Introduction to Difference in Differences Using Stata

Chuck Huber
StataCorp
chuber@stata.com

Oceania Stata Conference 2023
February 9, 2023

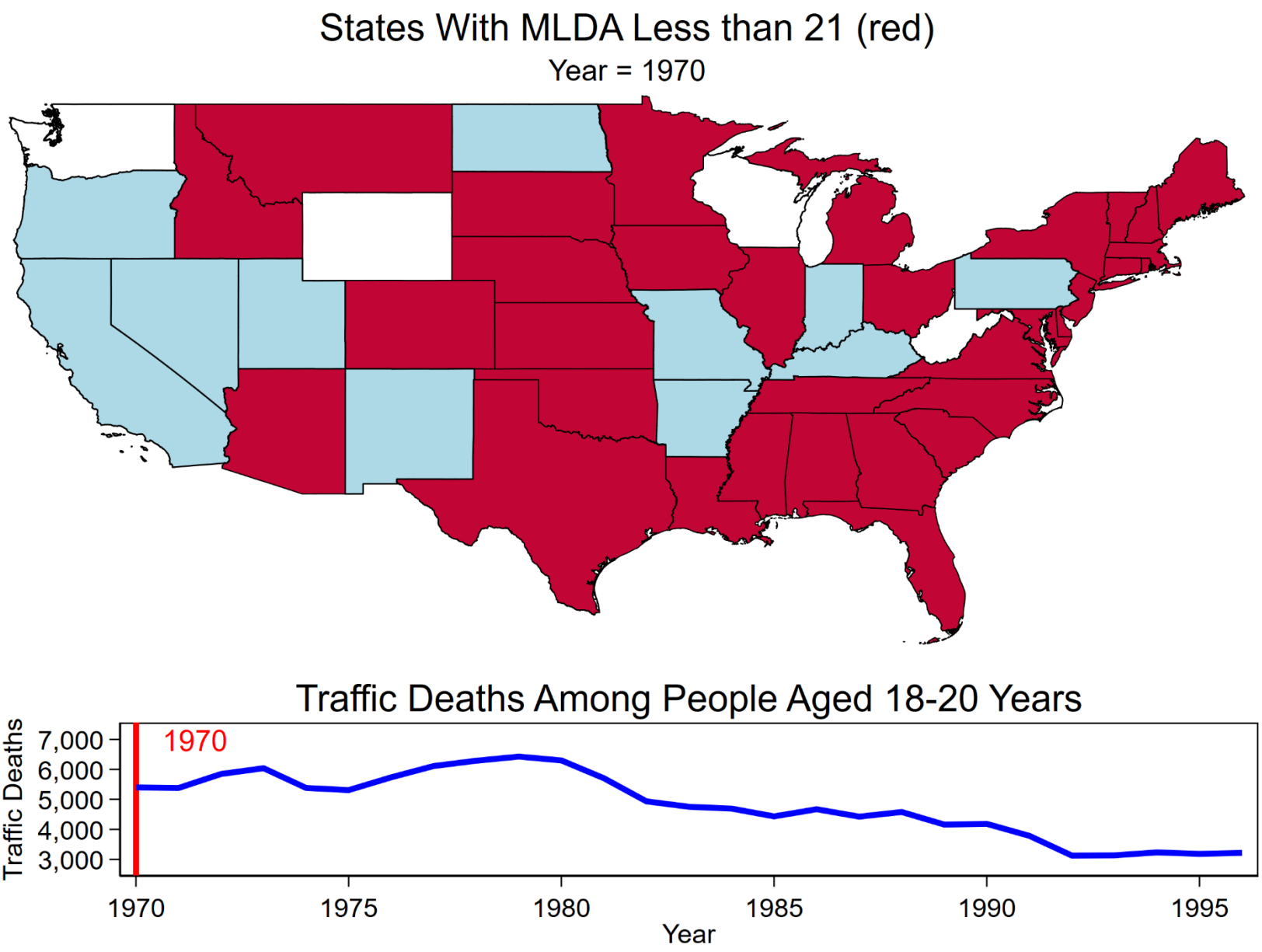
Download Website

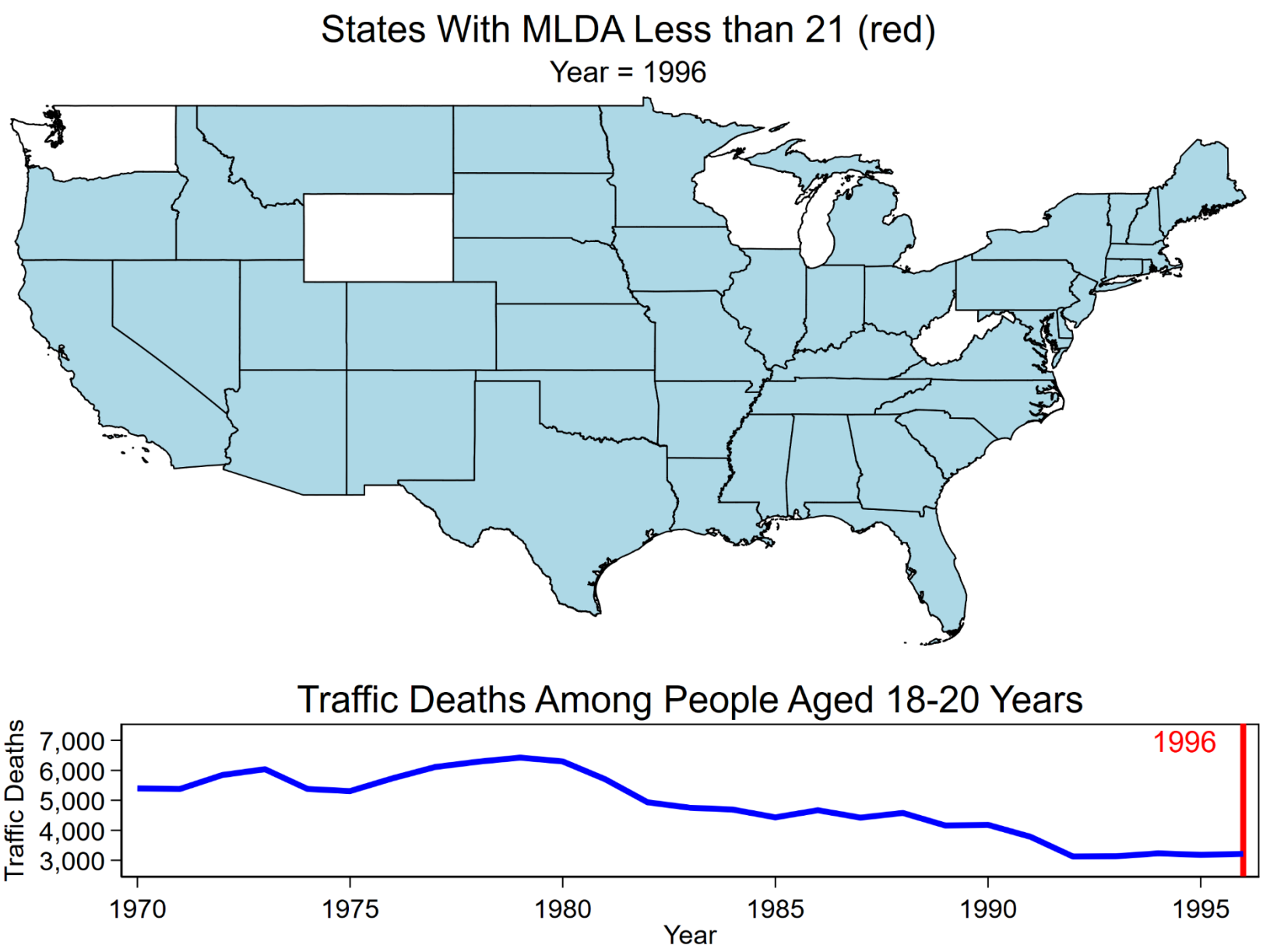
- You can download all of the slides, datasets and do-files here:

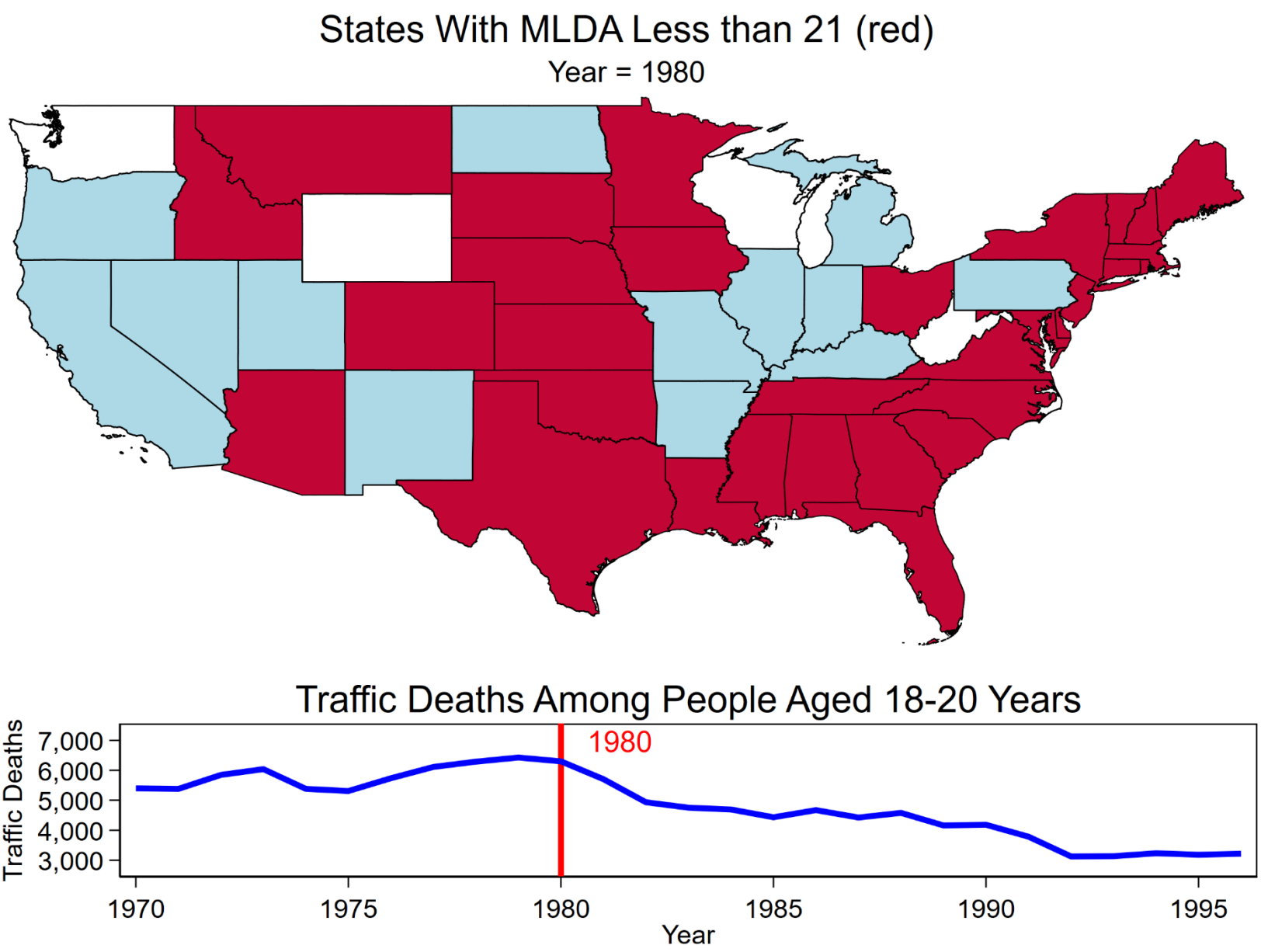
https://tinyurl.com/StataDID

Outline

- The Question
- An Intuitive Introduction
- Two Period, Two Groups Model
- Repeated Cross-Sectional Panel Data
- Longitudinal Panel Data
- More information


States With MLDA Less than 21 (red)
Year = 1970
Traffic Deaths Among People Aged 18-20 Years
Traffic Deaths
7,000
6,000
5,000
4,000
3,000
1970
1970
1975
1980
1985
1990
1995
Year

