# Difference in difference & Synthetic Control Method : Application to sin food taxes.

Isaac Meza

November 5, 2021

# 1 DiD

## 1.1 Methodology

Data consists of an unbalanced panel of 128056 households for which we have data (at least) for the period spanning 2013 and the first 3 months of 2014. Our main variables are

1. Kcal of SD
2. Kcal of non-SD
3. Kcal of HCF
4. Kcal of non-HCF
5. Total taxable expenditure

Data is at week-individual level but we collapse it (by mean) at the monthly-individual level. We smooth all series using a MA with 3 lags, 2 forward terms and current observation; so that the smoother applied (by individual) is

$$(1/6)[x_{t-3} + x_{t-2} + x_{t-1} + x_t + x_{t+1} + x_{t+2}]$$

The first task is to define a pure treatment/control group. We will do this by choosing an *optimal* partition of the total taxable expenditure distribution, as high spenders will be more likely to be more sensitive to a price change in SD and HCF, therefore we will define this as the treatment group. Total taxable expenditure is defined as total expenditure in SD and HCF.

The *optimal* partition is found by solving the following problem

$$
\begin{aligned}
\min_{H,L} \quad & \sum_{t=-12}^{-2} |\beta_t| \\
\text{s.t} \quad & \\
& (\beta_t)_{-12\leq t\leq 12} = \text{argmin}\left\{\left(y_{it} - \sum_{k=-12}^{12} \alpha_k \mathbb{1}(t=k) - \right.\right. \\
& \qquad \left.\left. \sum_{k=-12}^{12} \beta_k \mathbb{1}(i=T, k=t) + \gamma \mathbb{1}(i=T) - \lambda_i \right)^2 \right\} \\
& T = \mathbb{1}(x_i \geq H) \\
& C = \mathbb{1}(x_i \leq L) \\
& \min(x_i) \leq L \leq H \leq \max(x_i)
\end{aligned}
$$

Note that the coefficients $\beta$ solve for a fixed effects regression including time calendar dummies and leads and lags in treatment effect[1] Moreover $(\beta_t)_{t=-12}^{-2}$ gives a 'test' on parallel trends, so what we are looking for is to find the optimal partition of Treatment/Control group so that a parallel trend is preserved. We do this in order to try to capture 'true' treatment effects.

Also note that in principle the partition is allowed to be non-symmetrical or to not span the whole distribution.

We use as a dependent variable $(y)$ total calories of SD and HCF and total calories of SD. Variable $(x_i)$ corresponds to total taxable expenditure of individual $i$. Finally, $H$ and $L$ are the respective cuts on the distribution to define Treatment/Control groups. Note that event study corresponds to $t = 0$.

Results

Once we find the optimal "cuts" on the distribution of total taxable expenditure and define our pure Treatment/Control group, we graph

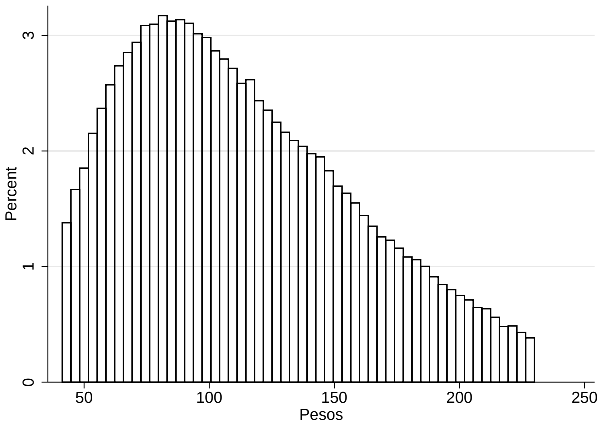
(a) The average calorie consumption throughout time by treatment group

[1] As recommended by Borusyak and Jaravel (2016), but unlike McCrary (2007) and most event study papers, include all relative time dummies in the regression rather than "binning" periods below $a$ or above $b$. Then we can just graph the periods from $a$ to $b$ if we want. But binning can cause bias if the trend isn't flat for periods less than $a$ or greater than $b$ (Borusyak and Jaravel, 2016). Note that when there is no pure control group, binning periods less than a or greater than b (i.e. imposing flat trend for those periods) is needed to pin down calendar time fixed effects, which is why Borusyak and Jaravel (2016) recommend having a pure control, which pin down the calendar time fixed effects without having to make these additional assumptions.

(b) The coefficients of the fixed effect regression with leads and lags

The following graph shows the distribution (cut at the 95th percentile) of the total taxable expenditure.

