.5064
Total	.014039406	555	.000025296	Root MSE	=	.00353

fatalityrate	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
sb_useage	-.0003651	.0017689	-0.21	0.837	-.0038397	.0031096
vmt	-2.21e-09	3.48e-09	-0.63	0.526	-9.04e-09	4.63e-09
speed65	.0003267	.0005106	0.64	0.523	-.0006763	.0013296
speed70	.0023998	.0005776	4.15	0.000	.0012651	.0035345
drinkage21	-.0011715	.0009095	-1.29	0.198	-.0029581	.0006151
ba08	-.0017607	.0004874	-3.61	0.000	-.0027182	-.0008032
income	-7.72e-07	5.77e-08	-13.38	0.000	-8.85e-07	-6.59e-07
age	-.0000408	.0001152	-0.35	0.724	-.0002671	.0001855
primary	.001822	.000758	2.40	0.017	.000333	.0033109
secondary	.0009606	.0005204	1.85	0.065	-.0000616	.0019827
year	-.0000667	.0000863	-0.77	0.439	-.0002362	.0001027
_cons	.1695551	.1701217	1.00	0.319	-.1646207	.5037309

Note: There is also the command *areg* which fits a linear regression absorbing one categorical factor. areg is designed for datasets with many groups, but not a number of groups that increases with the sample size.
The *reg* regression which does ordinary least squares regression of independent observations. -areg- is used with data that includes multiple observations on the same entities, and fits a regression model that allows each entity to have its own intercept, but all other coefficients are the same. The effects estimated by -areg- are within entity effects; those estimated by -reg- are between. So, for the reason that areg is a more appropriated regression for our type of dataset which is a panel data we decide to use it later on in our analysis.

Fixed Effect model

This type of regression model is used when we are interested in analyzing the impact of a variable that varies over time. In our case, this could be taken into account for different variables. Such that, *fatalityrate* or *sb_usage*. So, FE model explores the relationship between the predictors and outcome variables.
Hence, when we use this type of model, we assume that something within the individual may impact or bias the predictor or outcome variables, and we need to check this. FE removes the effect of those time-invariant characteristics so we can assess the net effect of the predictors on the outcome variable.
Moreover, the time-invariant characteristics are unique to the individual and should not be correlated with other individual characteristics.
To conclude, if the error terms are correlated, the FE is no suitable since inferences may not be correct and we need to model that relationship with other regression model, and this is the main rationale for the Hausman test.

The FE model equation is:

$$Y_it = B_1 * X_it + \alpha_i + u_it$$

Y_it = is the dependent variable (DV) where i = entity and t = time.
a_i (i=1….n) is the unknown intercept for each entity (n entity-specific intercepts).
Xit represents one independent variable (IV),
β1 is the coefficient for that IV,
uit is the error term

```
. xtreg fatalityrate sb_useage vmt speed65 speed70 drinkage21 ba08 income age primary secondary year, fe

Fixed-effects (within) regression               Number of obs      =       556
Group variable: fips                            Number of groups   =        51

R-sq:                                           Obs per group:
     within  = 0.7204                                         min =         8
     between = 0.1115                                         avg =      10.9
     overall = 0.0061                                         max =        15

                                                F(11,494)          =    115.70
corr(u_i, Xb)  = -0.6679                        Prob > F           =    0.0000

------------------------------------------------------------------------------
fatalityrate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   sb_useage |  -.0049444   .0012651    -3.91   0.000      -.00743   -.0024588
         vmt |  -5.62e-08   1.84e-08    -3.06   0.002    -9.23e-08   -2.01e-08
     speed65 |  -.0001069    .000332    -0.32   0.748    -.0007592    .0005453
     speed70 |    .001874   .0003246     5.77   0.000     .0012362    .0025119
  drinkage21 |   -.000147   .0005185    -0.28   0.777    -.0011657    .0008718
        ba08 |  -.0004914   .0003709    -1.32   0.186    -.0012202    .0002373
      income |   6.01e-07   1.41e-07     4.27   0.000     3.25e-07    8.78e-07
         age |   .0015661   .0004398     3.56   0.000     .0007021    .0024301
     primary |   .0011654   .0008605     1.35   0.176    -.0005253     .002856
   secondary |   .0002832   .0003362     0.84   0.400    -.0003774    .0009438
        year |   -.001269   .0001668    -7.61   0.000    -.0015967   -.0009413
       _cons |   2.484901   .3207987     7.75   0.000     1.854603    3.115199
-------------+----------------------------------------------------------------
     sigma_u |  .00674803
     sigma_e |  .00169539
         rho |  .94062523   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(50, 494) = 37.39                     Prob > F = 0.0000
```

Random model

In this case, the variation across entities is assumed to be random and uncorrelated with the predictor or independent variables included in the model. When there is the chance to believe that differences across entities have some influence on our dependent variable then we should use random effects. An advantage of random effects is that you can include time-invariant variables. In the fixed effects model these variables are absorbed by the intercept.

The Random effect equation is:

$$Y_{it} = \beta X_{it} + \alpha + u_{it} + \varepsilon_{it}$$

e_it = within-entity error
u_it = between entity error

```
. xtreg fatalityrate sb_useage vmt speed65 speed70 drinkage21 ba08 income age primary secondary year, re

Random-effects GLS regression                   Number of obs     =       556
Group variable: fips                            Number of groups  =        51

R-sq:                                           Obs per group:
     within  = 0.6992                                         min =         8
     between = 0.0442                                         avg =      10.9
     overall = 0.2763                                         max =        15

                                                Wald chi2(11)     =   1104.99
corr(u_i, X)   = 0 (assumed)                    Prob > chi2       =    0.0000
```

fatalityrate	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sb_useage	-.0056765	.0013066	-4.34	0.000	-.0082373	-.0031156
vmt	-2.54e-08	9.42e-09	-2.70	0.007	-4.39e-08	-6.93e-09
speed65	-.0001965	.0003437	-0.57	0.568	-.00087	.0004771
speed70	.0015491	.0003379	4.59	0.000	.0008869	.0022113
drinkage21	-.0000259	.0005322	-0.05	0.961	-.0010689	.0010172
ba08	-.0009419	.0003781	-2.49	0.013	-.001683	-.0002009
income	-9.08e-08	1.06e-07	-0.85	0.393	-2.99e-07	1.18e-07
age	.0002188	.000262	0.84	0.404	-.0002948	.0007324
primary	.0015567	.0008112	1.92	0.055	-.0000333	.0031467
secondary	.0003039	.0003487	0.87	0.383	-.0003795	.0009873
year	-.0004763	.0001143	-4.17	0.000	-.0007004	-.0002522
_cons	.9664871	.2231792	4.33	0.000	.5290639	1.40391
sigma_u	.00326321					
sigma_e	.00169539					
rho	.78744562	(fraction of variance due to u_i)				

Pro and Cons

Pooled OLS: OLS can simply be implemented on a computer using available algorithms from linear algebra. The implementation is so efficient that can be applied even to problems with hundreds of features and tens of thousand data points. OLS produces an easily interpretable solution.

FE model: We use fixed-effects (FE) when we are only interested in analyzing the impact of variables that vary over time. FE explore the relationship between predictor and outcome variables within an entity (country, person, company, etc.). Each entity has its own individual characteristics that may or may not influence the predictor variables. When using FE we assume that something within the individual may impact or bias the predictor or outcome variables and we need to control for this. This is the rationale behind the assumption of the correlation between entity's error term and predictor variables. FE remove the effect of those time-invariant characteristics so we can assess the net effect of the predictors on the outcome variable.

Random model: The rationale behind random effects model is that, unlike the fixed effects model, the variation across entities is assumed to be random and uncorrelated with the predictor or independent variables included in the model. So, an advantage of random effects is that we can include time-invariant variables.

Pooled OLS: OLS can work in the wrong way when some points have extremely large or small values for the dependent variable compared to the rest of the dataset. This is because the objective of the least-squares methods is to minimize the sum of squared error. Moreover, when some independent variables are strongly correlated to each other, OLS can give bad results.