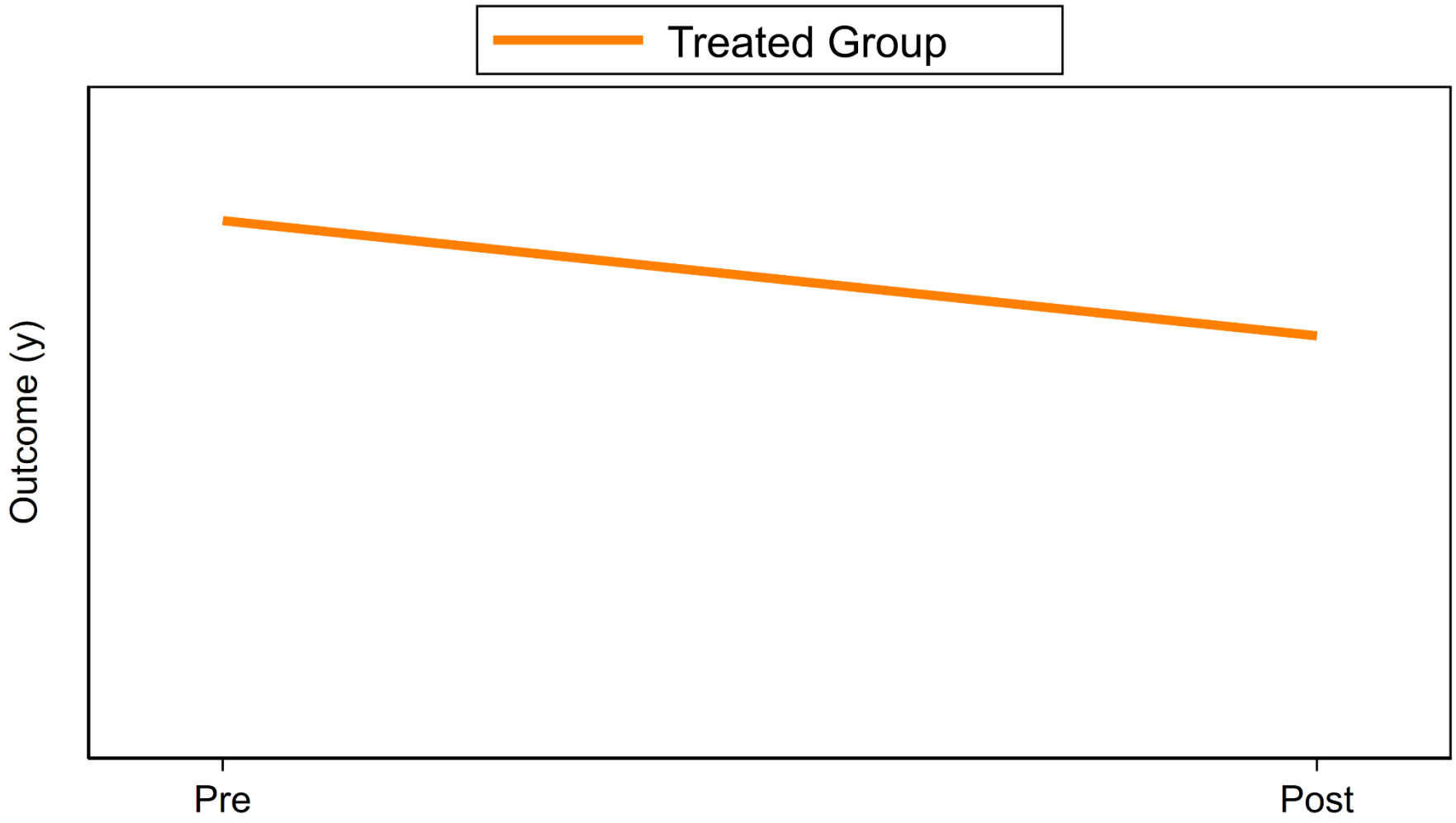

States With MLDA Less than 21 (red)
Year = 1996
Traffic Deaths Among People Aged 18-20 Years
Traffic Deaths
7,000
6,000
5,000
4,000
3,000
1996
1970
1975
1980
1985
1990
1995
Year


States With MLDA Less than 21 (red)
Year = 1980
Traffic Deaths Among People Aged 18-20 Years
Traffic Deaths
7,000
6,000
5,000
4,000
3,000
1980
1970
1975
1980
1985
1990
1995
Year

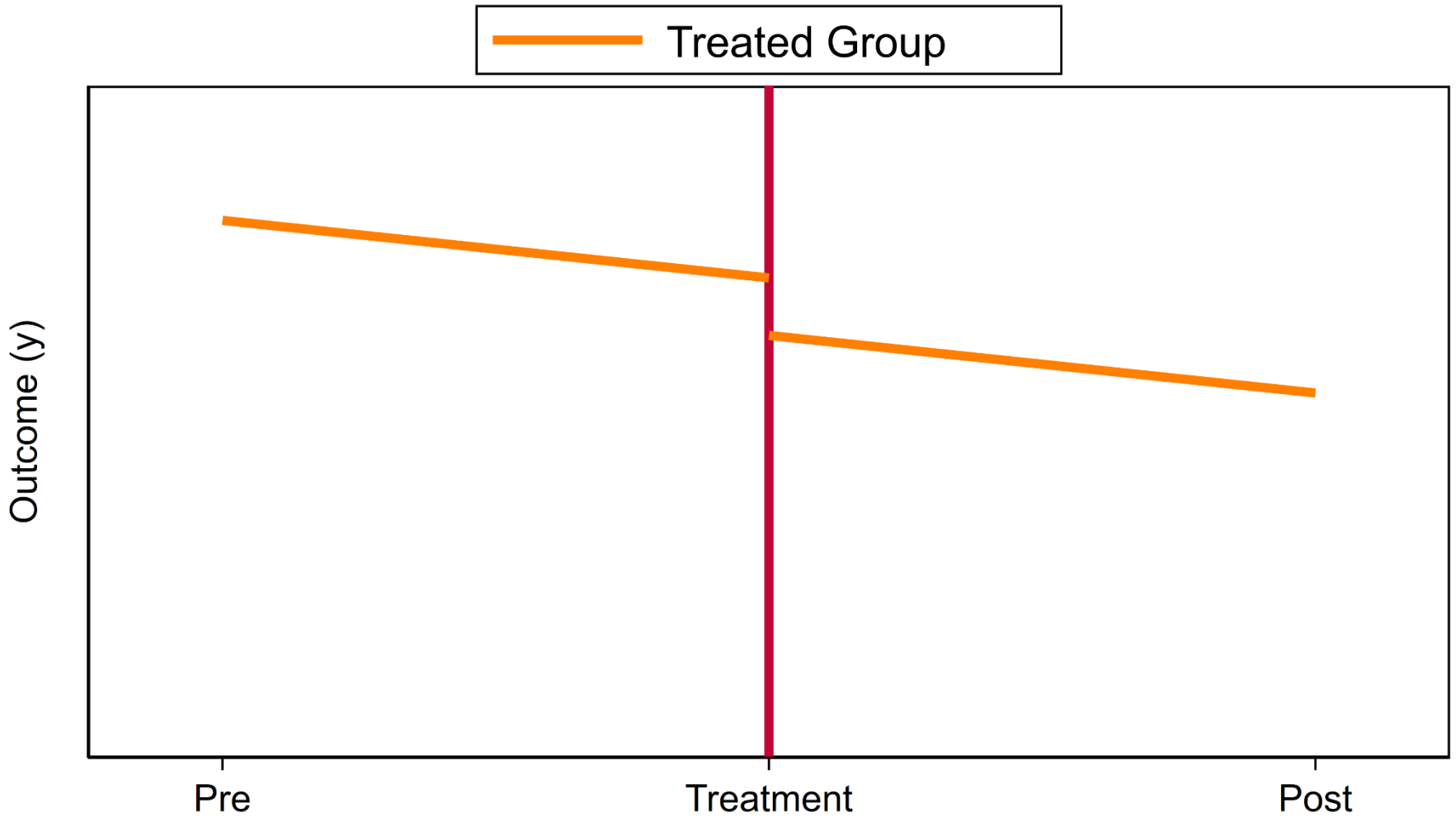
Outline

- The Question
- An Intuitive Introduction
- Two Period, Two Groups Model
- Repeated Cross-Sectional Panel Data
- Longitudinal Panel Data
- More information

