Difference in Differences Analysis: Seattle Soda Tax

Professor Song Yao
Olin Business School

Customer Analytics

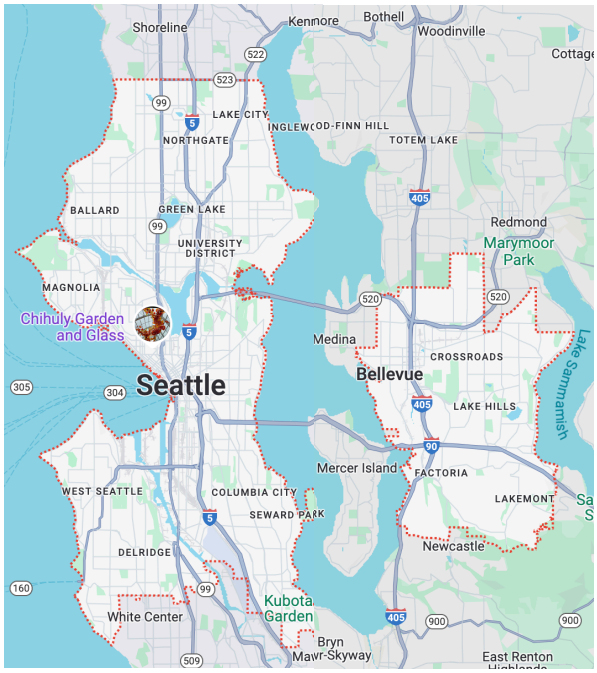
Soda Tax

- **Objective**: Reduce soda consumption; increase tax revenue
- 1.75-cent-per-ounce "soda tax"
- Taking effect on Jan 1, 2018 in Seattle, WA
- Increased the price of soda substantially

Does the tax really lower soda consumptions?

"Two Regions, Two Periods"

- Before and after the tax
- Seattle vs. Bellevue
  - Seattle has the soda tax, but Bellevue does not
  - But ***why not some nearby city?***



Data

- Households living in either Seattle or Bellevue
- Household weekly soda consumptions
- 9/1/2017 to 2/28/2018

```
# load the data
url = 'https://songyao21.github.io/course_data/seattle_soda_tax.csv'
soda_tax = pd.read_csv(url)

# randomly show 10 rows of the dataframe
soda_tax.sample(n=10, random_state=42)
```

	household_id	month	week	city	seattle	tax	ounces	log_ounces
76297	2935	12	14	bellevue	0	0	35.189999	3.560762
439351	16899	9	4	seattle	1	0	60.529999	4.103139
392104	15081	2	25	seattle	1	1	37.400002	3.621671
460616	17717	9	1	seattle	1	0	52.709999	3.964805
979	38	12	18	bellevue	0	1	30.440001	3.415758
90424	3478	2	23	bellevue	0	1	31.700001	3.456317
484005	18616	12	16	seattle	1	0	62.070000	4.128263
57323	2205	1	20	bellevue	0	1	30.900000	3.430756
422495	16250	1	22	seattle	1	1	44.869999	3.803769
366773	14107	12	18	seattle	1	1	43.150002	3.764683

Changes in consumption within Seattle before and after the tax

```
# 1.1 simple regression comparing seattle before and after tax
# using only seattle data
model = smf.ols('log_ounces ~ tax + C(month)',
                data=soda_tax[soda_tax['seattle'] == 1]).fit()
print(model.summary().tables[1])
```

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4.0421	0.001	3270.218	0.000	4.040	4.044
C(month)[T.2]	-0.0910	0.001	-127.518	0.000	-0.092	-0.090
C(month)[T.9]	-0.0678	0.001	-51.528	0.000	-0.070	-0.065
C(month)[T.10]	0.0112	0.001	8.405	0.000	0.009	0.014
C(month)[T.11]	0.0445	0.001	33.358	0.000	0.042	0.047
C(month)[T.12]	0.0399	0.001	35.379	0.000	0.038	0.042
tax	-0.2296	0.001	-203.455	0.000	-0.232	-0.227

$$\log(consumptions) = Intercept + Sep + Oct + Nov + Dec + Feb$$
$$\log(consumptions) = Intercept + Sep + Oct + Nov + Dec + Feb + Tax$$

Difference in consumption between Seattle and Bellevue

```
# 1.2 simple regression comparing seattle vs. bellevue after tax
# using only post-tax data
model = smf.ols('log_ounces ~ seattle + C(month)',
                data=soda_tax[soda_tax['tax'] == 1]).fit()
print(model.summary().tables[1])
```

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.5140	0.000	7669.520	0.000	3.513	3.515
C(month)[T.2]	-0.0864	0.001	-160.234	0.000	-0.087	-0.085
C(month)[T.12]	0.0389	0.001	45.626	0.000	0.037	0.041
seattle	0.2966	0.001	583.485	0.000	0.296	0.298

$$\log(consumptions) = Intercept + Dec + Feb$$
$$\log(consumptions) = Intercept + Dec + Feb + Seattle$$