FE model: The fixed-effects model controls for all time-invariant differences between the individuals, so the estimated coefficients of the fixed-effects models cannot be biased because of omitted time-invariant characteristics. One side effect of the features of fixed-effects models is that they cannot be used to investigate time-invariant causes of the dependent variables. Technically, time-invariant characteristics of the individuals are perfectly collinear with the person (or entity) dummies.

Random model: Random effects assume that the entity's error term is not correlated with the predictors which allows for time-invariant variables to play a role as explanatory variables. In random-effects, we need to specify those individual characteristics that may or may not influence the predictor variables. The problem with this is that some variables may not be available therefore leading to omitted variable bias in the model.

T-test

We used the t-test method in order to understand if the variables were significant for the analysis, considering a 95% confidence interval to reject the nullhypothesis on individually null OLS estimators. It means that variables with a p-value higher than 0,05 were taken out of the model, starting from the variable with the higher p-value, because the null-hypothesis could not be rejected. This method confers statistical significance to the resulting regression since the p-value is based on t-test.
The t-values also show the importance of a variable in the model: we can obtain it by dividing the coefficient by its standard error.

A **t-test** is a type of inferential statistic used to determine if there is a significant difference between the means of two groups, which may be related in certain features. It is a method of testing hypotheses about the mean of a small sample drawn from a normally distributed population when the population standard deviation is unknown. In particular, it is usually used to determine whether the mean of a dependent variable is statistically equivalent or not to the mean of another dependent variable or to a certain value.

```
. reg fatalityrate sb_useage vmt speed65 speed70 drinkage21 ba08 income age primary secondary year
```

Source	SS	df	MS		
Model	.007246243	11	.000658749	Number of obs	= 556
Residual	.006793163	544	.000012487	F(11, 544)	= 52.75
				Prob > F	= 0.0000
Total	.014039406	555	.000025296	R-squared	= 0.5161
				Adj R-squared	= 0.5064
				Root MSE	= .00353

fatalityrate	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
sb_useage	-.0003651	.0017689	-0.21	0.837	-.0038397	.0031096
vmt	-2.21e-09	3.48e-09	-0.63	0.526	-9.04e-09	4.63e-09
speed65	.0003267	.0005106	0.64	0.523	-.0006763	.0013296
speed70	.0023998	.0005776	4.15	0.000	.0012651	.0035345
drinkage21	-.0011715	.0009095	-1.29	0.198	-.0029581	.0006151
ba08	-.0017607	.0004874	-3.61	0.000	-.0027182	-.0008032
income	-7.72e-07	5.77e-08	-13.38	0.000	-8.85e-07	-6.59e-07
age	-.0000408	.0001152	-0.35	0.724	-.0002671	.0001855
primary	.001822	.000758	2.40	0.017	.000333	.0033109
secondary	.0009606	.0005204	1.85	0.065	-.0000616	.0019827
year	-.0000667	.0000863	-0.77	0.439	-.0002362	.0001027
_cons	.1695551	.1701217	1.00	0.319	-.1646207	.5037309

R-squared and others observations

R-squared is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable. R-squared values range from 0 to 1. If it is close to 0, it means that there is no correlation between the dependent variable and the independent variable, instead, if it is close to 1 it indicates a strong correlation.
R-squared gives an estimate of the relationship between movements of a dependent variable based on independent variable movements. However, it doesn't tell if the model we have chosen is good or bad or will it tell us whether the data and predictions are distorted. A high or low R-square isn't necessarily good or bad, as it doesn't convey the reliability of the model. A better estimate taking into account degrees of freedom is the adjusted R-squared which increases if the t-statistic of a newly added regressor is greater than one in absolute value.

In the previous table given by the command: *reg ...*
We can see that:

- R^2 = 0.5161 which indicates that the 51.61% of the variance in science scores can be predicted from the variables *sb_useage, vmt, speed65 speed70, drinkage21, ba08, income, age, primary, secondary and year.*
- Prob>F = 0.0000 which indicates that If this number is < 0.05 then our model is **ok**. This is a test (F) to see whether all the coefficients in the model are different than zero.
- Adj R-squared = 0.5061 shows the same as R-sqr but adjusted by the number of cases and number of variables. So, as we can see this is approximately equal to R^2.

Multicollinearity

Collinearity means that two variables are near perfect linear combinations of another one and, when more than two variables are involved, it is often called multicollinearity. In order to see if there was multicollinearity between the variables, we used the command *vif* (variance inflation factor) only for the significant independent variables. We also used the tolerance, defined as 1/*vif* to check on the degree of collinearity: a 1/*vif* value lower than 0.1 implies that the variable could be considered as a linear combination of other independent variables.
In our case

```
. vif
```

Variable	VIF	1/VIF
year	4.00	0.249767
income	3.62	0.276384
secondary	2.19	0.456569
fatalityrate	2.07	0.483902
speed65	1.88	0.531291
primary	1.79	0.557323
drinkage21	1.34	0.745500
speed70	1.34	0.747371
ba08	1.25	0.801897
age	1.21	0.826042
vmt	1.15	0.866125
Mean VIF	1.99	

Hausman test

```
.       hausman fixed random

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |     fixed        random       Difference          S.E.
-------------+------------------------------------------------------------
    sb_useage |  -.0056444    -.0055468      -.0000975        .0002051
         year |  -.0005287     -.000531       2.27e-06        7.10e-06
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(2) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =        0.85
                Prob>chi2 =      0.6542
```

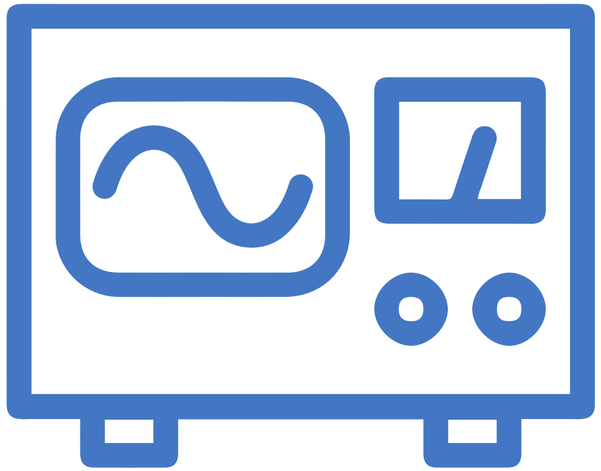
$$H = (\hat{\boldsymbol{\beta}}^{RE} - \hat{\boldsymbol{\beta}}^{FE})'\left[Var(\hat{\boldsymbol{\beta}}^{RE}) - Var(\hat{\boldsymbol{\beta}}^{FE})\right]^{-1}(\hat{\boldsymbol{\beta}}^{RE} - \hat{\boldsymbol{\beta}}^{FE}),$$

In panel data analysis, the Hausman test can help us to choose between fixed-effects model or a random-effects model. The null hypothesis is that the preferred model is random effects; The alternate hypothesis is that the model as fixed effects. Essentially, the tests look to see if there is a correlation between the unique errors and the regressors in the model. The null hypothesis is that there is no correlation between the two.

Interpreting the result from a Hausman test is fairly straightforward: if the p-value is small (less than 0.05), reject the null hypothesis.

Breusch Pagan Test

The Breusch-Pagan-Godfrey Test is a test for heteroskedasticity of errors in a linear regression model. Heteroscedasticity means “differently scattered”; this is opposite to homoscedastic, which means “same scatter.” Homoskedasticity in regression is an important assumption and if it is violated, we can't use a regression analysis. Hence, this test is used in order to understand if the variance of the errors from regression is dependent on the values of the independent variables. So, in this case, there is heteroskedasticity.



```
.           estat hettest, iid rhs

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: sb_useage vmt speed65 speed70 drinkage21 ba08 income age primary secondary year
```

White test

```
.           estat imtest, white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity
```

After evaluating the Breusch-Pagan Test result, the White test was implemented: an extremely common test for heteroskedasticity which begins by allowing the heteroskedasticity process to be a function of one or more of your independent variables. It’s similar to the Breusch-Pagan test, but the White test allows the independent variable to have a nonlinear and interactive effect on the error variance.