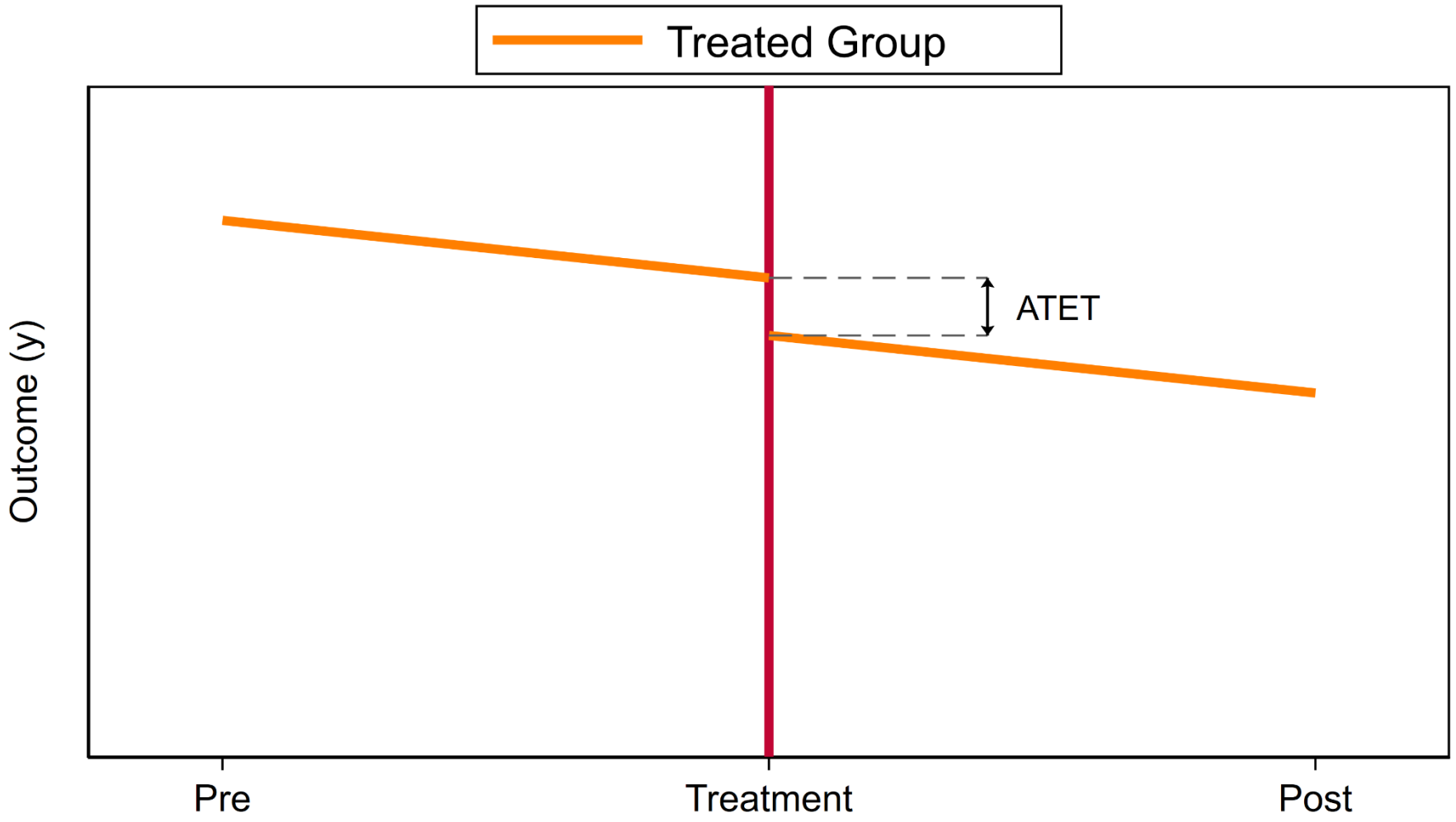
A Conceptual Introduction



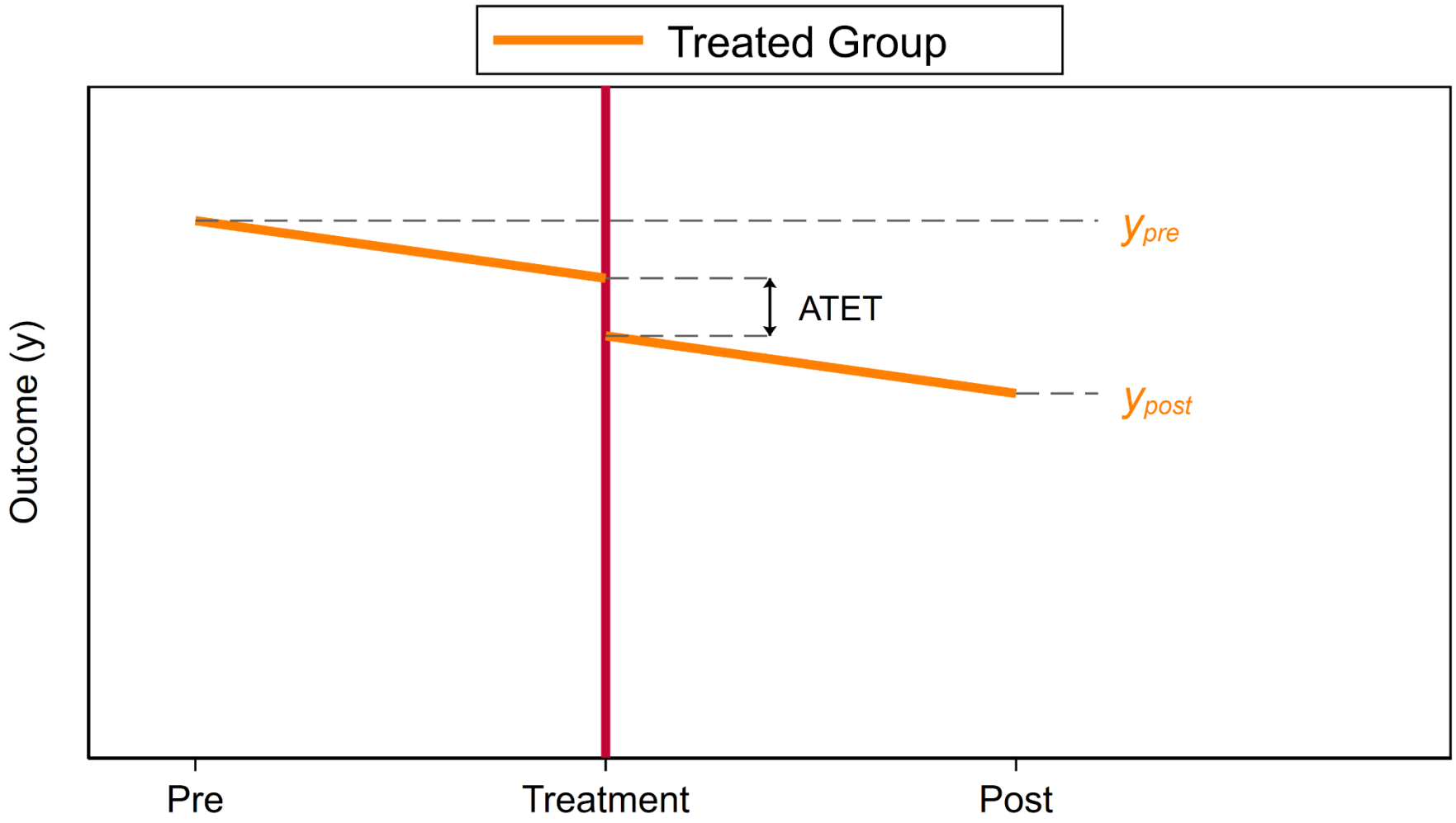
A Conceptual Introduction



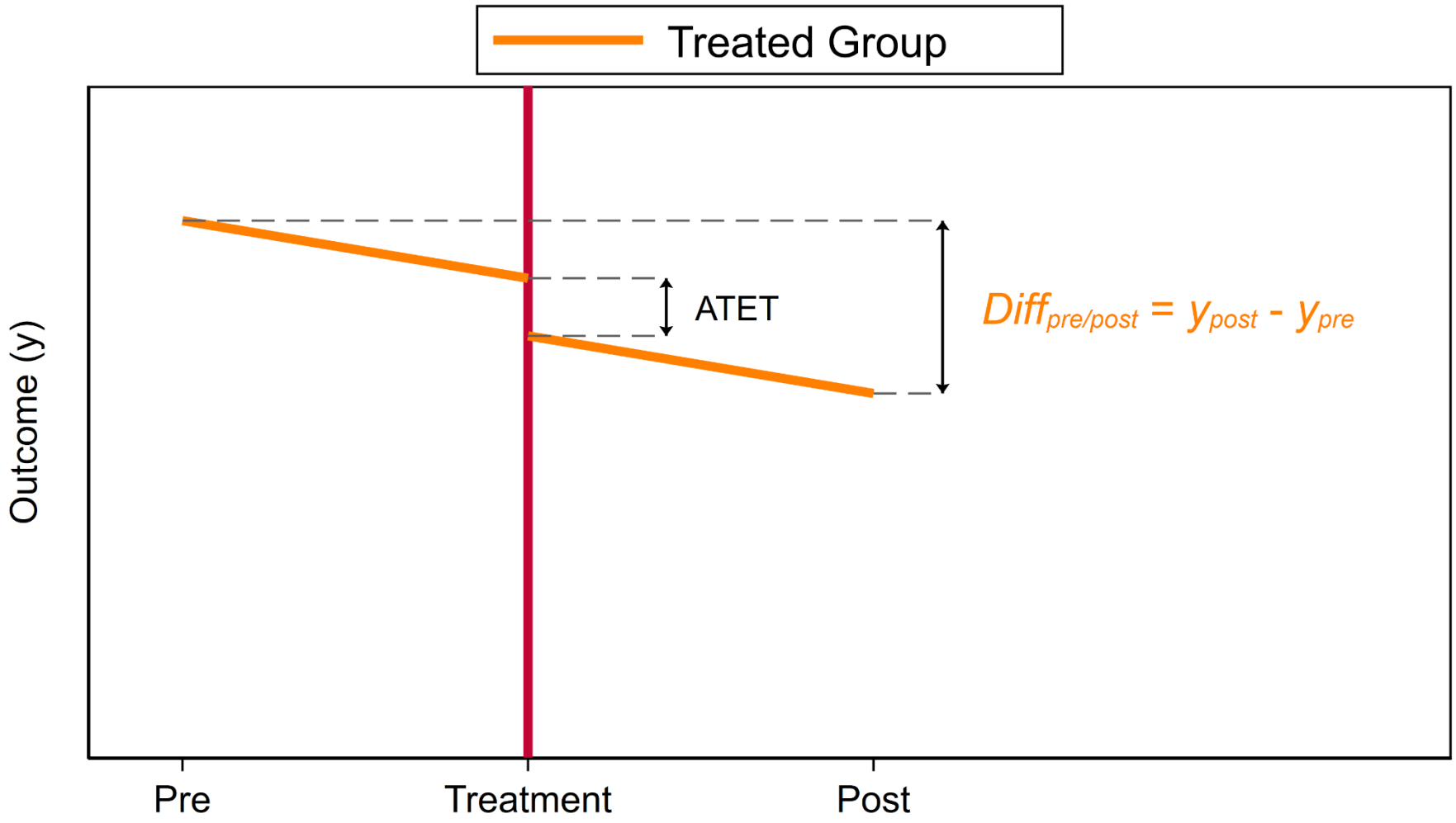
A Conceptual Introduction



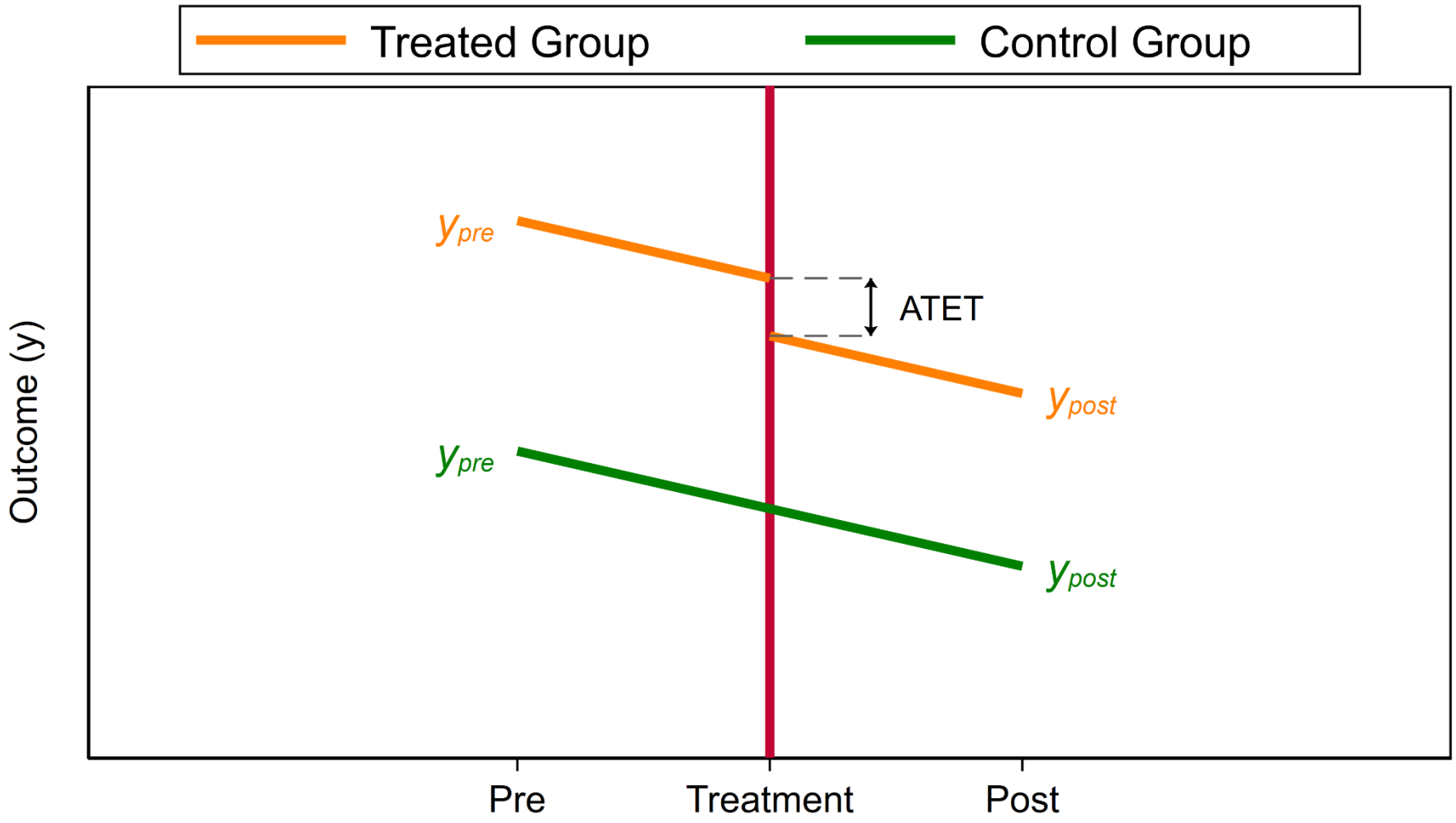
A Conceptual Introduction



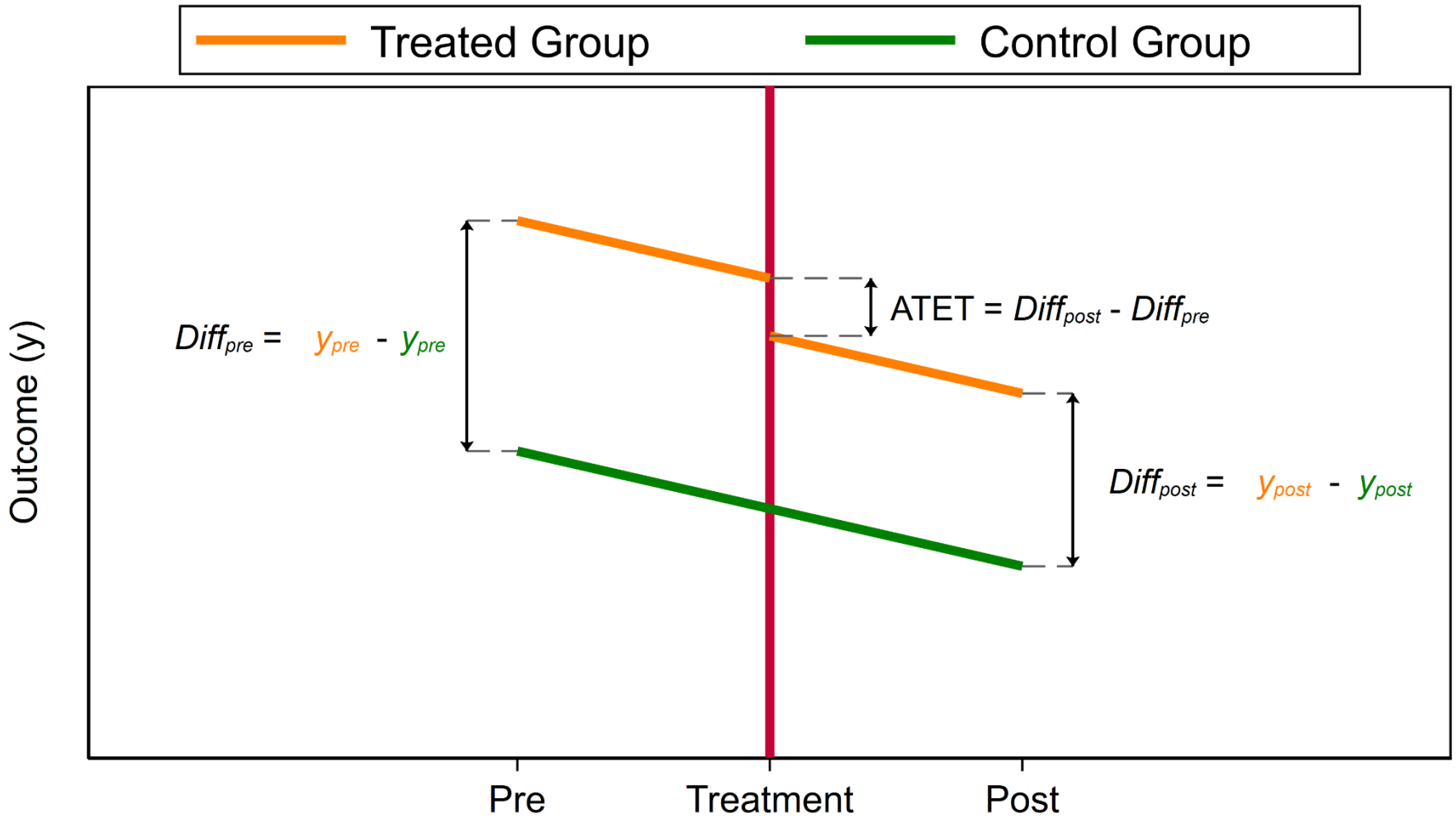
A Conceptual Introduction



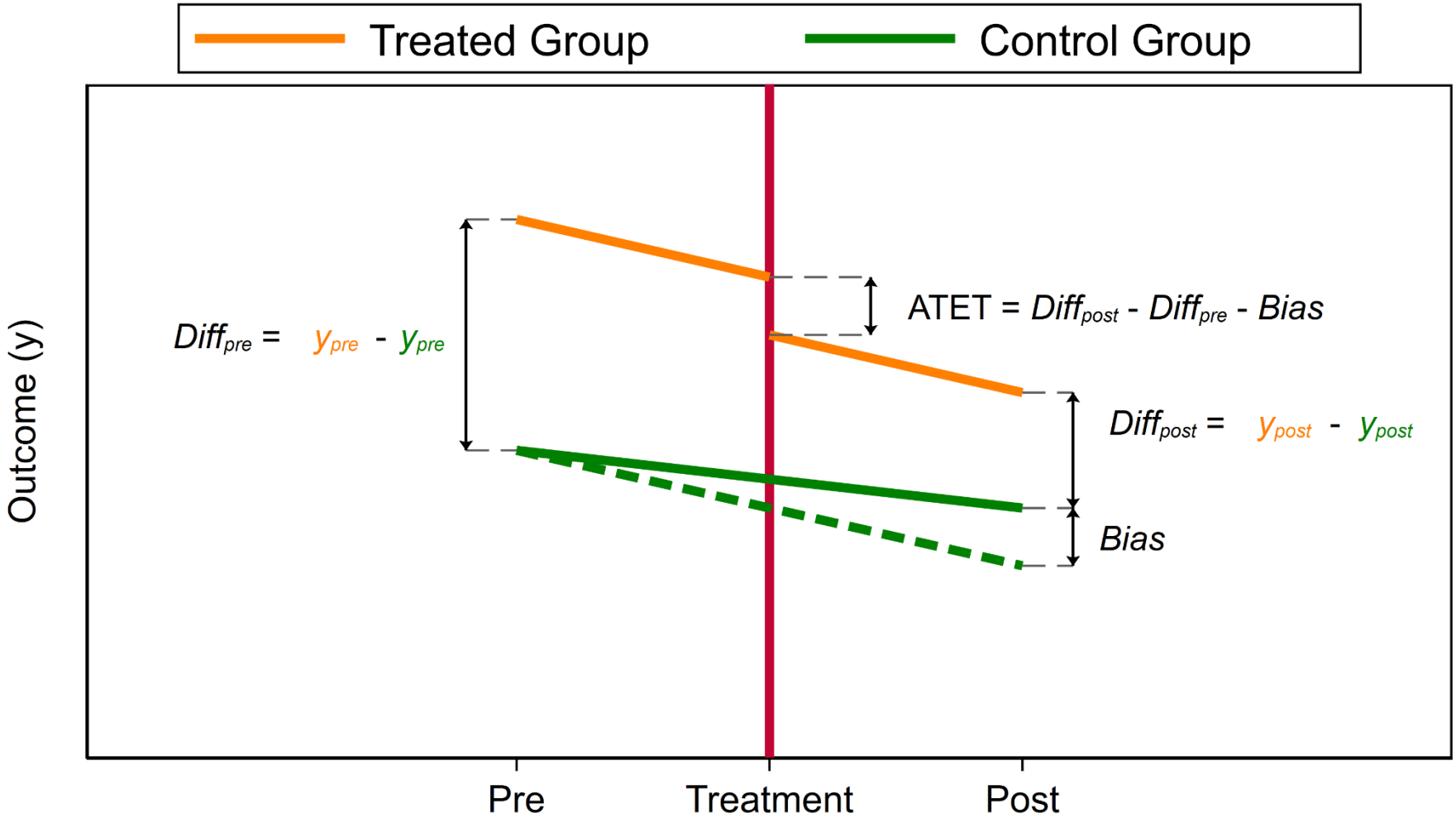
A Conceptual Introduction



A Conceptual Introduction

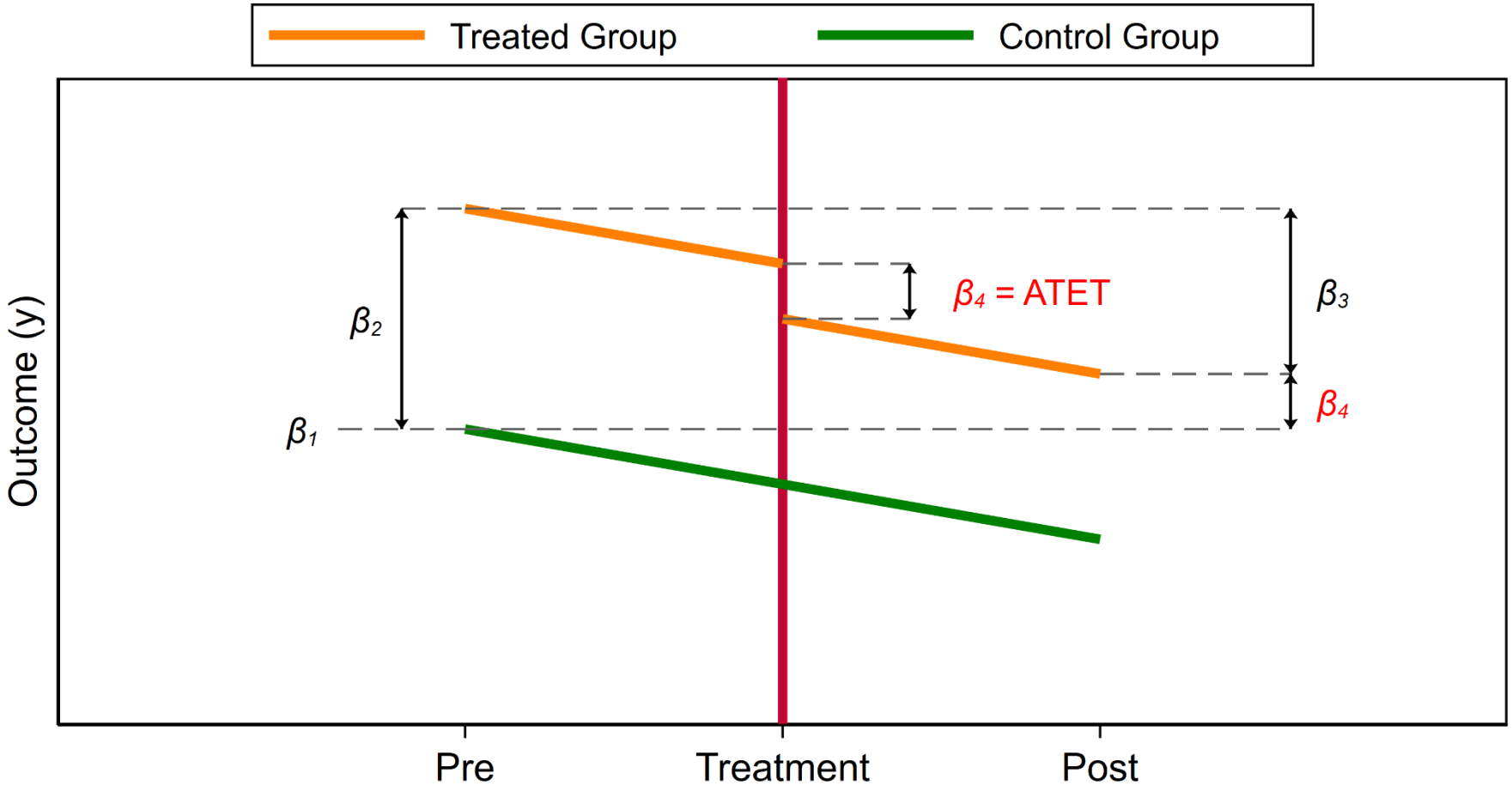


A Conceptual Introduction



