Treatment Effects and Endogeneity

Today's Outline

- Treatment effects
- Difference-in-Differences

Treatment Effects

- Treatment effect models aim to measure the difference in some outcome of interest between a treatment and control group
- A *difference estimator* is a model that uses an indicator variable to distinguish between the two groups
- We call it a difference estimator because we assume any difference observed after the treatment is caused by the treatment

$$y_i = \beta_o + \beta_1 d_i + e_i$$

$$\text{Where } d_i = 0 \text{ if an individual is in the control group.}$$

What is the danger of assuming any difference observed after treatment is caused by the treatment?

Treatment Effects Example

- Below we construct a synthetic example where our random assignment assumptions hold
- Individuals are randomly assigned to a high milk diet and their heights are recorded after treatment
- By construction we should detect a difference between the groups of ~.5 inches

$$height_i = \beta_o + \beta_1 milk_i + e_i$$

Where $d_i = 0$ if an individual is not on the high milk diet.

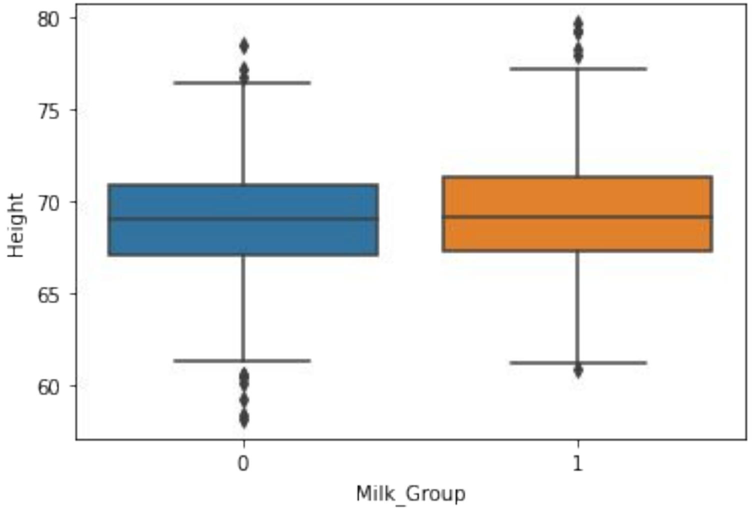
```python
# create synthetic data with two different meaned groups
# The first thousand observations will be the control and the second thousand are the treatment
heights = np.append(np.random.normal(69, 3, 1000), np.random.normal(69.5, 3, 1000))
```

```python
# Build Dataframe
height_data = pd.DataFrame([heights]).T
height_data.columns =["Height"]

# Assign the treatment and control groups an indicator variable
height_data["Milk_Group"] = 0
height_data.loc[1000: ,"Milk_Group"]= 1
```

Treatment Effects Example

- To calculate the difference we simply fit the model specified on the previous slide
- The only regressor is the dummy variable
- The coefficient on the dummy represents the treatment effect




```
treat_model = smf.ols('Height ~ Milk_Group', height_data).fit()
treat_model.summary()
```

OLS Regression Results

Dep. Variable:	Height	**R-squared:**	0.006
Model:	OLS	**Adj. R-squared:**	0.005
Method:	Least Squares	**F-statistic:**	11.11
Date:	Fri, 18 Nov 2022	**Prob (F-statistic):**	0.000876
Time:	09:05:19	**Log-Likelihood:**	-5077.6
No. Observations:	2000	**AIC:**	1.016e+04
Df Residuals:	1998	**BIC:**	1.017e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	69.0151	0.097	711.847	0.000	68.825	69.205
Milk_Group	0.4569	0.137	3.333	0.001	0.188	0.726

Omnibus:	0.300	**Durbin-Watson:**	2.036
Prob(Omnibus):	0.861	**Jarque-Bera (JB):**	0.361
Skew:	-0.017	**Prob(JB):**	0.835
Kurtosis:	2.943	**Cond. No.**	2.62

Treatment Effects With Fixed Effects and More Predictors

- If samples are randomly assigned between treatment and control we don't need additional regressors
- Sometimes other regressors may improve the estimator
- We can also include fixed effects if we expect that there are differences between groups but not *within* groups
  - In the example below we add "county" variables
  - In this example we may think it's possible that randomization occurs within counties but not on the counties themselves
  - Fixed effects will control for unobserved characteristics common to all individuals in a given county

```python
# add in some predictors
mom_height = np.random.normal(63.5, 2.5, 2000)
dad_height = np.random.normal(69, 3, 2000)
milk_treatment = np.random.choice([0,1], 2000)



height = .5 * mom_height + .5 * dad_height + .5*milk_treatment + np.random.normal(0,1, 2000)

data = pd.DataFrame([height, mom_height, dad_height, milk_treatment]).T
data.columns = ["height", "mom_height", "dad_height", "milk_treatment"]

# add in some counties
counties = ["Santa Barbara", "Ventura", "LA", "Orange", "Alameda", "Merced", "Riverside" "San Diego"]
data["county"] = np.random.choice(counties,2000)
```

```python
data.merge(pd.get_dummies(data.county), left_index = True, right_index = True)
```