4.Results

Starting point

```
. reg fatalityrate sb_useage speed65 speed70 ba08 drinkage21 income age fips year
```

Source	SS	df	MS		
Model	.008046335	9	.000894037	Prob > F	0.0000
Residual	.005993071	546	.000010976	R-squared	0.5731
Total	.014039406	555	.000025296	Root MSE	.00331

Number of obs	=	556				
F(9, 546)	=	81.45				
Prob > F	=	0.0000				
R-squared	=	0.5731				
Adj R-squared	=	0.5661				
Root MSE	=	.00331				

fatalityrate	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
sb_useage	.0031036	.0012101	2.56	0.011	.0007267	.0054805
speed65	.0000384	.0004543	0.08	0.933	-.0008539	.0009308
speed70	.0022043	.000535	4.12	0.000	.0011533	.0032553
ba08	-.0017968	.0004374	-4.11	0.000	-.0026559	-.0009376
drinkage21	-.0012781	.0008483	-1.51	0.132	-.0029445	.0003883
income	-8.83e-07	5.33e-08	-16.55	0.000	-9.87e-07	-7.78e-07
age	.0000289	.0001077	0.27	0.789	-.0001827	.0002405
fips	-.000085	9.51e-06	-8.93	0.000	-.0001036	-.0000663
year	5.06e-06	.0000778	0.06	0.948	-.0001478	.0001579
_cons	.0279582	.153482	0.18	0.856	-.2735294	.3294458

At the beginning of our study, we started with a simple regression with all the variables.

So, we looked at the values and understood there are some problems in this simple OLS model. In fact, this OLS model doesn't take into account the differences between the states. Hence, the *sb_useage* is positive, indicating that the more seatbelts are used the more *fatalityrate* increase.

Obviously, these values are totally wrong and we need to proceed in another way...

Fixing the OLS regression

```
. reg fatalityrate sb_useage speed65 speed70 ba08 drinkage21 income age i.fips i.year

      Source |       SS           df       MS      Number of obs   =       556
-------------+----------------------------------   F(71, 484)      =     70.23
       Model |  .012797208        71  .000180242   Prob > F        =    0.0000
    Residual |  .001242198       484  2.5665e-06   R-squared       =    0.9115
-------------+----------------------------------   Adj R-squared   =    0.8985
       Total |  .014039406       555  .000025296   Root MSE        =     .0016

------------------------------------------------------------------------------
fatalityrate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   sb_useage |  -.0035126   .0011229    -3.13   0.002    -.0057189   -.0013062
     speed65 |  -.0005822   .0004201    -1.39   0.166    -.0014076    .0002433
     speed70 |   .0011277   .0003487     3.23   0.001     .0004426    .0018128
        ba08 |  -.0007522   .0003462    -2.17   0.030    -.0014324   -.0000719
  drinkage21 |  -.0009714   .0005307    -1.83   0.068    -.0020142    .0000714
      income |   5.03e-07   1.45e-07     3.47   0.001     2.18e-07    7.87e-07
         age |   .0017208    .000396     4.35   0.000     .0009426     .002499
```

This regression starts to fix our problems. In fact, the *sb_useage* is negative, which means that more people are using the seatbelts the low is the fatalityrate.
Anyway, there is a strange value, which is the *income* that has to be modified in order to obtain uniform tables and results. So, we need to generate a new variable: *logincome*

Fixing the OLS regression 2

```
. reg fatalityrate sb_useage speed65 speed70 ba08 drinkage21 logincome age i.fips i.year

      Source |       SS           df       MS      Number of obs   =       556
-------------+----------------------------------   F(71, 484)      =     68.76
       Model |  .01277308         71  .000179903   Prob > F        =    0.0000
    Residual |  .001266326       484  2.6164e-06   R-squared       =    0.9098
-------------+----------------------------------   Adj R-squared   =    0.8966
       Total |  .014039406       555  .000025296   Root MSE        =    .00162

------------------------------------------------------------------------------
fatalityrate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   sb_useage |  -.0037186   .0011328    -3.28   0.001    -.0059445   -.0014926
     speed65 |  -.0007833   .0004241    -1.85   0.065    -.0016166       .00005
     speed70 |   .0008042   .0003402     2.36   0.018     .0001358    .0014725
        ba08 |  -.0008225   .0003516    -2.34   0.020    -.0015134   -.0001316
  drinkage21 |  -.0011337   .0005353    -2.12   0.035    -.0021855   -.0000819
   logincome |   .0062643   .0038683     1.62   0.106    -.0013363     .013865
         age |    .001318   .0003834     3.44   0.001     .0005648    .0020713
```

Adopting the *logincome* in this regression we have a more clear and uniform data. So, this would be our model to start our regression study!

Fixed Effect & Random Effect

```
. xtreg fatalityrate sb_useage speed65 speed70 ba08 drinkage21 logincome age i.fips i.year, fe
```

```
. xtreg fatalityrate sb_useage speed65 speed70 ba08 drinkage21 logincome age i.fips i.year, re
```

```
. estimates store random

.
. *** 4. Comparison between the fixed and the random model
. estimates table fixed random ols, star stats(N r2 r2_a)
```

Variable	fixed	random	ols
sb_useage	-.00371856**	-.00371856**	-.00371856**
speed65	-.00078331	-.00078331	-.00078331
speed70	.00080417*	.00080417*	.00080417*
ba08	-.00082249*	-.00082249*	-.00082249*
drinkage21	-.00113368*	-.00113368*	-.00113368*
logincome	.00626435	.00626435	.00626435
age	.00131804***	.00131804***	.00131804***

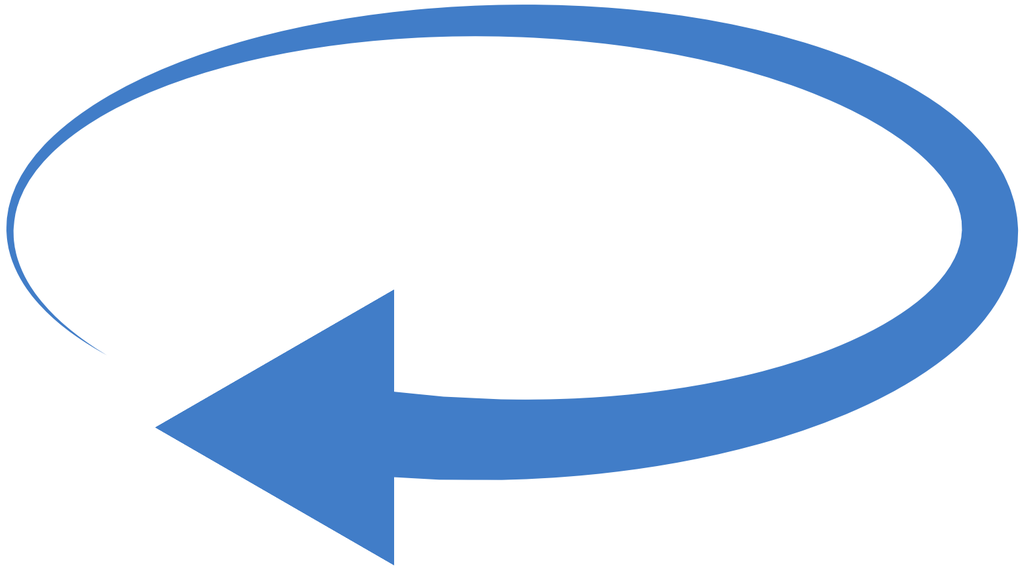
We estimate and store the regressions and compare all our regression models. So, as we can see we obtain the same results for each regression.