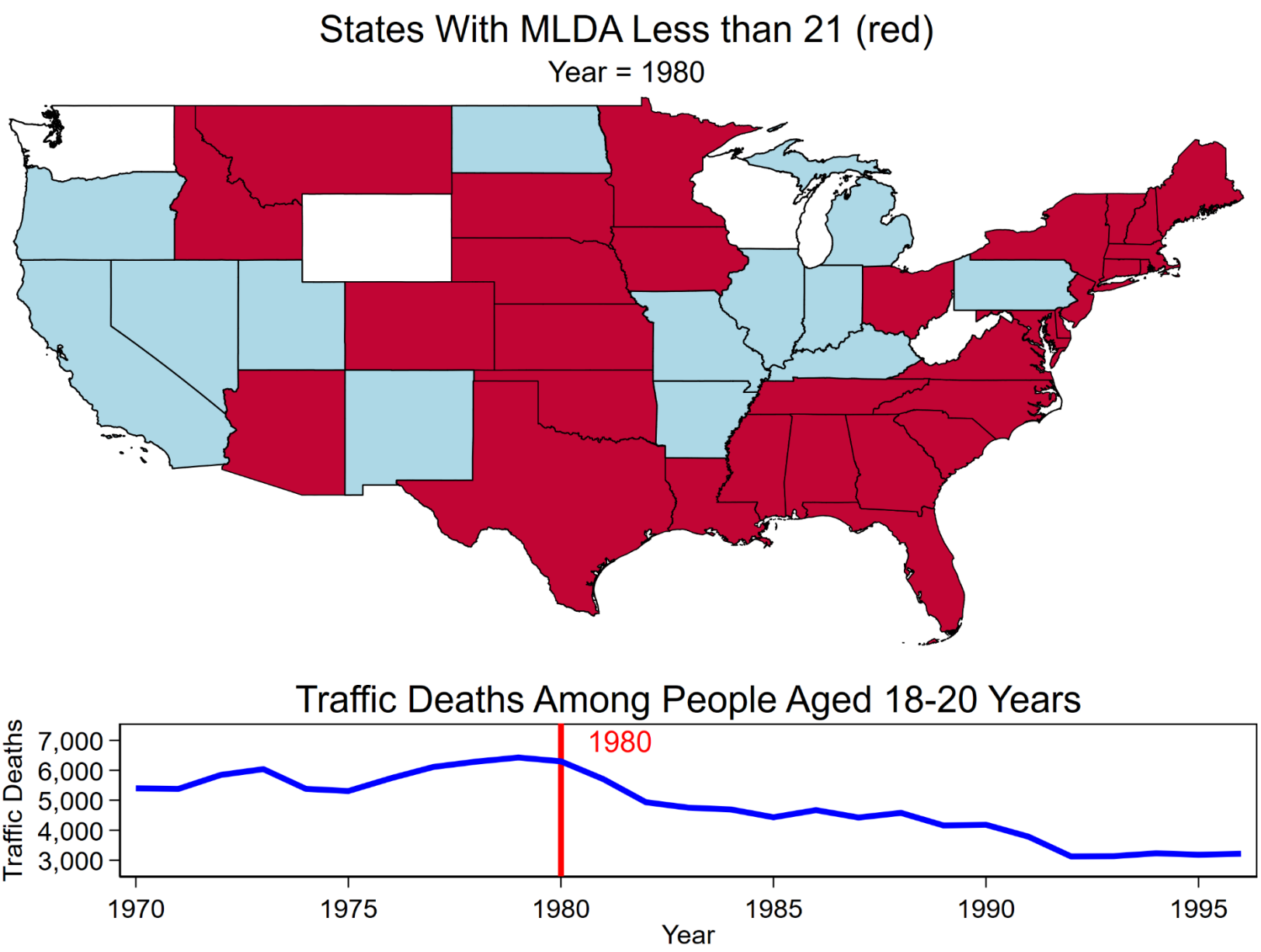
A Conceptual Introduction

$$E(y_{it}|Group, Time) = \beta_1 + \beta_2 Group + \beta_3 Time + \beta_4 GroupxTime$$

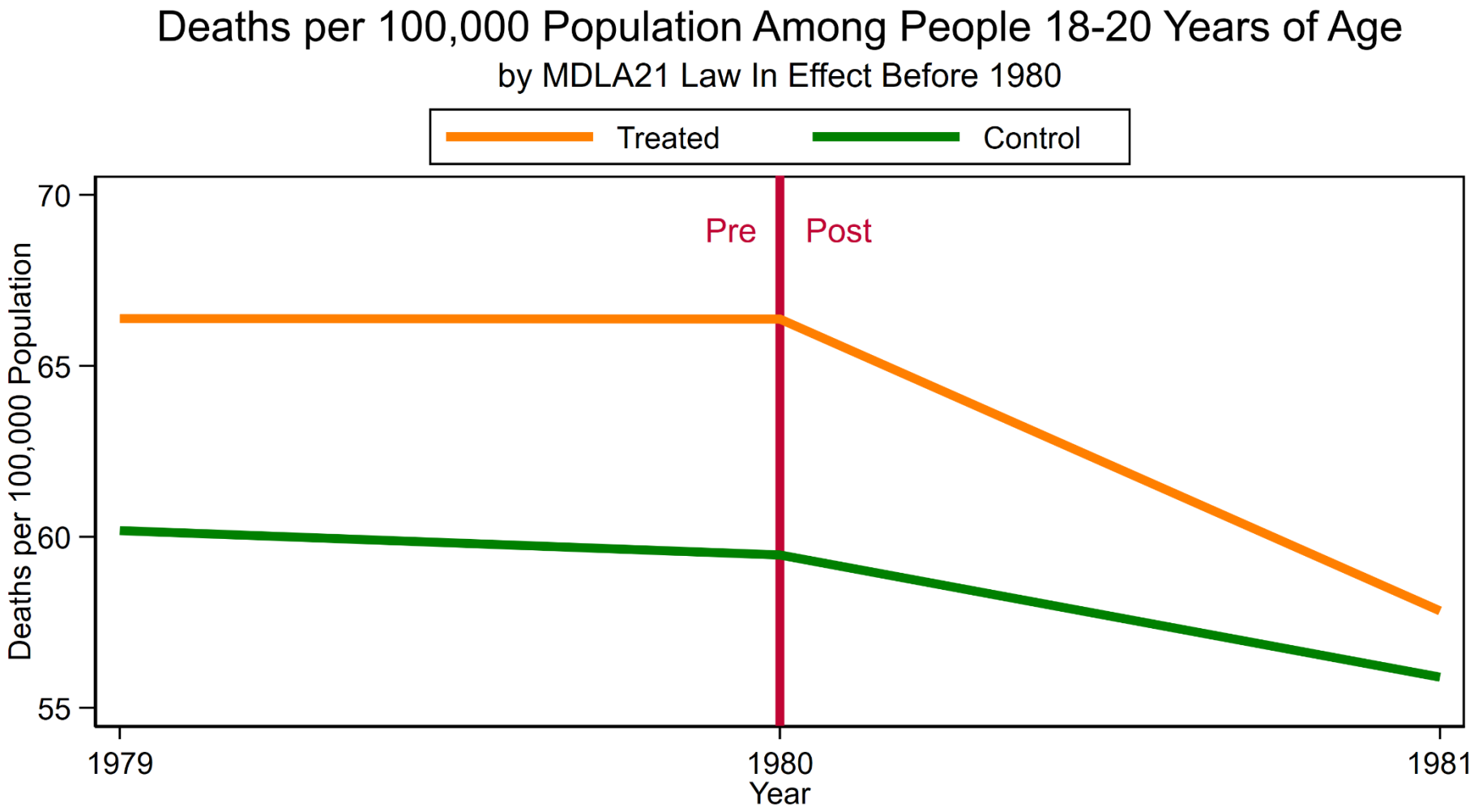


Outline

- ✓ The Question
- ✓ An Intuitive Introduction
- Two Period, Two Groups Model
- Repeated Cross-Sectional Panel Data
- Longitudinal Panel Data
- More information


States With MLDA Less than 21 (red)
Year = 1980
Traffic Deaths Among People Aged 18-20 Years
Traffic Deaths
7,000
6,000
5,000
4,000
3,000
1980
1970
1975
1980
1985
1990
1995
Year