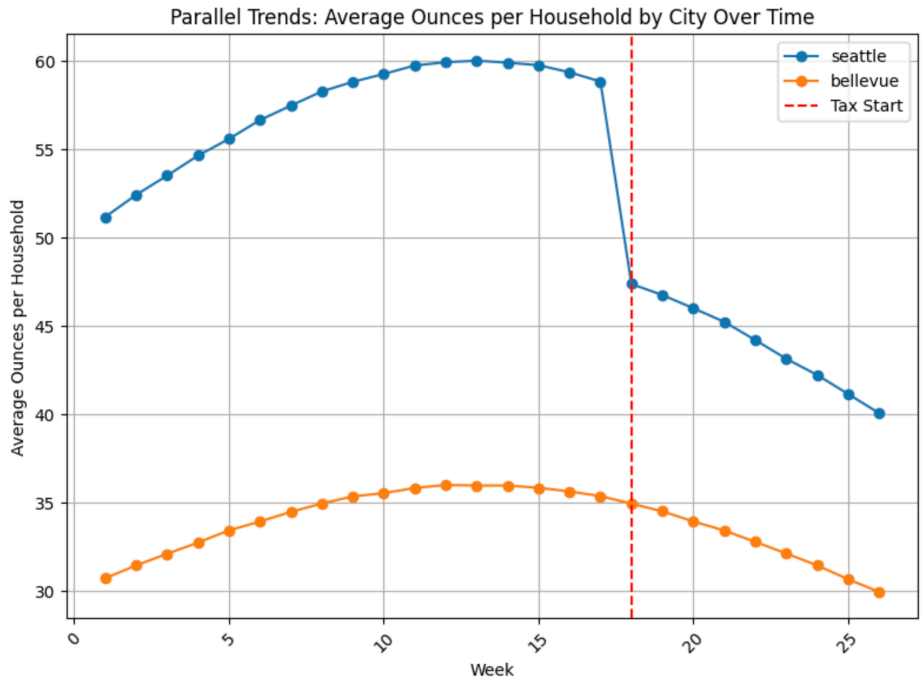
Diff-in-Diff

```
# 2. run the diff-in-diff regression
model = smf.ols('log_ounces ~ seattle + tax + seattle_tax + C(month)',
                data=soda_tax).fit()
print(model.summary().tables[1])
```

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.5324	0.001	4137.888	0.000	3.531	3.534
C(month)[T.2]	-0.0864	0.000	-178.678	0.000	-0.087	-0.085
C(month)[T.9]	-0.0690	0.001	-77.425	0.000	-0.071	-0.067
C(month)[T.10]	0.0098	0.001	10.834	0.000	0.008	0.012
C(month)[T.11]	0.0430	0.001	47.587	0.000	0.041	0.045
C(month)[T.12]	0.0389	0.001	50.878	0.000	0.037	0.040
seattle	0.5109	0.000	1540.535	0.000	0.510	0.512
tax	-0.0184	0.001	-22.620	0.000	-0.020	-0.017
seattle_tax	-0.2144	0.001	-380.252	0.000	-0.215	-0.213

$$\log(consumptions) = ... + tax$$
$$\log(consumptions) = ... + tax + seattle + seattle\_tax$$

Parallel Trends: Visual Inspection



Parallel Trends

```
parallel_trend2 = smf.ols('log_ounces ~ seattle * C(month)',
                          data=soda_tax[soda_tax['tax']==0]).fit()
print(parallel_trend2.summary().tables[1])
```

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.4634	0.000	8581.204	0.000	3.463	3.464
C(month)[T.10]	0.0786	0.001	129.867	0.000	0.077	0.080
C(month)[T.11]	0.1118	0.001	184.659	0.000	0.111	0.113
C(month)[T.12]	0.1081	0.001	178.576	0.000	0.107	0.109
seattle	0.5108	0.001	894.920	0.000	0.510	0.512
seattle:C(month)[T.10]	0.0004	0.001	0.468	0.640	−0.001	0.002
seattle:C(month)[T.11]	0.0005	0.001	0.634	0.526	−0.001	0.002
seattle:C(month)[T.12]	−0.0004	0.001	−0.455	0.649	−0.002	0.001

$$\log(consumptions) = Intercept + Oct + Nov + Dec$$

$$\log(consumptions) = Intercept + Oct + Nov + Dec + Seattle + Seattle\_Oct + Seattle\_Nov + Seattle\_Dec$$

Placebo Test

- What is a "placebo"?
  - A fake treatment that is intended to have no effect
  - Testing a vaccine or a new drug
    - Treatment group, receiving the real drug
    - Control group, no drug
    - Placebo group, receiving a fake drug that looks like the real one
- Placebo test in diff-in-diff
  - Move the "treatment" to a period or a region where the real treatment did not happen
  - There should be no significant effect.
    - If there is a significant effect, then the diff-in-diff is problemartic

Placebo test

First 3 weeks of Sep, 2017—No tax yet in Seattle

```
# 4. Create placebo tax column for three weeks in September (before tax)
soda_tax['placebo_tax'] = (soda_tax['week'] <= 3).astype(int)
parallel_trend2 = smf.ols('log_ounces ~ seattle * placebo_tax + C(month)',
                          data=soda_tax[soda_tax['tax']==0]).fit()
print(parallel_trend2.summary().tables[1])
```

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.4945	0.000	7330.157	0.000	3.494	3.495
C(month)[T.10]	0.0477	0.001	87.276	0.000	0.047	0.049
C(month)[T.11]	0.0809	0.001	148.142	0.000	0.080	0.082
C(month)[T.12]	0.0768	0.001	140.545	0.000	0.076	0.078
seattle	0.5109	0.000	1515.670	0.000	0.510	0.512
placebo_tax	-0.0519	0.001	-74.018	0.000	-0.053	-0.051
seattle:placebo_tax	2.786e-05	0.001	0.035	0.972	-0.002	0.002

No tax yet! So, the interaction should not be significant!