Hausman Test

```
. hausman fixed random

Note: the rank of the differenced variance matrix (8) does not equal the number o
      computing the test.  Examine the output of your estimators for anything u
      similar scale.

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |    fixed       random       Difference          S.E.
-------------+------------------------------------------------------------
  sb_useage  |  -.0037186    -.0037186        1.21e-13            .
    speed65  |  -.0007833    -.0007833       -2.41e-13            .
    speed70  |   .0008042     .0008042       -4.21e-14            .
       ba08  |  -.0008225    -.0008225       -1.54e-13            .
 drinkage21  |  -.0011337    -.0011337        1.55e-13            .
  logincome  |   .0062643     .0062643       -8.46e-12            .
        age  |    .001318      .001318       -2.75e-13            .
       year  |
       1984  |  -.0004319    -.0004319        8.55e-13            .
       1985  |  -.0010707    -.0010707        1.31e-12            .
       1986  |  -.0005777    -.0005777        1.76e-12            .
       1987  |  -.0008722    -.0008722        2.42e-12            .
       1988  |   -.001885     -.001885        3.00e-12            .
       1989  |  -.0041766    -.0041766        3.58e-12            .
       1990  |   -.005266     -.005266        4.08e-12            .
       1991  |  -.0066622    -.0066622        4.35e-12            .
       1992  |   -.008518     -.008518        4.80e-12            .
       1993  |  -.0089399    -.0089399        5.14e-12            .
       1994  |  -.0096297    -.0096297        5.53e-12            .
       1995  |  -.0101123    -.0101123        5.96e-12            .
       1996  |  -.0110766    -.0110766        6.41e-12            .
       1997  |  -.0116075    -.0116075        6.82e-12            .
------------------------------------------------------------------------------
                    b = consistent under Ho and Ha; obtained from xtreg
     B = inconsistent under Ha, efficient under Ho; obtained from xtreg

  Test:  Ho:  difference in coefficients not systematic

              chi2(8) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                      =    -0.00    chi2<0 ==> model fitted on these
                                    data fails to meet the asymptotic
```

This test was done in order to decide if we had to use the random or the Fixed Effect model.

So, due to the result is negative: -0.00 we decided that a simple regression (modified as we did in the previous slides) would bee the one.



Heteroskedasticity

```
.                estat hettest, iid rhs /* ---> reject homoskedasticity because = 0.0000*/

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: sb_useage speed65 speed70 ba08 drinkage21 logincome age 1983b.year 1984.y
                    1996.year 1997.year 1b.fips 2.fips 4.fips 5.fips 6.fips 8.fips 9.fips 10.
                    26.fips 27.fips 28.fips 29.fips 30.fips 31.fips 32.fips 33.fips 34.fips 3
                    50.fips 51.fips 53.fips 54.fips 55.fips 56.fips

         chi2(71)     =   155.16
         Prob > chi2  =   0.0000
```

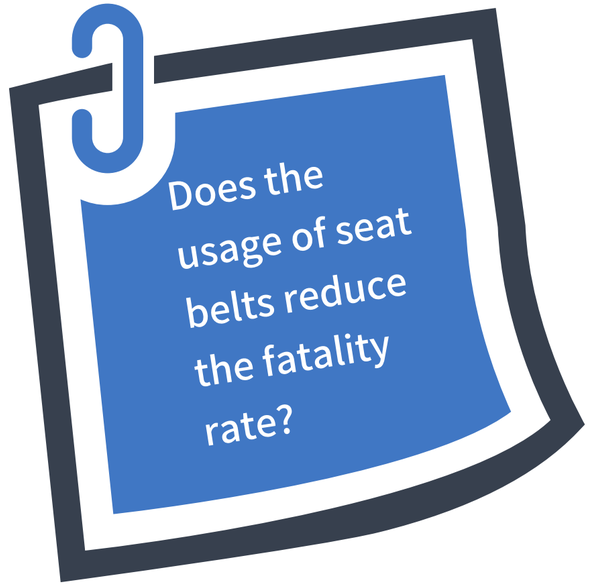
White Breush-Pagan test we checked that the Prob > chi2 = 0.0000
So, the null hypothesis rejected and Heteroskedasticity is assumed, so this means that we have to use the command ***robust*** in our regression!

In the next slides, we studied our regression model adopting all the analysis we did right now. So, using the robust command and also the areg regression which leads us to a more elegant resolution our questions.

1st Question

```
. reg fatalityrate sb_useage speed65 speed70 ba08 drinkage21 logincome age, robust

Linear regression                               Number of obs     =        556
                                                F(7, 548)         =      90.96
                                                Prob > F          =     0.0000
                                                R-squared         =     0.5493
                                                Root MSE          =      .0034

------------------------------------------------------------------------------
             |               Robust
fatalityrate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    sb_useage |   .0040684   .0012323     3.30   0.001     .0016478    .0064889
      speed65 |   .0001479   .0004076     0.36   0.717    -.0006527    .0009486
      speed70 |   .0024045   .0004721     5.09   0.000     .0014771    .0033319
         ba08 |  -.0019246   .0003612    -5.33   0.000     -.002634   -.0012151
   drinkage21 |   .0000799   .0009872     0.08   0.936    -.0018593     .002019
    logincome |  -.0181444    .001086   -16.71   0.000    -.0202776   -.0160111
          age |  -7.22e-06   .0001644    -0.04   0.965    -.0003302    .0003158
        _cons |   .1965469   .0092503    21.25   0.000     .1783766    .2147172
------------------------------------------------------------------------------
```

We used the *robust* command in because as we will see later on we are in presence of heteroskedasticity.



1. In the previous analysis we did a log transformation to the *income* variable in order to obtained less skewed distribution. Helping us making patterns in the data more interpretable and for helping to meet the assumptions of inferential statistics.
2. So, the *sb_usage* coefficient = .0040684 lead to the assumption that an increase in this variable lead an increase also in the *fatalityrate*

Determine Robust Std. Err.

```
. reg fatalityrate sb_useage speed65 speed70 ba08 drinkage21 logincome age i.fips, cluster(fips)

Linear regression                               Number of obs     =        556
                                                F(6, 50)          =          .
                                                Prob > F          =          .
                                                R-squared         =     0.8867
                                                Root MSE          =    .00179

                                  (Std. Err. adjusted for 51 clusters in fips)
------------------------------------------------------------------------------
             |               Robust
fatalityrate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    sb_useage |  -.0057748   .001751    -3.30   0.002    -.0092919   -.0022577
      speed65 |   -.000425   .0004778   -0.89   0.378    -.0013847    .0005346
      speed70 |   .0012333   .0003654    3.38   0.001     .0004994    .0019671
         ba08 |  -.0013775   .0003935   -3.50   0.001    -.0021677   -.0005872
   drinkage21 |   .0007453   .0007536    0.99   0.327    -.0007684     .002259
    logincome |  -.0135144   .0025018   -5.40   0.000    -.0185394   -.0084894
          age |   .0009787   .0007826    1.25   0.217    -.0005933    .0025507
              |
         fips |
            2 |   .0100326   .0046542    2.16   0.036     .0006843    .0193809
            4 |   .0022479   .0006823    3.29   0.002     .0008774    .0036184
            5 |   .0012263   .0006766    1.81   0.076    -.0001327    .0025854
            6 |   .0015822   .0019286    0.82   0.416    -.0022914    .0054559
            8 |   -.000033   .0015549   -0.02   0.983    -.0031561    .0030901
------------------------------------------------------------------------------
```

reg fatalityrate sb_useage speed65 speed70 ba08 drinkage21 logincome age i.fips, robust

In this second regression, we estimated the Robust Standard Error for all the variables and checking for all the US States in the dataset. So, we are checking for those value in order to obtain unbiased std. errors of OLS coefficients under Heteroskedasticity (that we are confirming later in this paper).