The Two Period, Two Group Model



The Examples Dataset

```
. use MLDA21, clear
(Minimum Legal Drinking Age (MLDA) from Angrist and Pischke, 2014 (type 'notes'))

. describe FIPS state year MLDA21_year MLDA21 mrate_agegrp beertax

Variable      Storage   Display    Value
    name         type    format    label      Variable label
------------------------------------------------------------------------------------------------
FIPS            byte    %10.0g                State ID (Federal Information Processing Standard)
state           str15   %15s                *  State
year            int     %9.0g               *  Year
MLDA21_year     int     %10.0g                 Year in which MLDA21 law was enacted
MLDA21          byte    %15.0g     treat83     Pre- or Post- MLDA21 Law
mrate_agegrp    double  %9.2f               *  Mortality rate (deaths/100000) in the 18-20 age group
beertax         double  %9.2f                  Beer tax
```

The Examples Dataset

```
. summarize FIPS state year MLDA21_year MLDA21 mrate_agegrp beertax
```

Variable	Obs	Mean	Std. dev.	Min	Max
FIPS	1,242	27.13043	14.33362	1	51
state	0				
year	1,242	1983	7.792018	1970	1996
MLDA21_year	1,242	1973.239	21.16342	1933	1988
MLDA21	1,242	.5764895	.4943138	0	1
mrate_agegrp	1,242	52.14309	21.93833	8.231131	189.9421
beertax	1,226	.2367918	.236822	.012747	1.984536

```
. list state MLDA21_year year MLDA21        ///
>      if inlist(FIPS, 17)                  ///
>      , sepby(state) nolabel noobs abbrev(18)
```

state	MLDA21_year	year	MLDA21
Illinois	1980	1970	0
Illinois	1980	1971	0
Illinois	1980	1972	0
Illinois	1980	1973	0
Illinois	1980	1974	0
Illinois	1980	1975	0
Illinois	1980	1976	0
Illinois	1980	1977	0
Illinois	1980	1978	0
Illinois	1980	1979	0
Illinois	1980	1980	1
Illinois	1980	1981	1
Illinois	1980	1982	1
Illinois	1980	1983	1
Illinois	1980	1984	1
Illinois	1980	1985	1
Illinois	1980	1986	1
Illinois	1980	1987	1
Illinois	1980	1988	1
Illinois	1980	1989	1
Illinois	1980	1990	1
Illinois	1980	1991	1
Illinois	1980	1992	1
Illinois	1980	1993	1
Illinois	1980	1994	1
Illinois	1980	1995	1
Illinois	1980	1996	1

```
. // List the data for Iowa
. list state MLDA21_year year MLDA21        ///
>      if inlist(FIPS, 19)                  ///
>      , sepby(state) nolabel noobs abbrev(18)
```

state	MLDA21_year	year	MLDA21
Iowa	1986	1970	0
Iowa	1986	1971	0
Iowa	1986	1972	0
Iowa	1986	1973	0
Iowa	1986	1974	0
Iowa	1986	1975	0
Iowa	1986	1976	0
Iowa	1986	1977	0
Iowa	1986	1978	0
Iowa	1986	1979	0
Iowa	1986	1980	0
Iowa	1986	1981	0
Iowa	1986	1982	0
Iowa	1986	1983	0
Iowa	1986	1984	0
Iowa	1986	1985	0
Iowa	1986	1986	1
Iowa	1986	1987	1
Iowa	1986	1988	1
Iowa	1986	1989	1
Iowa	1986	1990	1
Iowa	1986	1991	1
Iowa	1986	1992	1
Iowa	1986	1993	1
Iowa	1986	1994	1
Iowa	1986	1995	1
Iowa	1986	1996	1

Setting Up The Data

```
. keep if inrange(year, 1979, 1981)
(1,104 observations deleted)
```

Keep the data for 1979, 1980, and 1981

```
.
. list state MLDA21_year year MLDA21                ///
>      if year!=1980 & inlist(FIPS, 17,18,19)       ///
>      , sepby(state) nolabel noobs abbrev(18)
```

state	MLDA21_year	year	MLDA21
Illinois	1980	1979	0
Illinois	1980	1981	1
Indiana	1934	1979	1
Indiana	1934	1981	1
Iowa	1986	1979	0
Iowa	1986	1981	0

The law went into effect during the time interval

The law was in effect during the entire interval

The law was never in effect during the entire interval

Setting Up The Data

```
. // Create a variable for time
. generate time = (year>1980)

.
. list state MLDA21_year year MLDA21 time          ///
>      if year!=1980 & inlist(FIPS, 17,18,19)      ///
>      , sepby(state) nolabel noobs abbrev(18)
```

state	MLDA21_year	year	MLDA21	time
Illinois	1980	1979	0	0
Illinois	1980	1981	1	1
Indiana	1934	1979	1	0
Indiana	1934	1981	1	1
Iowa	1986	1979	0	0
Iowa	1986	1981	0	1

Time equals 0 prior to the intervention in 1980 and time equals 1 after the intervention in 1980

Setting Up The Data

```
// Create a variable for treatment status
gen treated = .
replace treated = 0 if MLDA21_year >1980
replace treated = 0 if MLDA21_year <=1980 & year<1980
replace treated = 1 if MLDA21_year <=1980
```

```
. list state MLDA21_year year MLDA21 time treated ///
>      if year!=1980 & inlist(FIPS, 17,18,19)     ///
>      , sepby(state) nolabel noobs abbrev(18)
```

state	MLDA21_year	year	MLDA21	time	treated
Illinois	1980	1979	0	0	1
Illinois	1980	1981	1	1	1
Indiana	1934	1979	1	0	1
Indiana	1934	1981	1	1	1
Iowa	1986	1979	0	0	0
Iowa	1986	1981	0	1	0

Treated equals 1 if the MLDA21 law was in effect in that state during 1980 or sooner. Treated equals 0 if the law went into effect after 1980.

Setting Up The Data

```
. // Create a variable for the interaction of time and treatment
. gen TimeTreat = time*treated

.
. list state MLDA21_year year MLDA21 time treated TimeTreat  ///
>      if year!=1980 & inlist(FIPS, 17,18,19)                ///
>      , sepby(state) nolabel noobs abbrev(18)
```

state	MLDA21_year	year	MLDA21	time	treated	TimeTreat
Illinois	1980	1979	0	0	1	0
Illinois	1980	1981	1	1	1	1
Indiana	1934	1979	1	0	1	0
Indiana	1934	1981	1	1	1	1
Iowa	1986	1979	0	0	0	0
Iowa	1986	1981	0	1	0	0

TimeTreat equals time x treated.
It will be the interaction term when we use `regress`.
It will be the treatment variable when we use `didregress`!!!!

DiD Using `regress`

```
. regress mrate_agegrp time treated TimeTreat, vce(cluster FIPS)

Linear regression                               Number of obs     =        138
                                                F(3, 45)          =       5.76
                                                Prob > F          =     0.0020
                                                R-squared         =     0.0233
                                                Root MSE          =     22.879

                                    (Std. err. adjusted for 46 clusters in FIPS)
------------------------------------------------------------------------------
             |               Robust
mrate_agegrp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        time |   -3.92868   2.045822    -1.92   0.061    -8.049176    .1918168
     treated |    6.549406  6.480692     1.01   0.318    -6.503378    19.60219
  TimeTreat  |   -4.611695   3.32785    -1.39   0.173    -11.31433    2.090939
       _cons |    59.82451  3.878595     15.42  0.000     52.01262     67.6364
------------------------------------------------------------------------------
```

The Average Treatment Effect Among the Treated (ATET) is the coefficient for TimeTreat.

The MLDA21 law caused a decrease of 4.6 people per 100,000 population in the 18-20 year old age group between 1979 and 1981.

DiD Using `regress`

```
. regress mrate_agegrp i.time##i.treated, vce(cluster FIPS)

Linear regression                               Number of obs     =        138
                                                F(3, 45)          =       5.76
                                                Prob > F          =     0.0020
                                                R-squared         =     0.0233
                                                Root MSE          =     22.879

                                    (Std. err. adjusted for 46 clusters in FIPS)
------------------------------------------------------------------------------
             |               Robust
mrate_agegrp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      1.time |   -3.92868   2.045822    -1.92   0.061    -8.049176    .1918168
   1.treated |   6.549406   6.480692     1.01   0.318    -6.503378    19.60219
             |
time#treated |
         1 1 |  -4.611695    3.32785    -1.39   0.173    -11.31433    2.090939
             |
       _cons |   59.82451   3.878595    15.42   0.000     52.01262     67.6364
------------------------------------------------------------------------------
```

DiD Using `regress`

```
. regress mrate_agegrp i.year TimeTreat i.FIPS, vce(cluster FIPS)

Linear regression                               Number of obs     =        138
                                                F(2, 45)          =          .
                                                Prob > F          =          .
                                                R-squared         =     0.8714
                                                Root MSE          =     10.187

                                    (Std. err. adjusted for 46 clusters in FIPS)
```

mrate_agegrp	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
year						
1980	-.5155614	2.882514	-0.18	0.859	-6.321242	5.29012
1981	-4.18646	3.313667	-1.26	0.213	-10.86053	2.487608
TimeTreat	-4.611695	4.083392	-1.13	0.265	-12.83607	3.612679
FIPS						
2	10.37779	2.88e-13	3.6e+13	0.000	10.37779	10.37779
4	3.268883	2.88e-13	1.1e+13	0.000	3.268883	3.268883
5	5.474611	1.361131	4.02	0.000	2.733153	8.216069

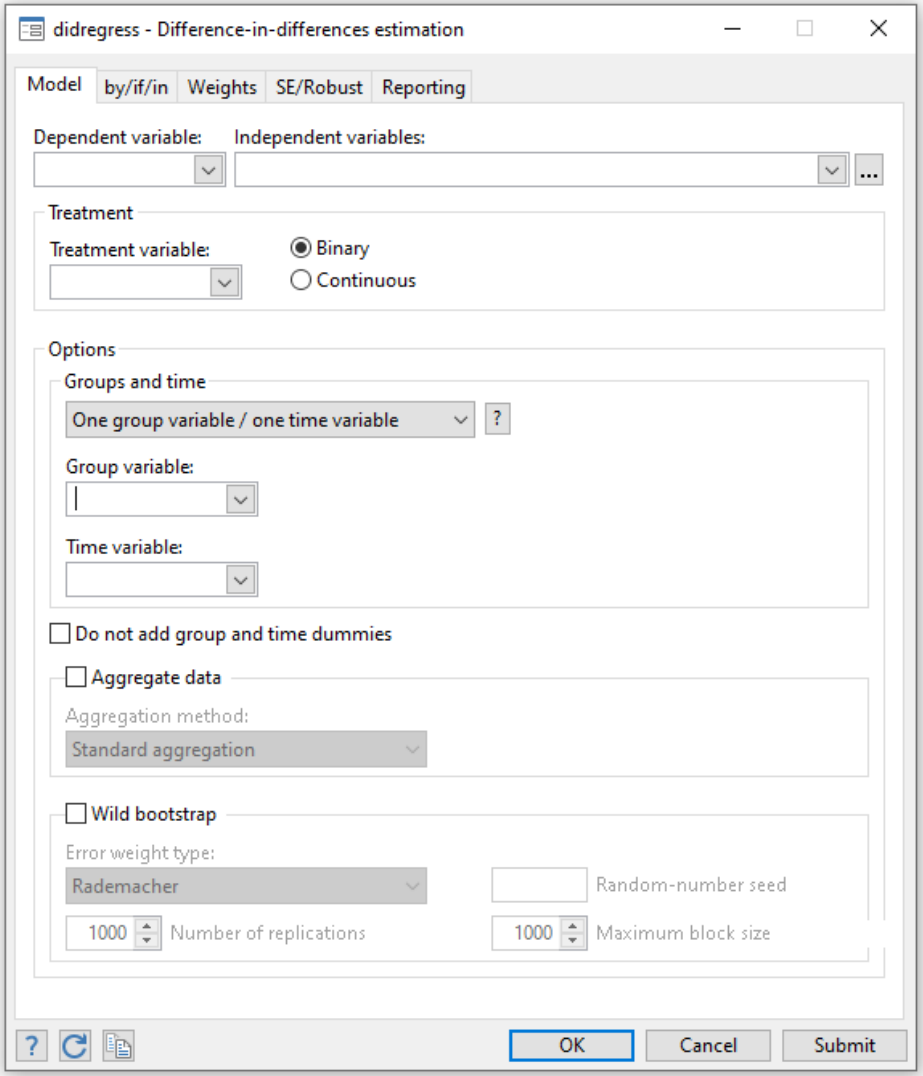
(FIPS output truncated)

DiD Using `regress`

```
. regress mrate_agegrp i.year TimeTreat i.FIPS, vce(cluster FIPS)

Linear regression                               Number of obs     =        138
                                                F(2, 45)          =          .
                                                Prob > F          =          .
                                                R-squared         =     0.8714
                                                Root MSE          =     10.187

                                    (Std. err. adjusted for 46 clusters in FIPS)
```

mrate_agegrp	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
year						
1980	-.5155614	2.882514	-0.18	0.859	-6.321242	5.29012
1981	-4.18646	3.313667	-1.26	0.213	-10.86053	2.487608
TimeTreat	-4.611695	4.083392	-1.13	0.265	-12.83607	3.612679
FIPS						
2	10.37779	2.88e-13	3.6e+13	0.000	10.37779	10.37779
4	3.268883	2.88e-13	1.1e+13	0.000	3.268883	3.268883
5	5.474611	1.361131	4.02	0.000	2.733153	8.216069

DiD Using `regress`



DiD Using `didregress`

```
didregress (ovar omvarlist)              ///
           (tvar[, continuous]),         ///
           group(groupvars)              ///
           [time(timevar) options]
```

```
didregress (mrate_agegrp)     ///
           (TimeTreat),       ///
           group(FIPS)        ///
           time(year)
```

TimeTreat = Time x Treat !!!!

