Module 3

Day 1: Differences-in-Differences

Goals for Today

- Card and Krueger
  - Regression analysis
  - OLS and Fixed Effects
  - Differences-in-Differences
- Review of:
  - Cleaning data
  - Using `lm()`
  - Presenting results with `stargazer()`

The 1994 Card and Krueger Study

- A \$0.80 minimum wage increase by New Jersey in 1992 prompted researchers to revisit the debate on the effect of the minimum wage on labor markets.

The 1994 Card and Krueger Study

- A $0.80 minimum wage increase by New Jersey in 1992 prompted researchers to revisit the debate on the effect of the minimum wage on labor markets.
- David Card and Alan Krueger collected information from fast-food restaurants in New Jersey and eastern Pennsylvania.

The 1994 Card and Krueger Study

- A $0.80 minimum wage increase by New Jersey in 1992 prompted researchers to revisit the debate on the effect of the minimum wage on labor markets.
- David Card and Alan Krueger collected information from fast-food restaurants in New Jersey and eastern Pennsylvania.
- They found no indication of a reduction in employment due to the minimum wage increase.

The 1994 Card and Krueger Study

- A $0.80 minimum wage increase by New Jersey in 1992 prompted researchers to revisit the debate on the effect of the minimum wage on labor markets.
- David Card and Alan Krueger collected information from fast-food restaurants in New Jersey and eastern Pennsylvania.
- They found no indication of a reduction in employment due to the minimum wage increase.
- An even bigger impact of the study was the introduction of differences-in-differences.
  - Intuitive tool suitable for quasi-experimental studies

The 1994 Card and Krueger Study

- A $0.80 minimum wage increase by New Jersey in 1992 prompted researchers to revisit the debate on the effect of the minimum wage on labor markets.
- David Card and Alan Krueger collected information from fast-food restaurants in New Jersey and eastern Pennsylvania.
- They found no indication of a reduction in employment due to the minimum wage increase.
- An even bigger impact of the study was the introduction of differences-in-differences.
  - Intuitive tool suitable for quasi-experimental studies
- We downloaded the data from *David Card's personal website*.

The 1994 Card and Krueger Study

- As is argued in the paper, New Jersey and Pennsylvania are similar enough states geographically, and economically, that they lend themselves to comparison.
  - NJ and PA present a natural experiment with a "control" group (no wage increase in PA) and a "treatment" group (NJ minimum wage increase).
  - Differences of course exist, but we can control for state-level differences using fixed effects.
  - Differences-in-differences will allow us to understand the effect of policy like a minimum wage increase because it provides insight into differences between the control and treament in more than one time period.

Reading in the Data: .dat files

Let's start by reading in the Card and Krueger data and doing some basic analysis in the same way that we have been since the beginning of class. We want to get an idea of what the data looks like before we perform any sort of regressions.

- Find the data file with the name 'public'
- This is saved as a .dat file: a generic data file format that can be read into R using the `fread()` function from the `data.table` package
- Once you've read in the data, take a look at the variable names and types

Reading in the Data: .dat files

- What are some issues you might foresee with these variables?
- *Why did many of our variables get read in as characters?*
- Try reading in the data again, using the code below to avoid turning variables into characters that should not be

```
cardKrueger <- fread("Data/public.dat",
                     na.strings = c("NA", "."))
```

Card and Krueger Data Codebook

Let's look at the codebook to rename and recode our variables.