	height	mom_height	dad_height	milk_treatment	county	Alameda	LA	Merced	Orange	RiversideSan Diego	Santa Barbara	Ventura
0	69.776905	66.720266	71.051910	1.0	Merced	0	0	1	0	0	0	0

Treatment Effects With Fixed Effects and More Predictors

- Results line up almost exactly as expected given the population model
- Milk treatment effect is .47 (close to .5)

```
treat_model2 = smf.ols('height ~ mom_height + dad_height+ milk_treatment+ county', data).fit()
treat_model2.summary()
```

OLS Regression Results

Dep. Variable:	height	R-squared:	0.802
Model:	OLS	Adj. R-squared:	0.801
Method:	Least Squares	F-statistic:	893.5
Date:	Thu, 17 Nov 2022	Prob (F-statistic):	0.00
Time:	21:35:42	Log-Likelihood:	-2805.8
No. Observations:	2000	AIC:	5632.
Df Residuals:	1990	BIC:	5688.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.8811	0.749	1.176	0.240	-0.589	2.351
county[T.LA]	-0.0024	0.083	-0.029	0.977	-0.165	0.160
county[T.Merced]	0.0542	0.085	0.637	0.524	-0.113	0.221
county[T.Orange]	0.0627	0.082	0.770	0.442	-0.097	0.223
county[T.RiversideSan Diego]	0.0916	0.083	1.105	0.269	-0.071	0.254
county[T.Santa Barbara]	0.1137	0.084	1.360	0.174	-0.050	0.278
county[T.Ventura]	0.0395	0.084	0.473	0.637	-0.124	0.203
mom_height	0.4766	0.009	54.238	0.000	0.459	0.494
dad_height	0.5084	0.007	70.906	0.000	0.494	0.522
milk_treatment	0.4743	0.044	10.728	0.000	0.388	0.561

Omnibus:	2.094	Durbin-Watson:	1.990
Prob(Omnibus):	0.351	Jarque-Bera (JB):	2.038
Skew:	0.048	Prob(JB):	0.361
Kurtosis:	3.123	Cond. No.	3.19e+03

Testing the Random Assignment Assumption

- We can use our regressors to predict whether someone is in a treatment or control group, then assignment is not random
- This simple test can tell us whether assignment is random
- If the F-statistic of the model is significant, then there is likely not random assignment

```
1 test_model = smf.ols('milk_treatment ~ mom_height + dad_height', data).fit()
2 test_model.summary()
```

OLS Regression Results

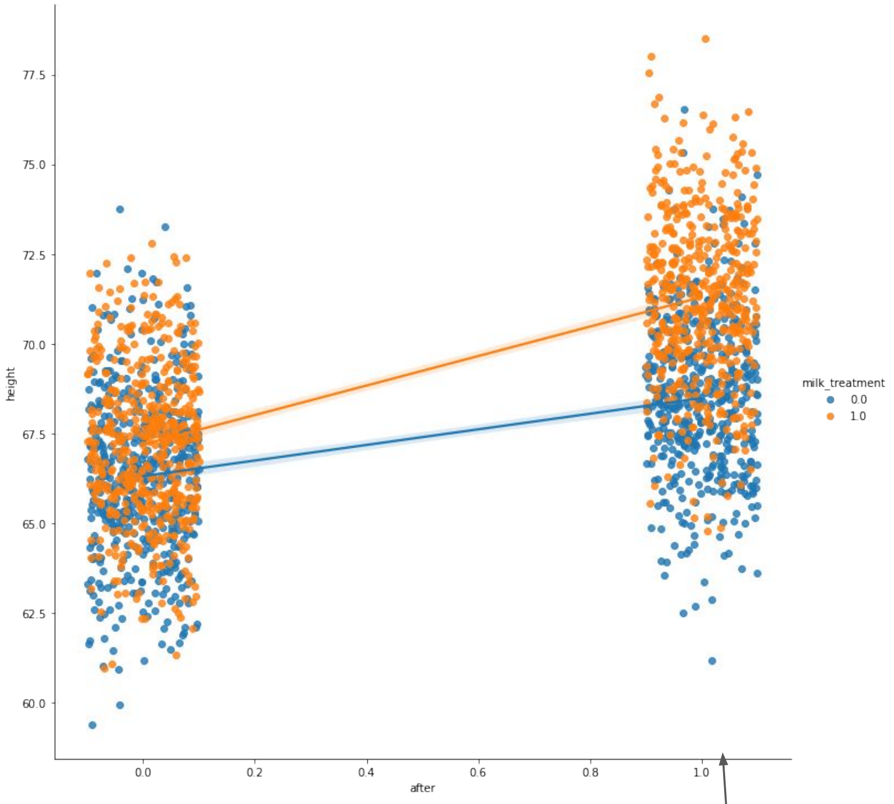
Dep. Variable:	milk_treatment	R-squared:	0.001
Model:	OLS	Adj. R-squared:	-0.000
Method:	Least Squares	F-statistic:	0.7239
Date:	Fri, 18 Nov 2022	Prob (F-statistic):	0.485
Time:	09:27:32	Log-Likelihood:	-1450.9
No. Observations:	2000	AIC:	2908.
Df Residuals:	1997	BIC:	2925.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.9556	0.379	2.521	0.012	0.212	1.699
mom_height	-0.0043	0.004	-0.957	0.339	-0.013	0.004
dad_height	-0.0027	0.004	-0.727	0.467	-0.010	0.005