DiD Using `didregress`

```
. didregress (mrate_agegrp)(TimeTreat), group(FIPS) time(year)
```

Number of groups and treatment time

Time variable: year
Control: TimeTreat = 0
Treatment: TimeTreat = 1

	Control	Treatment
Group		
FIPS	33	13
Time		
Minimum	1979	1981
Maximum	1979	1981

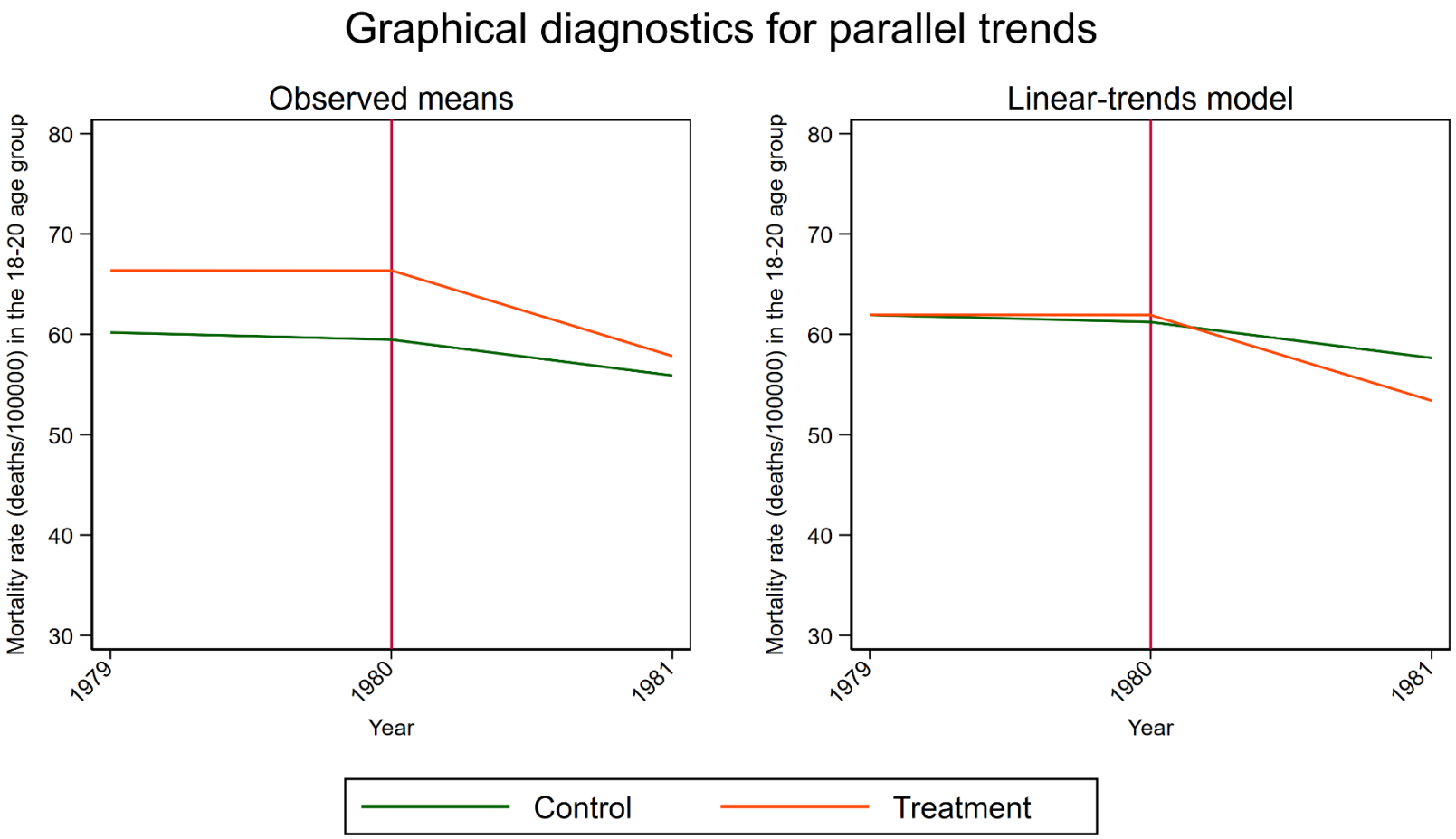
Difference-in-differences regression — Number of obs = 138
Data type: Repeated cross-sectional

(Std. err. adjusted for 46 clusters in FIPS)

mrate_agegrp	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
ATET						
TimeTreat (1 vs 0)	-4.611695	4.083392	-1.13	0.265	-12.83607	3.612679

Note: ATET estimate adjusted for group effects and time effects.

Checking the Parallel Trends Assumption




```
estat trendplots
```

Checking the Parallel Trends Assumption

```
. estat granger

Granger causality test
H0: No effect in anticipation of treatment

F(1, 45) =   0.02
Prob > F = 0.8871
```

didregress With Covariates

```
. didregress (mrate_agegrp beertax)(TimeTreat), group(FIPS) time(year)

Number of groups and treatment time

Time variable: year
Control:       TimeTreat = 0
Treatment:     TimeTreat = 1
```

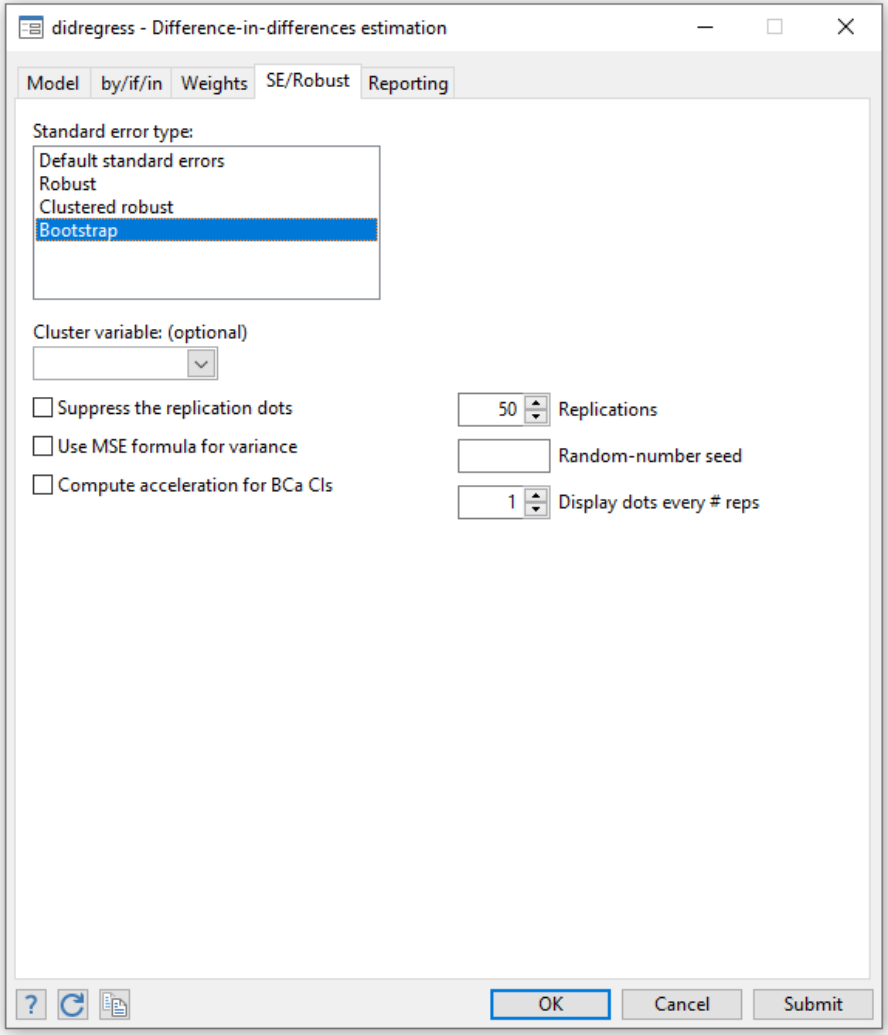
	Control	Treatment
Group		
FIPS	32	13
Time		
Minimum	1979	1981
Maximum	1979	1981

```
Difference-in-differences regression                 Number of obs = 135
Data type: Repeated cross-sectional

                                   (Std. err. adjusted for 45 clusters in FIPS)
```

mrate_agegrp	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
ATET						
TimeTreat (1 vs 0)	-5.140093	4.195906	-1.23	0.227	-13.59639	3.316199

```
Note: ATET estimate adjusted for covariates, group effects, and time effects.
```

Estimation of Standard Errors



Estimation of Standard Errors

Must account for the serial correlation of the outcome.

For a large number of groups

By default, `didregress` uses cluster–robust standard errors and uses critical values of a *t* distribution with $G - 1$ degrees of freedom, where G is the number of groups.

For more information: https://www.stata.com/manuals/tedidintro.pdf

Estimation of Standard Errors

For a small number of groups

wildbootstrap[(*wildopts*)] computes confidence intervals and *p*-values with the wild bootstrap. The wild bootstrap is constructed imposing the null hypothesis that the ATET is 0; that is, it is a restricted wild bootstrap. Confidence intervals are computed separately from the *p*-values. The bounds of the confidence interval are computed using a bisection optimization algorithm described in *Methods and formulas* in **[TE] didregress**. *wildopts* are errorweight(*edtype*), reps(#), rseed(#), and blocksize(#).

errorweight(*edtype*) defines the error weight used to draw residuals from the wild bootstrap. *edtype* is one of rademacher (the default), mammen, webb, normal, or gamma.

rademacher multiplies the residuals at each bootstrap replication with a randomly generated variable that takes the value of 1 with probability 0.5 and the value of -1 with probability 0.5. errorweight(rademacher) is the default.

mammen multiplies the residuals at each bootstrap replication with a randomly generated variable that takes the value of 1 - phi with probability phi/sqrt{5} and phi otherwise, where phi = (1+sqrt{5})/2.

webb multiplies the residuals at each bootstrap replication with a randomly generated variable that takes the values -sqrt{3/2}, -sqrt{2/2}, -sqrt{1/2}, sqrt{1/2}, sqrt{2/2}, and sqrt{3/2}, each with probability 1/6.

normal multiplies the residuals at each bootstrap replication with a randomly generated normal distribution variable with the first four moments given by 0, 1, 0, and 3.

gamma multiplies the residuals at each bootstrap replication with a randomly generated gamma distribution variable with shape parameter 4 and scale parameter 1/2.

reps(#) performs # wild bootstrap replications. The default is reps(1000).

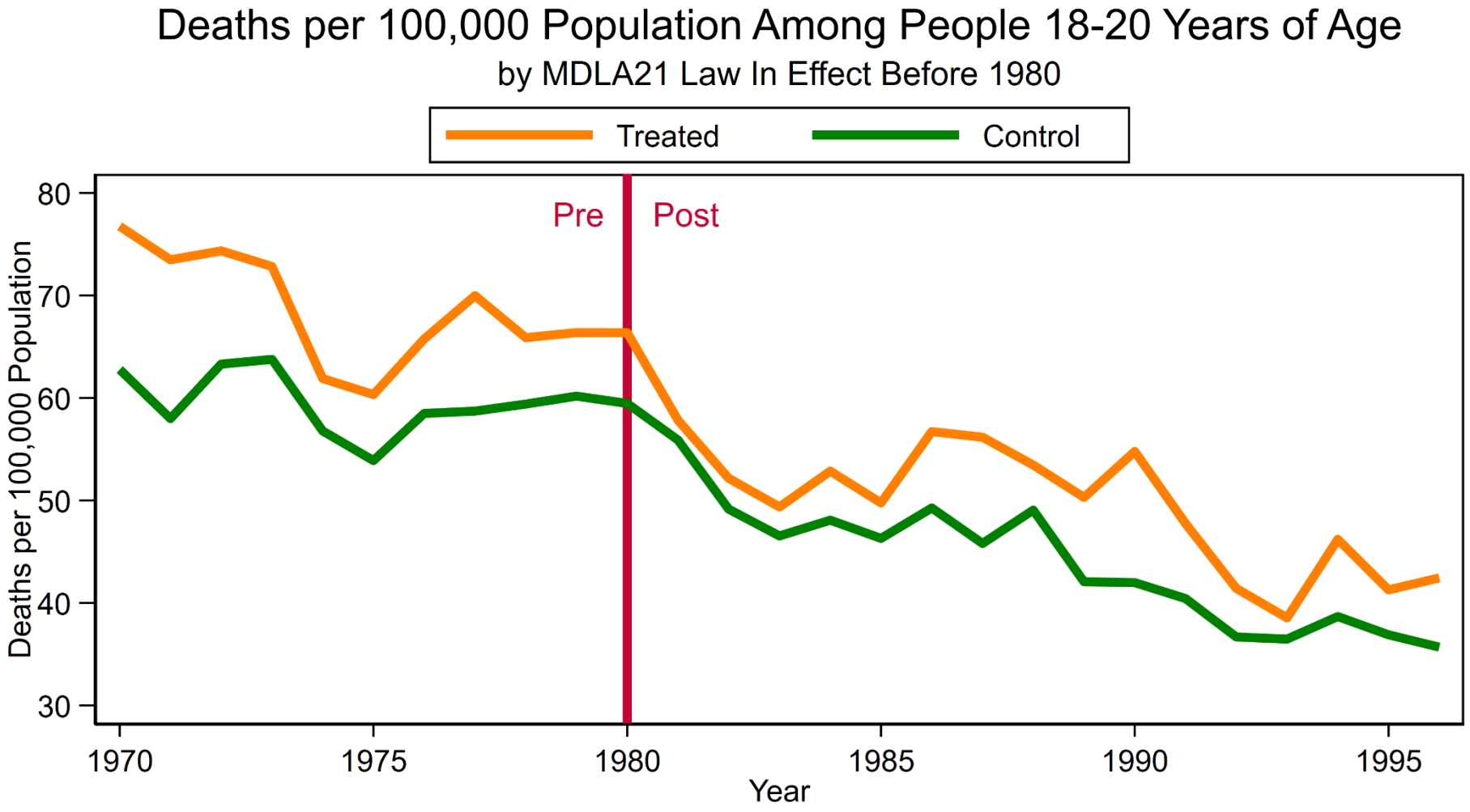
rseed(#) sets the random-number seed to #.

blocksize(#) specifies that the wild bootstrap be performed in blocks, with # replications per block. The wild bootstrap computation requires a matrix with dimensions (# groups) x (# replications). If this is too large, you can reduce the matrix to (# groups) x (# block size) and loop (# replications)/(# block size) times. When the same random seed is set, using a different block size does not change the numerical results; it only modifies the computation method. The block size must be less than or equal to the number of bootstrap replications.