- **ID variables:** SHEET, CHAIN, CO_OWNED, STATE
- **Location variables:** SOUTHJ, CENTRALJ, NORTHJ, PA1, PA2, SHORE
- **First interview variables:** NCALLS, EMPFT, EMPPT, NMGRS, WAGE_ST, INCTIME, FIRSTINC, BONUS, PCTAFF, MEALS, OPEN, HRSOPEN, PSODA, PFRY, PENTREE, NREGS, NREGS11
- **Second interview variables:** TYPE2, STATUS2, DATE2, NCALLS2, EMPFT2, EMPPT2, NMGRS2, WAGE_ST2, INCTIME2, FIRSTIN2, SPECIAL2, MEALS2, OPEN2R, HRSOPEN2, PSODA2, PFRY2, PENTREE2, NREGS2, NREGS112

In-Class Exercise

- Use the information from the codebook to create:
  - A categorical **chainName** variable:
    - "Burger King"
    - "KFC"
    - "Roys"
    - "Wendys"
  - A categorical **stateAbbrv** variable:
    - "NJ"
    - "PA"
  - A categorical **location** variable:
    - "Southern NJ"
    - "Central NJ"
    - "Northern NJ"
    - "Northeast suburbs of Philadelphia"
    - "Easton"
    - "NJ Shore"

Understanding the Movement in Wages

Now that we've cleaned our data, we can go back to our original questions about the dataset. We want to know *how minimum wage affects employment*, so let's start by looking at whether the policy change actually caused any change in the average wage in New Jersey.

Understanding the Movement in Wages

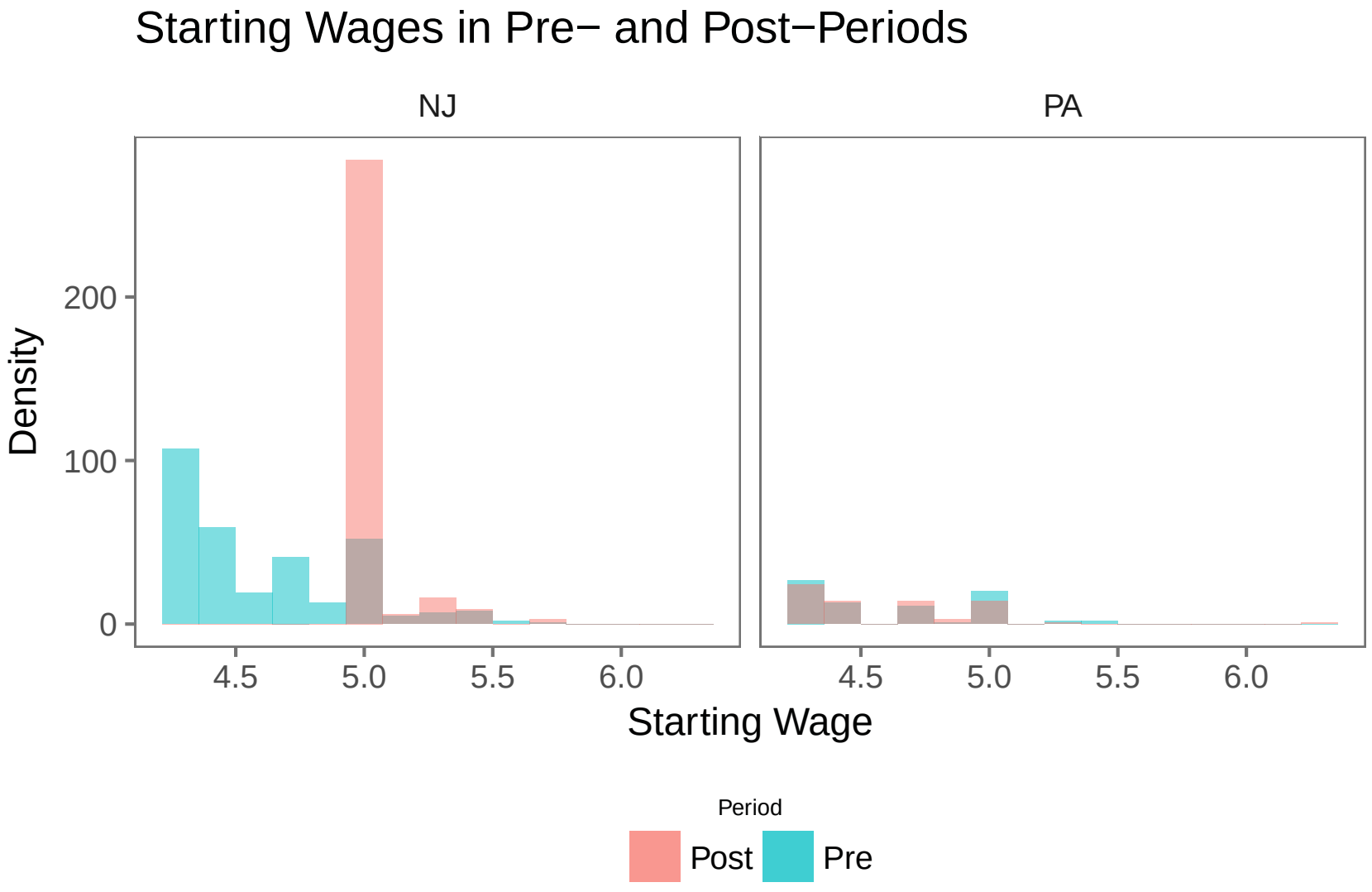
What is the average wage by state, by fast food chain, before and after the policy goes into effect?

```
## # A tibble: 8 x 4
##      chainName stateAbbrv avgWageBefore avgWageAfter
##          <chr>      <chr>          <dbl>        <dbl>
## 1 Burger King         NJ       4.533228     5.076364
## 2 Burger King         PA       4.521176     4.579697
## 3         KFC         NJ       4.575224     5.075758
## 4         KFC         PA       4.725000     4.683333
## 5        Roys         NJ       4.669241     5.063462
## 6        Roys         PA       4.794118     4.660667
## 7      Wendys         NJ       4.806829     5.135238
## 8      Wendys         PA       4.628000     4.600000
```