But, due to this regression suffers from omitted variables bias, we need to add the year

Don't estimate for all alphas

```
. areg fatalityrate sb_useage speed65 speed70 ba08 drinkage21 logincome age, cluster(fips) absorb(fips)

Linear regression, absorbing indicators         Number of obs     =        556
Absorbed variable: fips                         No. of categories =         51
                                                F(   7,     50)   =      87.90
                                                Prob > F          =     0.0000
                                                R-squared         =     0.8867
                                                Adj R-squared     =     0.8737
                                                Root MSE          =     0.0018

                                (Std. Err. adjusted for 51 clusters in fips)
------------------------------------------------------------------------------
             |               Robust
fatalityrate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   sb_useage |  -.0057748    .001751    -3.30   0.002    -.0092919   -.0022577
     speed65 |   -.000425   .0004778    -0.89   0.378    -.0013847    .0005346
     speed70 |   .0012333   .0003654     3.38   0.001     .0004994    .0019671
        ba08 |  -.0013775   .0003935    -3.50   0.001    -.0021677   -.0005872
  drinkage21 |   .0007453   .0007536     0.99   0.327    -.0007684     .002259
   logincome |  -.0135144   .0025018    -5.40   0.000    -.0185394   -.0084894
         age |   .0009787   .0007826     1.25   0.217    -.0005933    .0025507
       _cons |   .1209958   .0193262     6.26   0.000     .082178    .1598137
------------------------------------------------------------------------------
```

With this command, we can easily see that the *sb_useage* leads to a decrease in the *fatalityrate.* Moreover, this variable is significant for our study.
Anyway, checking the first slides we understood that there is also here something strange. Because the more dangerous states with the *fatalityrate* value very high, also the *sb_useage* is high.

So, this regression suffers from omitted variables bias, and we will solve this in the further slides...

Reg with all the variables

```
. reg fatalityrate sb_useage speed65 speed70 ba08 drinkage21 logincome age i.fips i.year, cluster(fips)

Linear regression                               Number of obs     =        556
                                                F(20, 50)         =          .
                                                Prob > F          =          .
                                                R-squared         =     0.9098
                                                Root MSE          =     .00162

                                   (Std. Err. adjusted for 51 clusters in fips)
------------------------------------------------------------------------------
             |               Robust
fatalityrate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    sb_useage |  -.0037186   .0015246    -2.44   0.018    -.0067808   -.0006563
      speed65 |  -.0007833   .0006093    -1.29   0.205    -.0020071    .0004405
      speed70 |   .0008042   .0004803     1.67   0.100    -.0001605    .0017688
         ba08 |  -.0008225   .0004656    -1.77   0.083    -.0017577    .0001127
   drinkage21 |  -.0011337   .0006534    -1.74   0.089    -.0024461    .0001787
    logincome |   .0062643   .0070367     0.89   0.378    -.0078693    .0203979
          age |    .001318   .0007287     1.81   0.076    -.0001455    .0027816
```

As we can see here we solved the problem of omitted variables, adding the i.fips and the i.year.

In the do-file il possible to check for all the years and the fips.

areg Regression

```
. areg fatalityrate sb_useage speed65 speed70 ba08 drinkage21 logincome age i.year, cluster(fips) absorb(fips)

Linear regression, absorbing indicators         Number of obs     =        556
Absorbed variable: fips                         No. of categories =         51
                                                F(  21,     50)   =      47.41
                                                Prob > F          =     0.0000
                                                R-squared         =     0.9098
                                                Adj R-squared     =     0.8966
                                                Root MSE          =     0.0016

                                (Std. Err. adjusted for 51 clusters in fips)
------------------------------------------------------------------------------
             |               Robust
fatalityrate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   sb_useage |  -.0037186   .0015246    -2.44   0.018    -.0067808   -.0006563
     speed65 |  -.0007833   .0006093    -1.29   0.205    -.0020071    .0004405
     speed70 |   .0008042   .0004803     1.67   0.100    -.0001605    .0017688
        ba08 |  -.0008225   .0004656    -1.77   0.083    -.0017577    .0001127
  drinkage21 |  -.0011337   .0006534    -1.74   0.089    -.0024461    .0001787
   logincome |   .0062643   .0070367     0.89   0.378    -.0078693    .0203979
         age |    .001318   .0007287     1.81   0.076    -.0001455    .0027816
             |
        year |
        1984 |  -.0004319   .0014475    -0.30   0.767    -.0033392    .0024754
        1985 |  -.0010707    .001853    -0.58   0.566    -.0047925    .0026512
        1986 |  -.0005777    .002109    -0.27   0.785    -.0048138    .0036583
        1987 |  -.0008722   .0026195    -0.33   0.741    -.0061336    .0043892
        1988 |   -.001885   .0030219    -0.62   0.536    -.0079547    .0041847
        1989 |  -.0041766   .0034205    -1.22   0.228    -.0110468    .0026936
        1990 |   -.005266   .0037186    -1.42   0.163     -.012735    .0022031
        1991 |  -.0066622   .0039487    -1.69   0.098    -.0145935     .001269
        1992 |   -.008518   .0041863    -2.03   0.047    -.0169265   -.0001095
        1993 |  -.0089399   .0044105    -2.03   0.048    -.0177988   -.0000811
        1994 |  -.0096297   .0048249    -2.00   0.051    -.0193207    .0000613
        1995 |  -.0101123   .0051428    -1.97   0.055    -.0204419    .0002172
        1996 |  -.0110766   .0054713    -2.02   0.048     -.022066   -.0000871
        1997 |  -.0116075   .0058129    -2.00   0.051    -.0232831    .0000681
             |
       _cons |  -.0779904   .0697046    -1.12   0.269    -.2179962    .0620155
------------------------------------------------------------------------------
```

Notes:

- **areg** is designed for datasets with many groups, but not a number of groups that increases with the sample size.
- **absorb(varname)** specifies the categorical variable, which is to be included in the regression as if it were specified by dummy variables. absorb() is required.
- **Cluster** generate command produces grouping variables after hierarchical clustering; see [MV] cluster generate. These variables can then be used in other Stata commands, such as those that tabulate, summarize, and provide graphs. For instance, you might use a cluster generate to create a grouping variable.

As we can see, if we add in our regression the *fips* (also in the previous slide), the coeff. value of the *sb_usage* leads to a reduction. Indicating that more dangerous is the State (higher *fatalityrate*) more high is the value of people using seatbelts (*sb_usage).* Hence, this suffers from OVB (omitted-variable bias). Which is in our case wher our model leves out one or more relevant variables (*fips)*.

VIF,multicollinearity

. vif

Variable	VIF	1/VIF
sb_useage	2.36	0.424010
speed65	2.41	0.415267
speed70	1.49	0.671406
ba08	1.14	0.876287
drinkage21	1.81	0.553564
logincome	3.17	0.315279
age	1.22	0.819172
year		
1984	3.63	0.275200
1985	8.20	0.121988
1986	10.68	0.093601
1987	11.89	0.084092
1988	12.62	0.079215
1989	13.10	0.076358
1990	20.93	0.047776
1991	21.42	0.046695
1992	21.82	0.045830
1993	22.13	0.045183
1994	22.50	0.044454
1995	23.34	0.042844
1996	23.67	0.042244
1997	24.07	0.041550
Mean VIF	12.08	

We use this command after doing the regression to check multicollinearity.

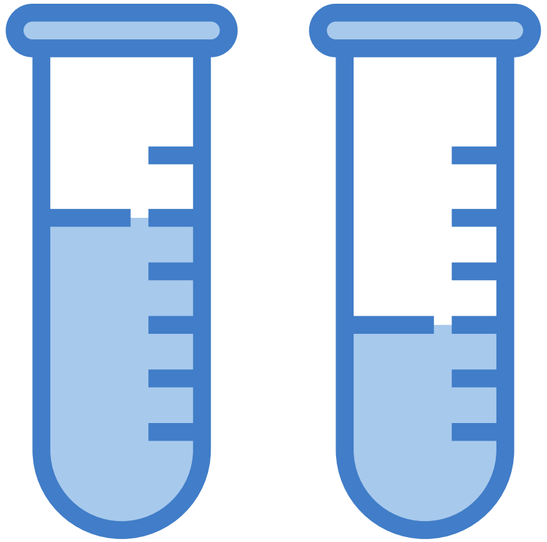
vif= *variance inflation factor.*

If it is ≥ 10 this value for the variable need further investigations. Even if tolerance (1/VIF) is 0.42, which is higher than 0.1. Therefore, it can be assumed that there is no collinearity.