Omnibus:	7176.049	Durbin-Watson:	1.972
Prob(Omnibus):	0.000	Jarque-Bera (JB):	332.369
Skew:	-0.004	Prob(JB):	6.72e-73
Kurtosis:	1.003	Cond. No.	3.18e+03

Difference in Differences

- Now let's suppose we have a more complicated situation
- The treatment and control groups are *not randomly assigned*
- Let's suppose in our previous example that people who are taller generally choose to be on high milk diets
- This means we start with a difference between the two groups that cannot be attributed to the treatment
- We also observe that over time people in both groups tend to grow taller



Before treatment

After treatment

Difference in Differences Estimation

- Want to estimate the treatment effect while accounting for the initial heterogeneity between the two groups
- This can be accomplished simply by regressing the output variable on:
  - A treatment dummy
  - A dummy indicating whether an observation was taken before or after treatment
  - An interaction term between the two dummies
- The treatment effect is then the coefficient on the interaction term
- Intuition:
  - *"After"* accounts for the change over time
  - "*Treatment*" accounts for the initial differences between the groups
  - The interaction accounts for the difference that can't be attributed to time or group

```
# Test is to regress our treatment on teh predictors
test_model = smf.ols('height ~ mom_height + dad_height + after*milk_treatment ', height_data).fit()
test_model.summary()
```

5]:

OLS Regression Results

Dep. Variable:	height	**R-squared:**	0.876
Model:	OLS	**Adj. R-squared:**	0.876
Method:	Least Squares	**F-statistic:**	2817.
Date:	Fri, 18 Nov 2022	**Prob (F-statistic):**	0.00
Time:	09:38:06	**Log-Likelihood:**	-2835.2
No. Observations:	2000	**AIC:**	5682.
Df Residuals:	1994	**BIC:**	5716.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	0.2657	0.770	0.345	0.730	-1.244	1.775
mom_height	0.5013	0.009	55.913	0.000	0.484	0.519
dad_height	0.4966	0.007	66.830	0.000	0.482	0.511
after	1.8670	0.064	28.988	0.000	1.741	1.993
milk_treatment	0.8374	0.063	13.366	0.000	0.715	0.960
after:milk_treatment	2.1333	0.090	23.815	0.000	1.958	2.309

Omnibus:	0.026	**Durbin-Watson:**	2.025
Prob(Omnibus):	0.987	**Jarque-Bera (JB):**	0.017
Skew:	0.007	**Prob(JB):**	0.992
Kurtosis:	3.002	**Cond. No.**	3.23e+03

What assumption is needed for DiD to be valid?

Exercise 1

A classic example of difference in differences studied the effect that a new garbage incinerator had on the prices of nearby homes.

1. Import *kielmc* from wooldridge
2. Estimate the treatment effect of *nearinc* on *rprice* while including age and age**2 as additional regressors. Is the treatment effect significant?
3. Is assignment to the treatment groups random?

Rumors that the incinerator would be built began after 1978, and construction started in 1981. If assignment is nonrandom then we can use whether or not a house was sold before 1978 or in 1981 and whether or not a house was near an incinerator to find the treatment effect.

4. Estimate a difference in differences model, where:
    a. *nearinc* is the treatment
    b. *y81* indicates if a house was sold before or after treatment
    c. Include age and age**2 as additional regressors

What is the treatment effect? Is it significant?