Understanding the Movement in Wages

- There seems to be significant movement in starting wage between these two periods in New Jersey, while Pennsylvania stays relatively constant for each chain.
- Let's investigate the movement in wages graphically also:
  - Make a histogram plot of wages in the two periods using color to indicate the time period.

Understanding the Movement in Wages



Understanding the Movement in Wages

What do we learn from our data and plot about the movement in wages between periods?

- Is this enough to know whether the minimum wage increase in New Jersey had an impact on the labor market?
- Are there differences between chains in the wages paid to employees?
  - How might this impact (or not impact) our analysis?
- Do you think stores with higher salaries have cheaper or more expensive food?

Understanding the Simultaneous Shifts in Employment

- For their measure of employment, Card and Krueger create a **totalEmp** variable that consists of full-time employees, managers, and half of part-time employees.
- Using `mutate()`, create **totalEmp** and **totalEmp2**; variables for total employees in both periods.

```
cardKrueger_clean <- cardKrueger_clean %>%
  mutate(totalEmp = #code to execute,
         totalEmp2 = #code to execute))
```

- What is the total number of employees by state, by fast food chain, before and after the policy goes into effect?

Understanding the Simultaneous Shifts in Employment

- Recall that `stargazer()` can produce tables presenting statistics. We can present our `numEmp` object using the option `summary = FALSE` in order to supress calculations of descriptive statistics on our sums of total employees.

Table 1: numEmp Summary

chainName	stateAbbrv	empBefore	empAfter
Burger King	NJ	2903.05	3096
Burger King	PA	970.75	917.75
KFC	NJ	856.75	933.5
KFC	PA	128.5	156
Roys	NJ	1874	1695.25
Roys	PA	335.5	268.75
Wendys	NJ	927.25	983
Wendys	PA	361.75	287.25

Note on Using `stargazer()` for Summary Tables

Main arguments used for summary statistics tables:

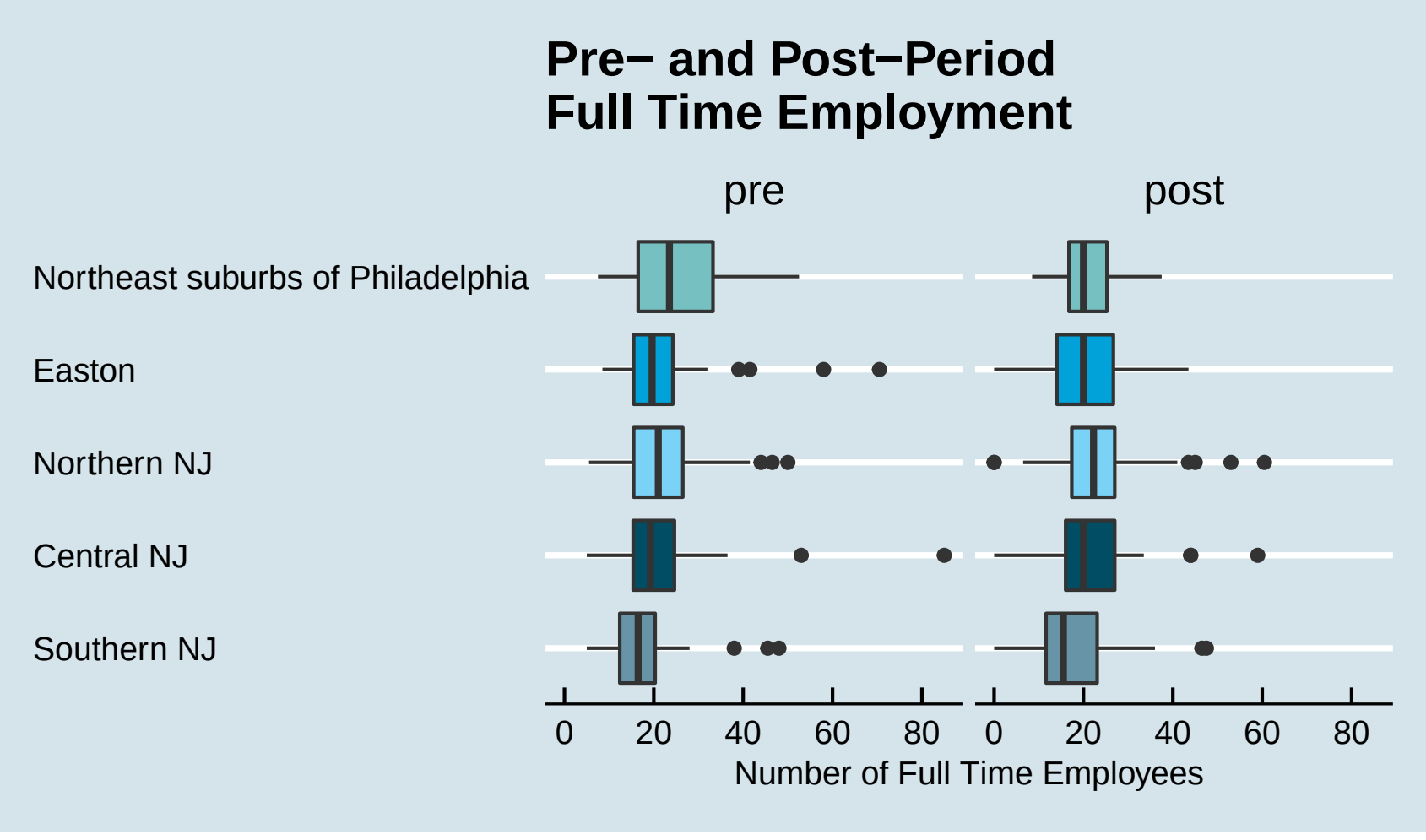
- **summary:** If TRUE, calculate number of observations, mean, median, max, and min of numeric variables in the input object. If FALSE, output the contents of the input object.
- **rownames:** If TRUE, output the rownames (by default these are the row number). If FALSE, suppress rownames from the output.
- **digits:** Lets you control how many decimal places to print.
- . . . : Many more! Check the manual

Understanding the Simultaneous Shifts in Employment

Make a histogram of employment by region before the minimum wage changes and after.