The tolerance value indicates to check on the degree of collinearity.

Breusch-Pagan & Test Param for Years

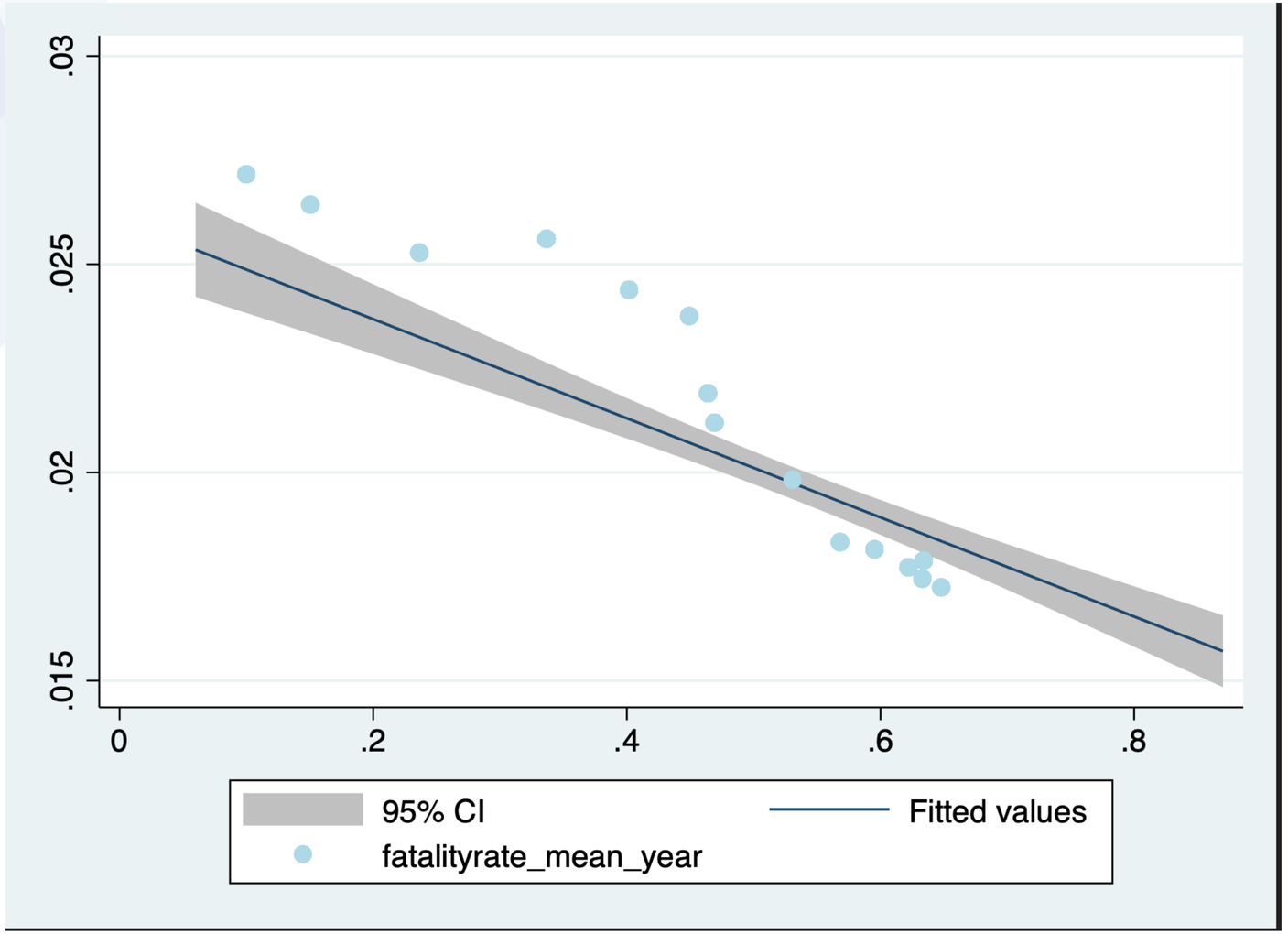
The Breusch-Pagan Test is obiously the same as before, so we did not reported here again, but you can find it in the .do file.



This test is a join test to see if the dummies for all years are equal to 0. If they are equal to 0 then, no time fixed effect is needed.
So, to do this we launched the command: *testparm i.year* after running the regression.

```
( 1)  1984.year = 0
( 2)  1985.year = 0
( 3)  1986.year = 0
( 4)  1987.year = 0
( 5)  1988.year = 0
( 6)  1989.year = 0
( 7)  1990.year = 0
( 8)  1991.year = 0
( 9)  1992.year = 0
(10)  1993.year = 0
(11)  1994.year = 0
(12)  1995.year = 0
(13)  1996.year = 0
(14)  1997.year = 0

      F( 14,    50) =    8.90
           Prob > F =    0.0000
```

Fatalityrate and sb_useage



We decided to show this graph only in order to understand that the *fatalityrate* decreased with higher values of *sb_usage* in the xline.

This command could also be run at the beginning in order to have a better understanding that these two variables are strictly correlated, and in the regression, the *Coefficient sb_useage* must be negative.

2nd Question

```
. reg sb_useage primary secondary speed65 speed70 ba08 drinkage logincome age i.year, cluster(fips) absorb(fips)

Linear regression, absorbing indicators         Number of obs     =        556
                                                F(21, 50)         =          .
                                                Prob > F          =          .
                                                R-squared         =     0.9016
                                                Adj R-squared     =     0.8869
                                                Root MSE          =     .05722

                                  (Std. Err. adjusted for 51 clusters in fips)
------------------------------------------------------------------------------
             |               Robust
   sb_useage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     primary |   .2055968   .0243489     8.44   0.000     .1566907     .254503
   secondary |   .1085184   .0140858     7.70   0.000     .0802262    .1368106
     speed65 |   .0228485   .0215529     1.06   0.294    -.0204417    .0661388
     speed70 |   .0120424   .0216313     0.56   0.580    -.0314054    .0554902
        ba08 |   .0037584    .018507     0.20   0.840     -.033414    .0409307
  drinkage21 |   .0107149   .0285425     0.38   0.709    -.0466144    .0680442
   logincome |   .0582708    .269387     0.22   0.830     -.482809    .5993506
         age |   .0138232   .0243016     0.57   0.572     -.034988    .0626345
             |
        year |
        1984 |   .0041178   .0299885     0.14   0.891    -.0561159    .0643514
        1985 |   .0575169   .0452117     1.27   0.209    -.0332935    .1483273
        1986 |   .1073527   .0579004     1.85   0.070    -.0089437     .223649
        1987 |   .1240647   .0810099     1.53   0.132    -.0386485    .2867779
        1988 |   .1390924   .1025384     1.36   0.181    -.0668621    .3450468
        1989 |   .1702325   .1186812     1.43   0.158    -.0681457    .4086106
        1990 |   .1897753   .1358066     1.40   0.168    -.0830004     .462551
        1991 |   .2370697    .143986     1.65   0.106    -.0521347    .5262741
        1992 |   .2633971   .1598977     1.65   0.106     -.057767    .5845612
        1993 |   .2824192   .1717693     1.64   0.106    -.0625896    .6274279
        1994 |   .2983722   .1826111     1.63   0.109    -.0684131    .6651575
        1995 |   .2959081   .1946357     1.52   0.135    -.0950292    .6868454
        1996 |   .2875641   .2086531     1.38   0.174    -.1315281    .7066562
        1997 |   .2977352   .2209318     1.35   0.184    -.1460193    .7414896
             |
       _cons |  -.893022   2.775641    -0.32   0.749     -6.46806    4.682016
------------------------------------------------------------------------------
```

How the first and secondary enforcement influence the behaviours of the drivers like seat belts usage.