Outline

- ✓ The Question
- ✓ An Intuitive Introduction
- ✓ Two Period, Two Groups Model
- Repeated Cross-Sectional Panel Data
- Longitudinal Panel Data
- More information

DiD For Repeated Cross-Sectional Data



DiD For Repeated Cross-Sectional Data

```
use MLDA21, clear

// Create a variable for time
generate time = (year>1980)

// Create a variable for treatment status
gen treated = .
replace treated = 0 if MLDA21_year >1980
replace treated = 0 if MLDA21_year <=1980 & year<1980
replace treated = 1 if MLDA21_year <=1980

// Create a variable for the interaction of time and treatment
gen TimeTreat = time*treated
```

```
. list state MLDA21_year year MLDA21 time treated TimeTreat  ///
>      if year!=1980 & inlist(FIPS, 17)                      ///
>      , sepby(state) nolabel noobs abbrev(18)
```

state	MLDA21_year	year	MLDA21	time	treated	TimeTreat
Illinois	1980	1970	0	0	1	0
Illinois	1980	1971	0	0	1	0
Illinois	1980	1972	0	0	1	0
Illinois	1980	1973	0	0	1	0
Illinois	1980	1974	0	0	1	0
Illinois	1980	1975	0	0	1	0
Illinois	1980	1976	0	0	1	0
Illinois	1980	1977	0	0	1	0
Illinois	1980	1978	0	0	1	0
Illinois	1980	1979	0	0	1	0
Illinois	1980	1981	1	1	1	1
Illinois	1980	1982	1	1	1	1
Illinois	1980	1983	1	1	1	1
Illinois	1980	1984	1	1	1	1
Illinois	1980	1985	1	1	1	1
Illinois	1980	1986	1	1	1	1
Illinois	1980	1987	1	1	1	1
Illinois	1980	1988	1	1	1	1
Illinois	1980	1989	1	1	1	1
Illinois	1980	1990	1	1	1	1
Illinois	1980	1991	1	1	1	1
Illinois	1980	1992	1	1	1	1
Illinois	1980	1993	1	1	1	1
Illinois	1980	1994	1	1	1	1
Illinois	1980	1995	1	1	1	1
Illinois	1980	1996	1	1	1	1

DiD For Repeated Cross-Sectional Data

```
. didregress (mrate_agegrp)(TimeTreat), group(FIPS) time(year)

Number of groups and treatment time

Time variable: year
Control:       TimeTreat = 0
Treatment:     TimeTreat = 1
```

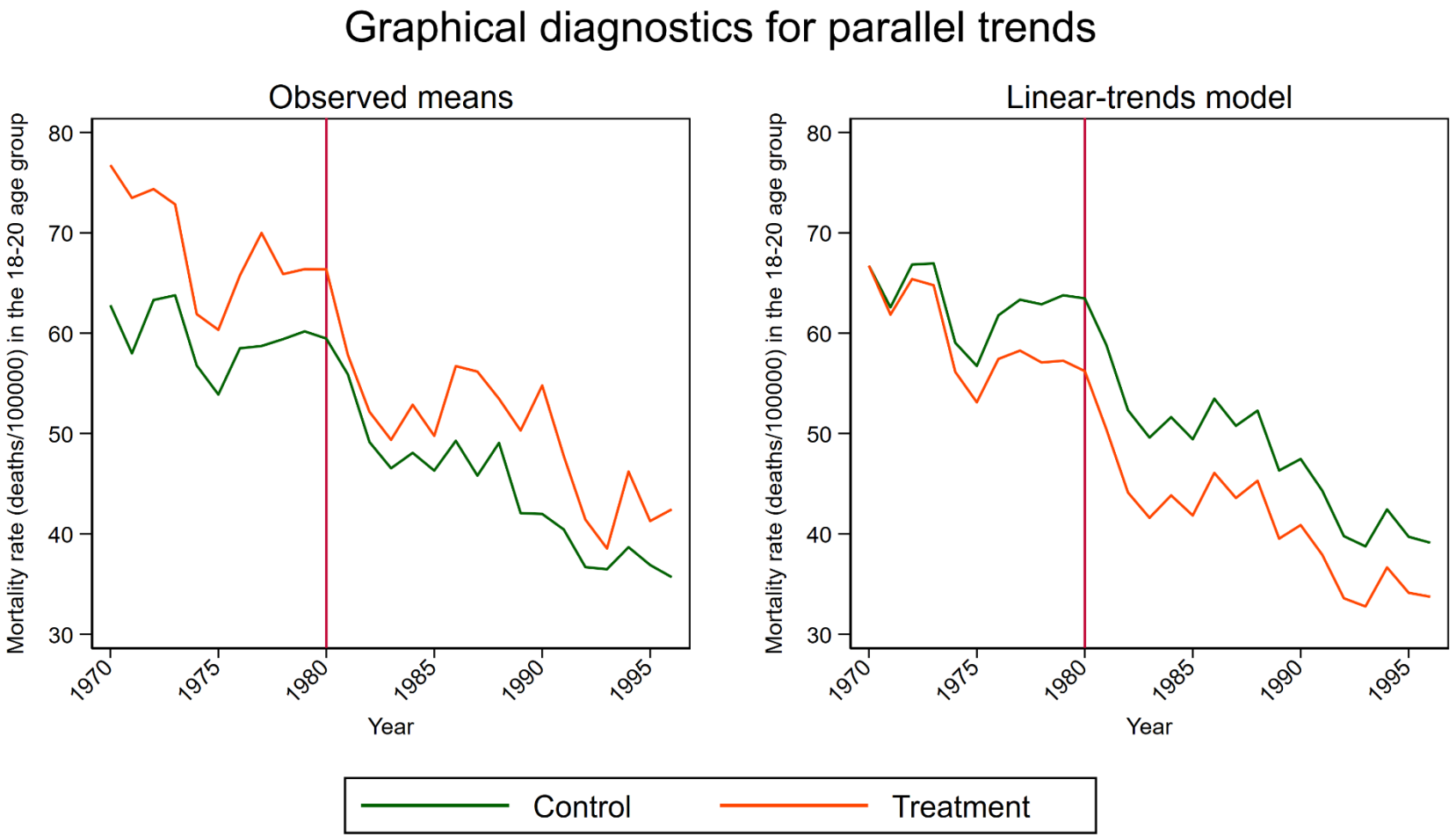
	Control	Treatment
Group		
FIPS	33	13
Time		
Minimum	1970	1981
Maximum	1970	1981

```
Difference-in-differences regression                 Number of obs = 1,242
Data type: Repeated cross-sectional

                                     (Std. err. adjusted for 46 clusters in FIPS)
```

mrate_agegrp	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
ATET						
TimeTreat (1 vs 0)	-3.268054	4.212866	-0.78	0.442	-11.7532	5.217093

```
Note: ATET estimate adjusted for group effects and time effects.
```

DiD For Repeated Cross-Sectional Data




```
estat trendplots
```

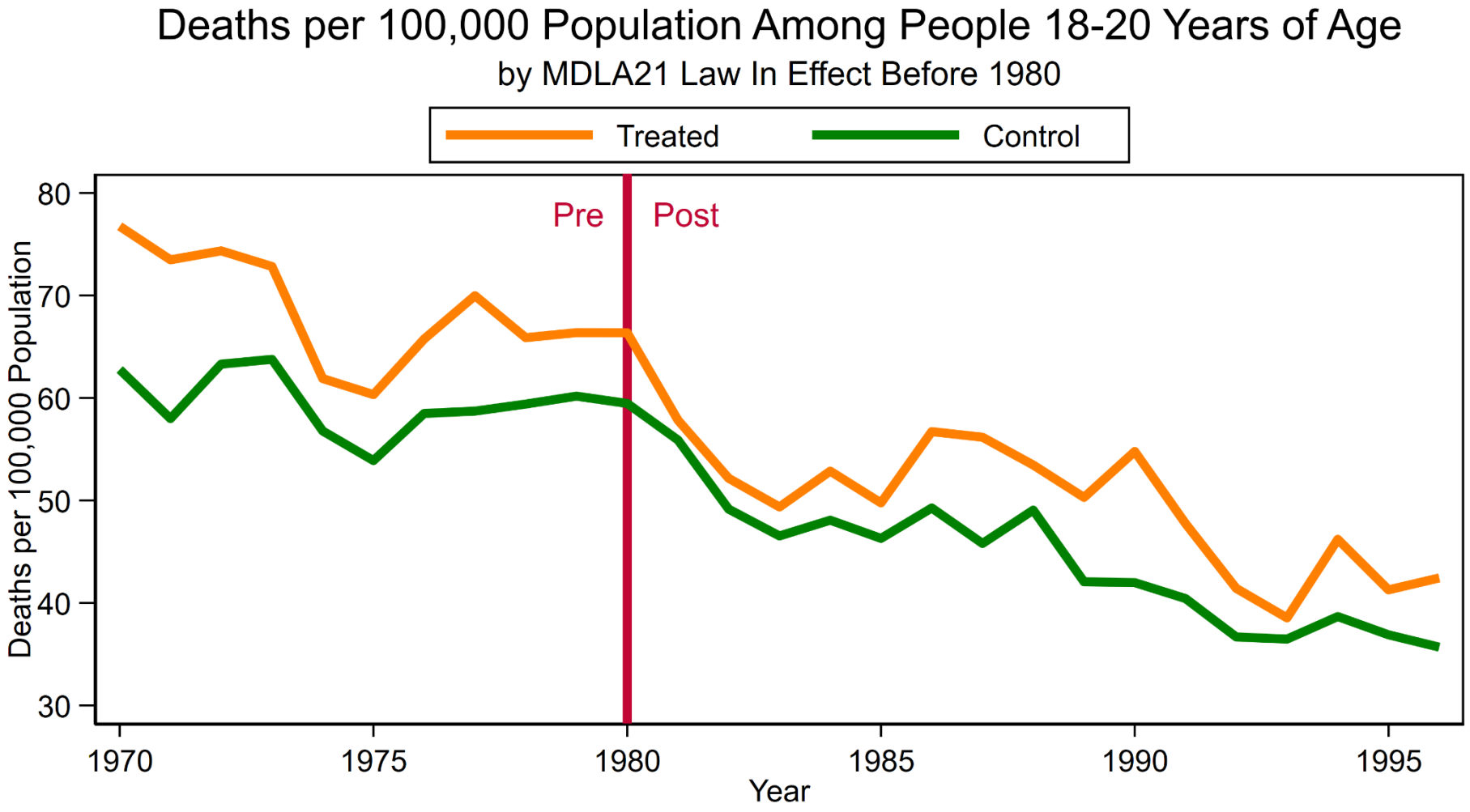
DiD For Repeated Cross-Sectional Data

```
. estat granger

Granger causality test
H0: No effect in anticipation of treatment

F(10, 45) =   0.95
 Prob > F = 0.4952
```

Outline

- ✓ The Question
- ✓ An Intuitive Introduction
- ✓ Two Period, Two Groups Model
- ✓ Repeated Cross-Sectional Panel Data
- Longitudinal Panel Data
- More information

DiD For Longitudinal Panel Data



DiD For Longitudinal/Panel Data

```
. xtset FIPS year, yearly

Panel variable: FIPS (strongly balanced)
 Time variable: year, 1970 to 1996
         Delta: 1 year
```

DiD For Longitudinal/Panel Data

```
. xtdidregress (mrate_agegrp)(TimeTreat), group(FIPS) time(year)

Number of groups and treatment time

Time variable: year
Control:       TimeTreat = 0
Treatment:     TimeTreat = 1
```