- We need the graph to show employment by subsets of both region and period (instead of just period, like in the wages histogram we made)
- In order to do so, we will need to create an indicator variable for period
- We can use `gather()` to reshape our data into a longer format with **variable**, **value**, and **period** columns, instead of individual columns for each variable

Understanding the Simultaneous Shifts in Employment



What does our plot tell us about differences in employment by region and period?

Causal Effect

We have seen a number of relationships between variables of interest in our data exploration exercise:

- States have different patterns in wage changes
- Chains offer different levels of wages to new employees
- Regions have varying levels of employees per store

What we have not seen, however, is any proof of **causal effect.** Our research goal is to exactly find the effect of wage on employment; to do this we must use regressions.

OLS Regression

We can start with a straightforward linear regression on pre-period employment and wages.

- What hypothesis is this regression testing?

```
preReg <- lm(EMPFT ~ WAGE_ST, cardKrueger_clean)
```

Use `stargazer()` to see how well our hypothesis performed on the pre-policy-change period.

OLS Regression

Table 2: Pre-period Regression

	EMPFT
WAGE_ST	3.193**
	(1.255)
Constant	−6.468
	(5.807)
Observations	385
R^2	0.017
Adjusted R^2	0.014
Note:	*p<0.1; **p<0.05; ***p<0.01

- Is this model doing a good job explaining variations in the employment variable?

OLS and Fixed Effects

- **Fixed effects** let you limit your analysis to within-group variation.
  - For our use case, there is no meaningful difference between a fixed effects estimator and dummy variable estimator; they are computationally equivalent.
- In a panel data setting fixed effects are most commonly used to account for heterogeneity **between** individual subjects.
  - This allows your model to get consistent estimates even with structural differences between groups/subjects.
- As an analogy, think about how `group_by()` in the `dplyr` package works.
  - Fixed effects are similar to running `lm()` after using `group_by()`

OLS and Fixed Effects

- Some reasons to use fixed effects:
  - The variation you see when pooling all observations together might not be of interest to you.
  - You believe a variable, or any transformation of one, should not enter your model linearly but still does explain variation.
  - There is a non-numeric variable that you want to include in your model.

OLS and Fixed Effects

- You cannot use fixed effects with a variable if there is not enough variation in each group of the variable.
  - For example, you could not use fixed effects to treat each observation of our dataset as a "group."
  - Each group would have just one observation, which is not enough variation to run a regression.

Specifying Fixed Effects in `lm()`

Scenario 1: If class of x is character, factor, logical, or otherwise only has two levels:

```
lm(y ~ x)
```

Scenario 2: If class of x is numeric and has more than two unique values, convert to factor:

```
lm(y ~ as.factor(x))
```

Which variables (other than wage) did we find to be impactful on employment?

- Are any of them categorical variables that could be used as fixed effects?

OLS, Fixed Effects, and Controls

Let's try running the pre-period regression, adding state fixed effects.

```
preReg_fixedState <- lm(EMPFT ~ WAGE_ST + stateAbbrv,
                                  cardKrueger_clean)
```

We can keep improving our pre-period model by adding control variables to the fixed effects to account for additional variation in employment. Here we include controls used in the Card and Krueger study.

```
preReg_controls <- lm(EMPFT ~ WAGE_ST +
                            stateAbbrv + HRSOPEN +
                            PSODA + PFRY + PENTREE,
                          cardKrueger_clean)
```

OLS, Fixed Effects, and Controls

Let's compare the results of the `preReg`, `preReg_fixedState`, and `preReg_controls` model using `stargazer()`.

```
stargazer(preReg, preReg_fixedState, preReg_controls,
          title = "Pre-period Regressions",
          header = FALSE,
          dep.var.caption = "",
          omit.stat = c("ser","f"),
          no.space = TRUE,
          intercept.bottom = FALSE)
```

Note on Using `stargazer()` for Regression Output

Main arguments used for regression tables:

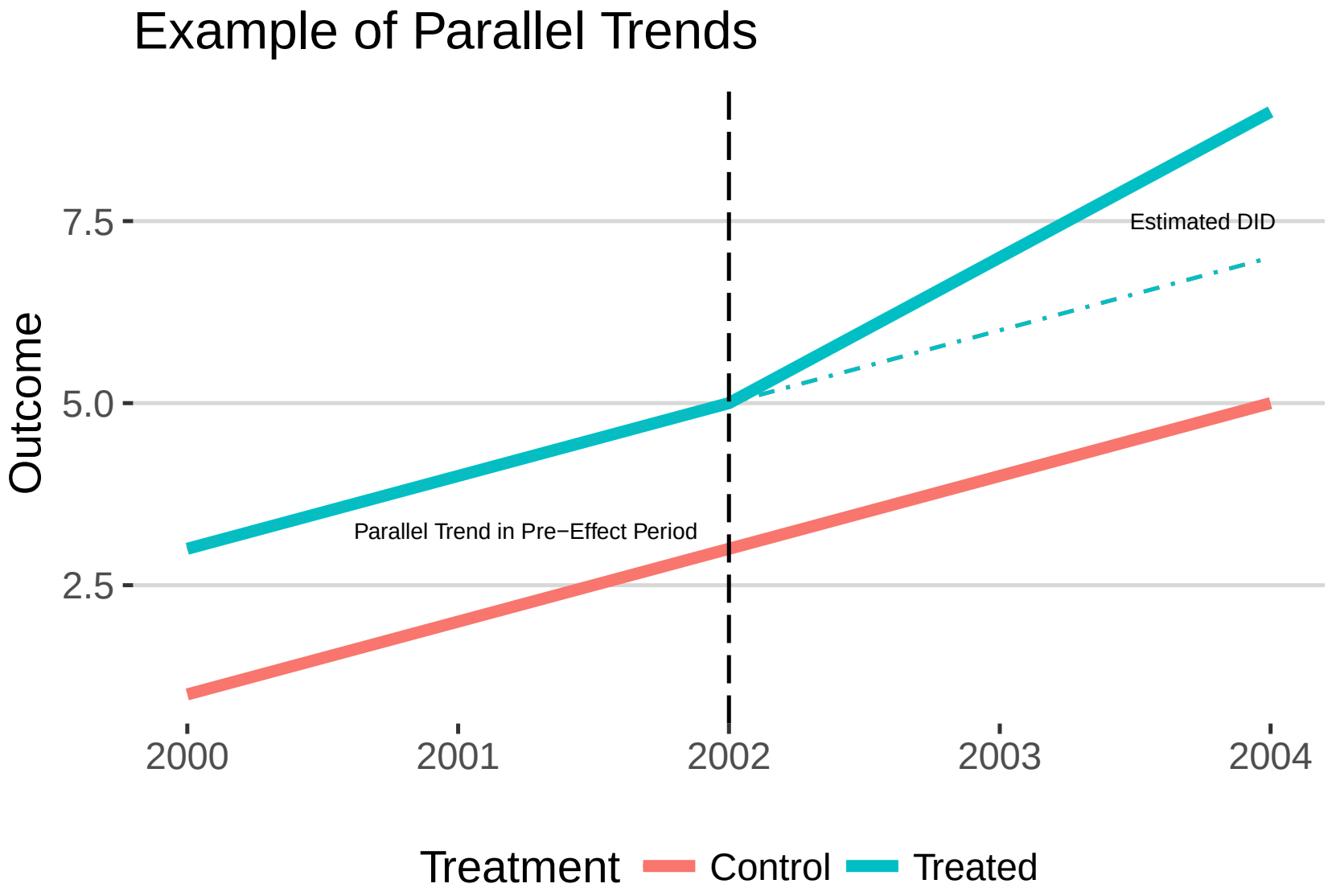
- **covariate.labels:** Labels for the variables listed on the far left, top to bottom
- **dep.var.labels:** Labels for the variables listed at the top of each column, left to right
- **title:** Title of table
- **keep/omit:** Suppress printing out variables you do not want
- **order:** Control the order of dependent variables (from top to bottom)
- **intercept.bottom:** If set to FALSE, the intercept term is put at the top; the default is TRUE and the intercept is listed last
- **column.labels:** Labels for columns
- . . . : And many more!