In this second regression, we estimated the Robust Standard Error for all the variables and checking for all the US States in the dataset.

So, we are checking for those value in order to obtain unbiased std. errors of OLS coefficients under Heteroskedasticity (that we are confirming later in this paper).

Areg, another way Robust Std. Err

```
. areg sb_useage primary secondary speed65 speed70 ba08 drinkage logincome age i.year, cluster(fips) absorb(fips)

Linear regression, absorbing indicators         Number of obs     =        556
Absorbed variable: fips                         No. of categories =         51
                                                F(  22,     50)   =     456.16
                                                Prob > F          =     0.0000
                                                R-squared         =     0.9016
                                                Adj R-squared     =     0.8869
                                                Root MSE          =     0.0572

                                  (Std. Err. adjusted for 51 clusters in fips)
------------------------------------------------------------------------------
             |               Robust
   sb_useage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     primary |   .2055968   .0243489     8.44   0.000     .1566907     .254503
   secondary |   .1085184   .0140858     7.70   0.000     .0802262    .1368106
     speed65 |   .0228485   .0215529     1.06   0.294    -.0204417    .0661388
     speed70 |   .0120424   .0216313     0.56   0.580    -.0314054    .0554902
        ba08 |   .0037584    .018507     0.20   0.840     -.033414    .0409307
  drinkage21 |   .0107149   .0285425     0.38   0.709    -.0466144    .0680442
   logincome |   .0582708    .269387     0.22   0.830     -.482809    .5993506
         age |   .0138232   .0243016     0.57   0.572     -.034988    .0626345
             |
        year |
        1984 |   .0041178   .0299885     0.14   0.891    -.0561159    .0643514
        1985 |   .0575169   .0452117     1.27   0.209    -.0332935    .1483273
        1986 |   .1073527   .0579004     1.85   0.070    -.0089437     .223649
        1987 |   .1240647   .0810099     1.53   0.132    -.0386485    .2867779
        1988 |   .1390924   .1025384     1.36   0.181    -.0668621    .3450468
        1989 |   .1702325   .1186812     1.43   0.158    -.0681457    .4086106
        1990 |   .1897753   .1358066     1.40   0.168    -.0830004     .462551
        1991 |   .2370697    .143986     1.65   0.106    -.0521347    .5262741
        1992 |   .2633971   .1598977     1.65   0.106     -.057767    .5845612
        1993 |   .2824192   .1717693     1.64   0.106    -.0625896    .6274279
        1994 |   .2983722   .1826111     1.63   0.109    -.0684131    .6651575
        1995 |   .2959081   .1946357     1.52   0.135    -.0950292    .6868454
        1996 |   .2875641   .2086531     1.38   0.174    -.1315281    .7066562
        1997 |   .2977352   .2209318     1.35   0.184    -.1460193    .7414896
             |
       _cons |  -.893022   2.775641    -0.32   0.749     -6.46806    4.682016
------------------------------------------------------------------------------
```

In this type of analysis is clear that the Coef. for almost every *year* is incrementing its value. So, this means that less usage of seatbelts in a year determine and increase in next year of primary and secondary enforcement.

For instance, in 1996 we registered a reduction of sb_useage led by a reduction of primary and secondary enforcement, that was easily solved in 1997 incrementing the primary and secondary enforcement.

Vif test

. vif

Variable	VIF	1/VIF
primary	1.81	0.552378
secondary	2.18	0.459664
speed65	2.66	0.375333
speed70	1.50	0.667333
ba08	1.20	0.836038
drinkage21	1.81	0.552506
logincome	2.90	0.344575
age	1.22	0.817387
year		
1984	3.63	0.275182
1985	8.18	0.122220
1986	10.64	0.094012
1987	11.86	0.084302
1988	12.61	0.079294
1989	13.02	0.076810
1990	20.80	0.048083
1991	21.15	0.047282
1992	21.49	0.046542
1993	21.75	0.045973
1994	22.10	0.045243
1995	22.98	0.043525
1996	23.38	0.042776
1997	23.77	0.042078
Mean VIF	11.48	

We use this command after doing the regression to check multicollinearity.

vif= *variance inflation factor.*

If it is ≥ 10 this value for the variable need further investigations. Even if the tolerance (1/VIF) is ≤ 0.1 also we need to take a closer look at that.

In this case, VIF=11.48. This represents a high correlation with other independent variables.

However, in this question, we need to understand primary and secondary enforcement and so we do not need other investigations.

moreover, this VIF high value is given by the i.year in the regression formula above.

So, this is another point in our favour to do not investigate anymore in this high value.

. vif

Variable	VIF	1/VIF
secondary	2.04	0.491305
primary	1.67	0.600047
logincome	1.55	0.645752
speed65	1.42	0.704933
drinkage21	1.36	0.733999
age	1.19	0.843634
ba08	1.15	0.867983
speed70	1.12	0.894701
Mean VIF	1.44	

Breusch-Pagan Test

```
.                 estat hettest, iid rhs

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: primary secondary speed65 speed70 ba08 drinkage21 logincome age 1983b.
                    1995.year 1996.year 1997.year 1b.fips 2.fips 4.fips 5.fips 6.fips 8.fi
                    24.fips 25.fips 26.fips 27.fips 28.fips 29.fips 30.fips 31.fips 32.fip
                    48.fips 49.fips 50.fips 51.fips 53.fips 54.fips 55.fips 56.fips

         chi2(72)     =    125.12
         Prob > chi2  =    0.0001
```

This test allows us to check if there is *homoscedasticity* is present in comparison with heteroskedasticity.
Hence, as we can see the P-value is lower than 0.05 and so the null hypothesis of homoskedasticity is rejected and heteroskedasticity assumed.

So, this is why we adopteed the robust regression in the previous regressoin model.

Test Param

```
.                  testparm i.year

 ( 1)  1984.year = 0
 ( 2)  1985.year = 0
 ( 3)  1986.year = 0
 ( 4)  1987.year = 0
 ( 5)  1988.year = 0
 ( 6)  1989.year = 0
 ( 7)  1990.year = 0
 ( 8)  1991.year = 0
 ( 9)  1992.year = 0
 (10)  1993.year = 0
 (11)  1994.year = 0
 (12)  1995.year = 0
 (13)  1996.year = 0
 (14)  1997.year = 0

       F( 14,     50) =    5.55
            Prob > F =    0.0000
```

This test is a join test to see if the dummies for all years are equal to 0. If they are equal to 0 then, no time fixed effect is needed.
So, to do this we launched the command: *testparm i.year* after running the regression.

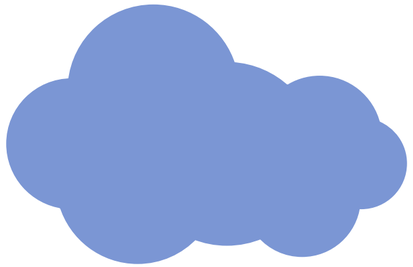
3rd Question

How the alcohol street regulations influence the speed limit and the fatality rate?

We can easily answer this question by looking at the previous regressions and understanding that all wee need to know it is already done by answering question n°1 and 2.

In fact, it's obvious that the alcohol regulation can't affect the speed limit variable due to are both independent one. But it is also clear that both these variables affect the fatality rate. And we can check for these results easily analyzing the previous regressions.