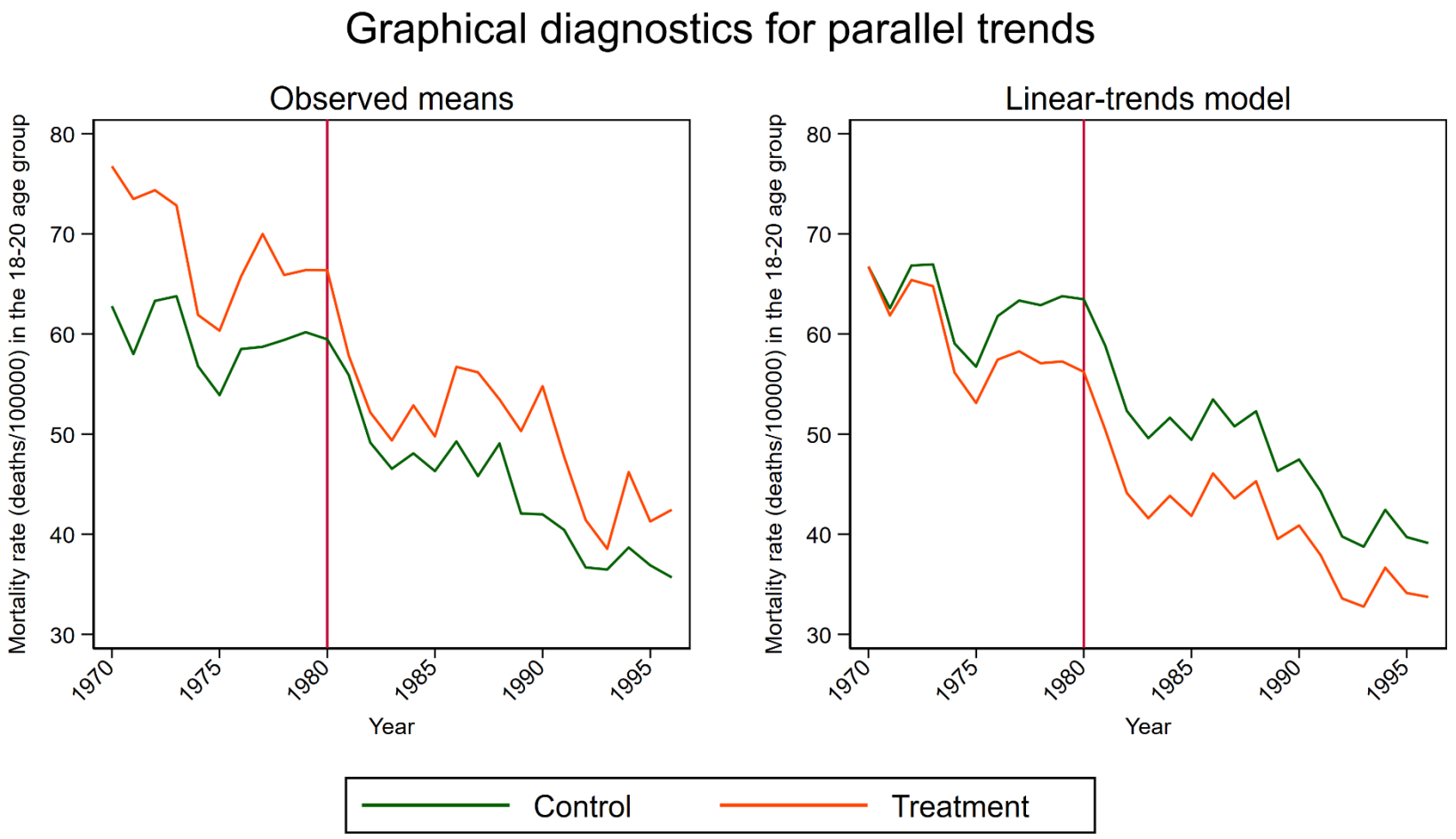
	Control	Treatment
Group		
FIPS	33	13
Time		
Minimum	1970	1981
Maximum	1970	1981

```
Difference-in-differences regression                 Number of obs = 1,242
Data type: Longitudinal

                                   (Std. err. adjusted for 46 clusters in FIPS)
```

mrate_agegrp	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
ATET						
TimeTreat (1 vs 0)	-3.268054	4.134048	-0.79	0.433	-11.59445	5.058346

```
Note: ATET estimate adjusted for panel effects and time effects.
```

DiD For Longitudinal/Panel Data




```
estat trendplots
```

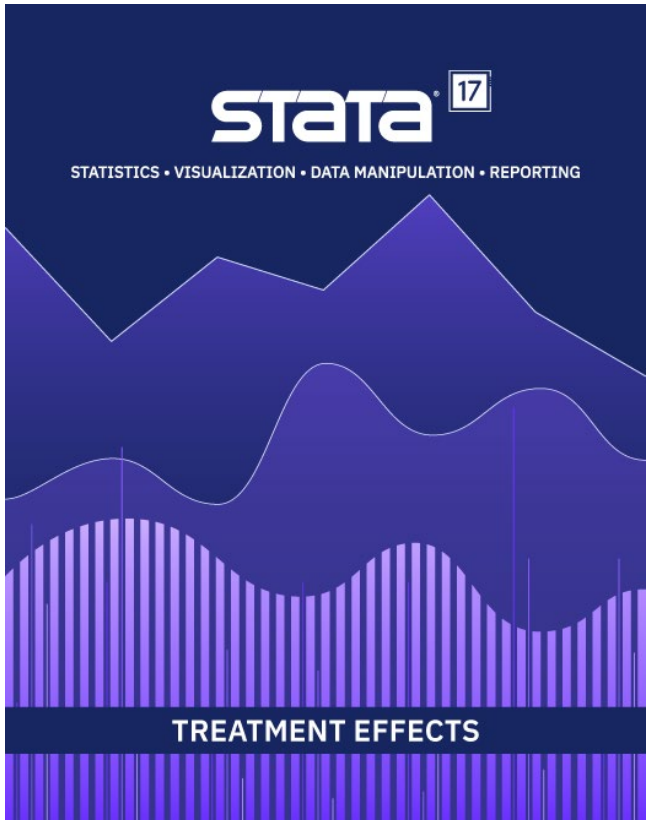
DiD For Longitudinal/Panel Data

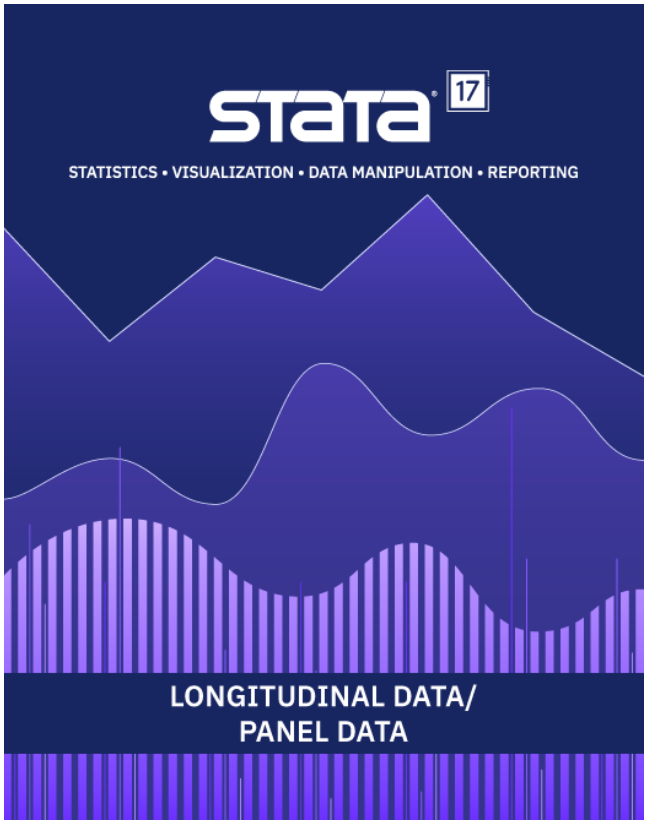
```
. estat granger

Granger causality test
H0: No effect in anticipation of treatment

F(10, 45) =   0.99
 Prob > F = 0.4650
```

Outline

- The Question
- An Intuitive Introduction
- Two Period, Two Groups Model
- Repeated Cross-Sectional Panel Data
- Longitudinal Panel Data
- More information

Read The Fine Manuals!

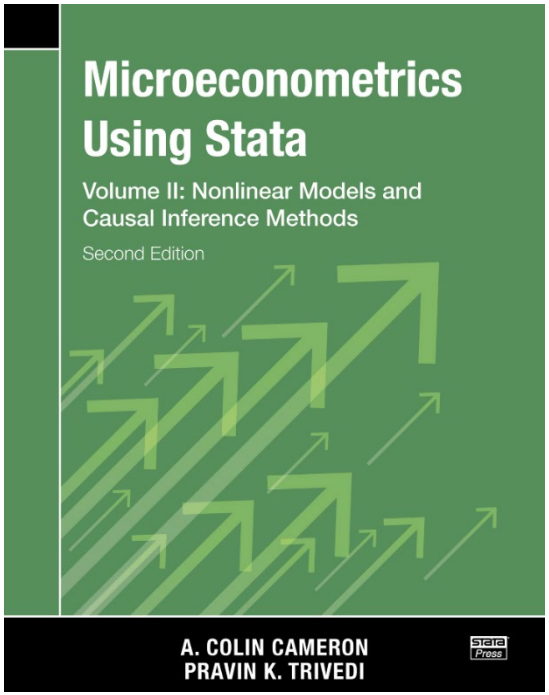


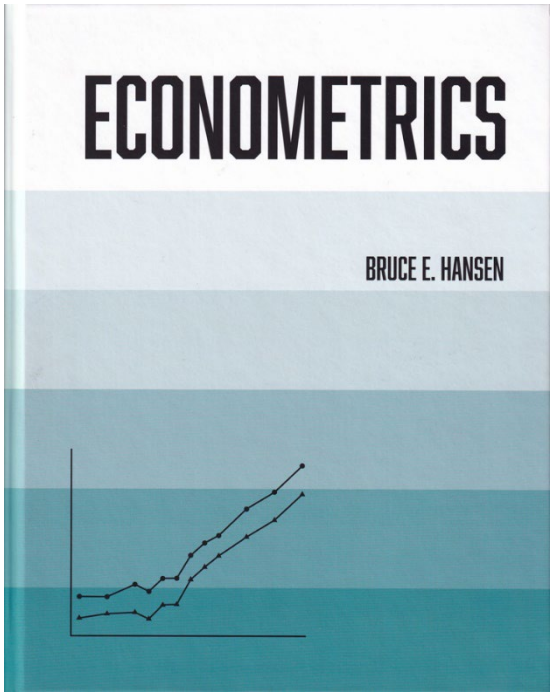


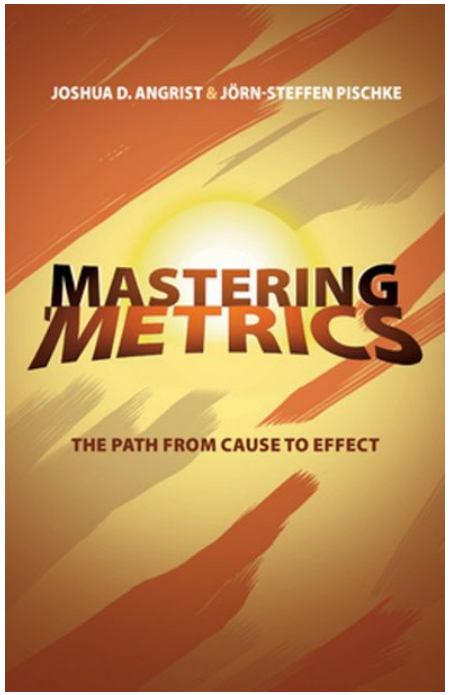
didregress https://www.stata.com/manuals/te.pdf

xtdidregress https://www.stata.com/manuals/xt.pdf

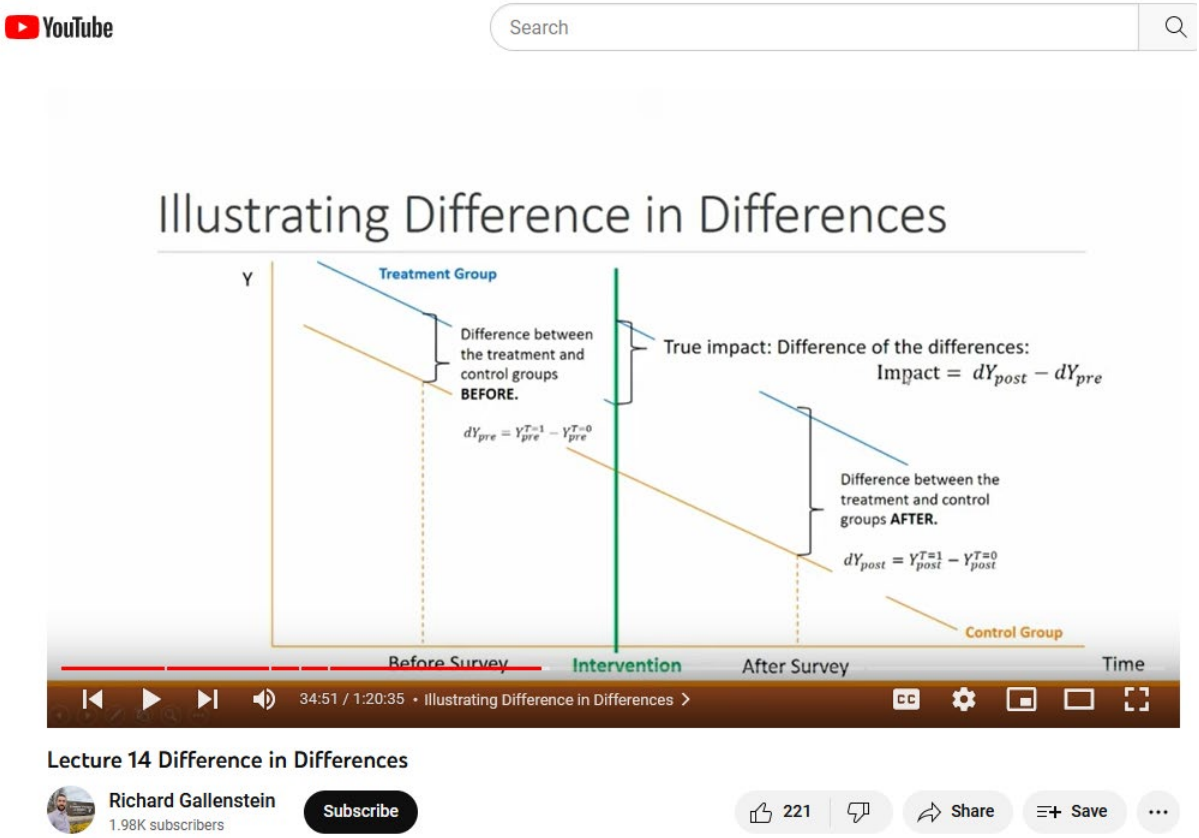
More Information







More Information



Richard Gallenstein's YouTube video about DiD

https://www.youtube.com/watch?v=rrgxR97QBpc

More Information

US National Highway Traffic Safety Administration Information

Miron and Tetelbaum, 2009

Thank you!

Questions?

You can download the slides, datasets, and do-files here:
https://tinyurl.com/StataDID

You can contact me anytime at chuber@stata.com