To find a description of EVERY argument to stargazer try the manual, `?stargazer`.

Difference-in-Differences (DID)

- An alternative way to approach determining a casual effect is using the difference-in-differences (DID) technique.
- With the DID technique you are able to make use of treatment and control groups to make causal inferences about the effect of some treatment.
- The method allows you to find differences between:
  - 1. Treatment and control groups
  - 2. Multiple time periods
- Random assignment is not necessary for DID, which makes it a useful tool for quasi-experiments.
- DID also accounts for (time-invariant) confounders, or variables that affect both the "x" and the "y" side of your equation.

The Major Assumption: Parallel Trends



The Major Assumption: Parallel Trends

In the example, we are able to clearly see how a treatment, which took effect in 2002, changed the trend line of the treatment group.

The additional difference in the treatment group in the post-treatment period (after 2002) from the baseline counterfactual (the dotted line) is the DID estimate.

- The trend observed in the control group must approximate the trend that would have been observed in the treatment group, had they not received the treatment.
- Variations in original characteristics between control and treatment groups are okay.
  - Variations in *trends* between groups are *NOT* okay.
  - We want to be able to observe the effect of the treatment.
- Violation of parallel trends leads to biased estimates.

Difference-in-Differences (DID)

Some requirements for DID:

- You must be able to observe both the treatment and control groups:
  - The treatment group so you can observe the result of treatment.
  - The control group so you have a baseline comparison to estimate the full effect of the treatment.
- The control should serve as a counterfactual - it ought to tell you how the treatment group would look like if no treatment ever occurred.

Recreating Card and Krueger Using DID

- To recreate Card and Krueger, we'll again be using the **totalEmp** variables that we created for each time period.
- In order to capture the difference in total employment between time periods, create a **changeTotalEmp** variable that measures the change from `totalEmp` to `totalEmp2` using `mutate()`.

```
cardKrueger_clean <-
  cardKrueger_clean %>%
  mutate(changeTotalEmp =  #code to execute)
```

Recreating Card and Krueger Using DID

- We want to recreate rows 1, 2, and 4 from Card and Krueger's Table 3:

Variable	Stores by state: PA (i)	NJ (ii)	Difference, NJ − PA (iii)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (0.51)	−2.89 (1.44)
2. FTE employment after, all available observations	21.17 (0.94)	21.03 (0.52)	−0.14 (1.07)
3. Change in mean FTE employment	−2.16 (1.25)	0.59 (0.54)	2.76 (1.36)
4. Change in mean FTE employment, balanced sample of stores[c]	−2.28 (1.25)	0.47 (0.48)	2.75 (1.34)

- What steps do we need to take to produce these rows?

Recreating Card and Krueger Using DID

- Using `group_by()` and `summarise()` can you get the same state results as Card and Krueger?

```
empAvgs <- # data %>%
  group_by(# grouping variable) %>%
  summarise(preAvgEmp = # code to be executed),
            postAvgEmp = # code to be executed),
            avgEmpChange = # code to be executed))
```

Recreating Card and Krueger Using DID

- Now you can also calculate the differences in state average employment to find the difference between the "control" and "treatment" groups.

```
## # A tibble: 3 x 4
##   stateAbbrv preAvgEmp postAvgEmp avgEmpChange
##        <chr>     <dbl>       <dbl>         <dbl>
## 1         NJ     20.44       21.03          0.47
## 2         PA     23.33       21.17         -2.28
## 3    NJ - PA     -2.89       -0.14          2.75
```

How Fixed Effects Relate to DID

- We can avoid calculating differences manually, like we did for the previous table, by using fixed effects estimation
- Sure - making the manual differences did not take too long, but we only had two time periods to deal with. How would we find the DID estimate manually if we observed multiple periods both pre and post treatment?

DID in lm

```
lm(value ~ period*stateAbbrv, data = empl_PrePost)
```