5.Conclusions



Fatality vs. Sb_useage

From this regression, we know that the *fatalityrate* is given by the regression formula:

$$Y_i t = B_1 * X_i t + \alpha_i + u_i t$$

So, *fatalityrate= -.0037186*
The *sb_useage* passed from 6% in 1984 to 87% in 1997. Showing an increase in this value of 81%

So, *fatalityrate* = .0037186 * delta_1984_1997_sbuseage = .00301207 which represent the fatalities permillion traffic miles. So, the life saved from this are: .00301207*mean_vmt = 124.84348

```
. areg fatalityrate sb_useage speed65 speed70 ba08 drinkage21 logincome age i.year, cluster(fips) absorb(fips)

Linear regression, absorbing indicators          Number of obs     =        556
Absorbed variable: fips                          No. of categories =         51
                                                 F(  21,     50)   =      47.41
                                                 Prob > F          =     0.0000
                                                 R-squared         =     0.9098
                                                 Adj R-squared     =     0.8966
                                                 Root MSE          =     0.0016

                                    (Std. Err. adjusted for 51 clusters in fips)
------------------------------------------------------------------------------
             |               Robust
fatalityrate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   sb_useage |  -.0037186   .0015246    -2.44   0.018    -.0067808   -.0006563
     speed65 |  -.0007833   .0006093    -1.29   0.205    -.0020071    .0004405
     speed70 |   .0008042   .0004803     1.67   0.100    -.0001605    .0017688
        ba08 |  -.0008225   .0004656    -1.77   0.083    -.0017577    .0001127
  drinkage21 |  -.0011337   .0006534    -1.74   0.089    -.0024461    .0001787
   logincome |   .0062643   .0070367     0.89   0.378    -.0078693    .0203979
         age |    .001318   .0007287     1.81   0.076    -.0001455    .0027816
```

```
egen max_sb_useage=max(sb_useage)
egen min_sb_useage=min(sb_useage)
gen delta_1984_1997_sbuseage = max_sb_useage-min_sb_useage
egen mean_sb_useage=mean(sb_useage)
egen mean_vmt=mean(vmt)
display max_sb_useage
display min_sb_useage
display mean_sb_useage
display mean_vmt

display .0037186*(delta_1984_1997_sbuseage)
display .00301207*mean_vmt
```

Sb_useage vs. primary and secondary

From this regression, we know that the *sb_useage* is given by the regression formula:

$$Y_i t = B_1 * X_i t + \alpha_i + u_i t$$

So, the coefficient of primary is = *.2055968*
while secondary coef. = 0.1085184

The *sb_useage* change its values because the more is adopted primary, the more people tend to use sbuseage in comparison with the secondary enforcement.
Due to, primary and secondary cannot exist both at the same time, we can conclude by showing the trend of *sb_useage* with the mean achieved by primary and secondary across years. Knowing that a more specific analysis can be done for each state in each year.

sb_useage = .2055968*mean_primary + .1085184*mean_secondary = **.07875683**

Another example for California in 1997 who had only seconday enforcement, so: sb_useage = .2055968*0 + .1085184*1 = .1085184
So, the fatality rate = .0037186 * .1085184 = .00040354
So the lifes prevented = .00040354 * 245259.1 = 98.971857

```
. reg sb_useage primary secondary speed65 speed70 ba08 drinkage logincome age i.year, cluster(fips) absorb(fips)

Linear regression, absorbing indicators         Number of obs     =        556
                                                F(21, 50)         =          .
                                                Prob > F          =          .
                                                R-squared         =     0.9016
                                                Adj R-squared     =     0.8869
                                                Root MSE          =     .05722

                                    (Std. Err. adjusted for 51 clusters in fips)
------------------------------------------------------------------------------
             |               Robust
   sb_useage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     primary |   .2055968   .0243489     8.44   0.000     .1566907    .254503
   secondary |   .1085184   .0140858     7.70   0.000     .0802262   .1368106
     speed65 |   .0228485   .0215529     1.06   0.294    -.0204417   .0661388
     speed70 |   .0120424   .0216313     0.56   0.580    -.0314054   .0554902
        ba08 |   .0037584    .018507     0.20   0.840     -.033414   .0409307
  drinkage21 |   .0107149   .0285425     0.38   0.709    -.0466144   .0680442
   logincome |   .0582708    .269387     0.22   0.830     -.482809   .5993506
         age |   .0138232   .0243016     0.57   0.572     -.034988   .0626345
             |
        year |
        1984 |   .0041178   .0299885     0.14   0.891    -.0561159   .0643514
        1985 |   .0575169   .0452117     1.27   0.209    -.0332935   .1483273
        1986 |   .1073527   .0579004     1.85   0.070    -.0089437    .223649
        1987 |   .1240647   .0810099     1.53   0.132    -.0386485   .2867779
        1988 |   .1390924   .1025384     1.36   0.181    -.0668621   .3450468
        1989 |   .1702325   .1186812     1.43   0.158    -.0681457   .4086106
        1990 |   .1897753   .1358066     1.40   0.168    -.0830004    .462551
        1991 |   .2370697    .143986     1.65   0.106    -.0521347   .5262741
        1992 |   .2633971   .1598977     1.65   0.106     -.057767   .5845612
        1993 |   .2824192   .1717693     1.64   0.106    -.0625896   .6274279
        1994 |   .2983722   .1826111     1.63   0.109    -.0684131   .6651575
        1995 |   .2959081   .1946357     1.52   0.135    -.0950292   .6868454
        1996 |   .2875641   .2086531     1.38   0.174    -.1315281   .7066562
        1997 |   .2977352   .2209318     1.35   0.184    -.1460193   .7414896
             |
       _cons |  -.893022   2.775641    -0.32   0.749     -6.46806    4.682016
------------------------------------------------------------------------------
```

Speed65, Speed70 vs Fatalityrate

So, *speed65*= *-.0007833*
while *speed70* = 0.0008042

The *fatalityrate* change its values because the more is adopted speed65 the more fatality decreases, while the more *speed70* is adopted, the more fatality increases.

Example for California in 1997 who had only *speed65 and speed70* limits, so:
fatality rate = .0037186 * (speed65*1 + speed70*1) = -7.772e-08
So the lifes prevented = 245259.1 * -7.772e-08 = -.01906154

To conclude, we can affirm that in California in the 1997 the speed70 limit generated almost 0.02 deaths in 245259 miles.

```
. areg fatalityrate sb_useage speed65 speed70 ba08 drinkage21 logincome age i.year, cluster(fips) absorb(fips)

Linear regression, absorbing indicators         Number of obs     =        556
Absorbed variable: fips                         No. of categories =         51
                                                F(  21,     50)   =      47.41
                                                Prob > F          =     0.0000
                                                R-squared         =     0.9098
                                                Adj R-squared     =     0.8966
                                                Root MSE          =     0.0016

                                  (Std. Err. adjusted for 51 clusters in fips)
------------------------------------------------------------------------------
             |               Robust
fatalityrate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    sb_useage |  -.0037186   .0015246    -2.44   0.018    -.0067808   -.0006563
      speed65 |  -.0007833   .0006093    -1.29   0.205    -.0020071    .0004405
      speed70 |   .0008042   .0004803     1.67   0.100    -.0001605    .0017688
         ba08 |  -.0008225   .0004656    -1.77   0.083    -.0017577    .0001127
   drinkage21 |  -.0011337   .0006534    -1.74   0.089    -.0024461    .0001787
    logincome |   .0062643   .0070367     0.89   0.378    -.0078693    .0203979
          age |    .001318   .0007287     1.81   0.076    -.0001455    .0027816
```

```
display  .0037186 * (.0007833 + -.0008042)
su year vmt if state == "CA"
display 245259.1 * -7.772e-08 // life prevented with the speed limits per 245259.1 miles
```

Importance of conclusions & limits

All the results reported are very important in order to decided in the future how strategies adopt in order to reduce the most the *fatalityrate* variable
In fact, as we reported in our study we observed that primary enforcement is a best options for states in order to increase the *sb_useage* and also reducing the *fatalityrate* .
Moreover, having a lower speed limit decrease too the *fatalityrate* . So, this means that (in our case California) needs to be more careful adopting speed70 limit.

Obviously the results obtained don't take into account other parameters which might be very important in this type of analysis. For instance the security level of the car.
To sum up, all the result obtained were important in order to reduce deaths, but it would be also important to have more data, taking into account different parameters, like car security level. Because all of us know thtat more technological cars are also more safer that the old ones.

Possibile Extensions

Of course it could be the possibility to extent our study.
For instance, as I said before the increase in security of cars could be new parameter to analyze, or the security of streets, or also the mean wheather in each country indentifying if there is any correlations with fatalities and rainy or snowy States.