Figure 1: Distribution taxable expenditure

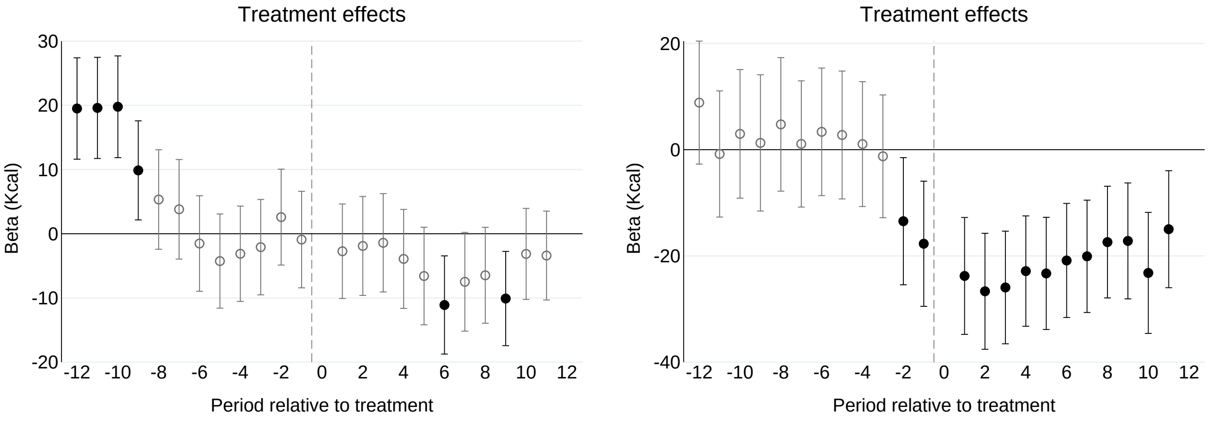


*Notes:* *Do file:* `dist_totalexp.do`

The DiD specification is the following:

$$y_{it} = \sum_{k=-12}^{12} \alpha_k \mathbb{1}(t = k) + \sum_{k=-12}^{12} \beta_k \mathbb{1}(i = T, k = t) + \gamma \mathbb{1}(i = T) - \lambda_i + \epsilon_{it}$$

Figure 2: Treatment effects


(a) HCF

(b) SD

*Notes: Do file:* `did.do` , `beta_coef_did.do`

# 2 SCM

## 2.1 Methodology

The data corresponding to the years 2013 and 2014 at the household-week level of the consumption of various products that were classified into food and beverages manually according to (Arturo's criterion). In order to have the same product classification, a product homologation was carried out between Mexico (MEX) and Central America (CAM), this with the purpose of having a better synthetic control for MEX households; keeping 11 drinks and 17 food. In this homologation, the units of consumption of the products were reviewed manually and for each one, as well as the price expressed in the same currency, taking care that there was consistency between the volume and price for the products between the two regions. For the missing values, an imputation was made according to a linear regression between volume and price with region fixed effects and linear time trend, for each of the products.

Likewise, two products were labeled as those subject to tax: soft drinks and cookies, since these are the products that were identified as those unambiguously subject to a tax, in addition to the fact that the tax rate was the highest for these products.

Weekly expenditure per household was added for each of the products and in order to smooth the series, it collapsed (on average) at the monthly level per household, so that the unit of observation remained as the average consumption and expenditure at the week for a month at the household level.

The panel was balanced so that we had households with data for the pre- and post-treatment date and the gaps between the data were filled in with a polygonal interpolation, that is, the average between the observation $t-1$ and $t+1$ to impute the value of the observation at $t$, when it was missing. Finally, a moving average with 3 lags, 2 terms ahead and the current observation was used, so that the smoother applied is

$$(1/6)[x_{t-3} + x_{t-2} + x_{t-1} + x_t + x_{t+1} + x_{t+2}]$$

This resulted in a balanced panel consisting of 7,401 households, of which 5447 were treatment units, that is, MEX households.

The next step consisted of finding clusters of households for separate treatment and control, that is, MEX and CAM:

Given the set $\{x_1, x_2, \ldots, x_n\}$ donde $x_i = (x_i^{-12}, x_i^{-11}, \ldots, x_i^{11})$ is a vector of consumption for the pre- and post-treatment periods, and $n$ corresponds to the number of households. The goal is to partition the $n$ observations into $k$ clusters $S := \{S_1, S_2, \ldots S_k\}$, so that the sum of squares within the clusters is minimized, formally:

$$\text{argmin}_S \sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu_i||^2$$

where $\mu_i$ is the median [2] of the points in $S_i$

This with the purpose of reducing the search space for the synthetic control, exploring the optimal weights between the clusters of CAM households and thus speeding up the computation of the optimization process. In addition, each cluster would reduce, by construction, the variance or dispersion of the variables of interest among the households that compose it and thus we would be left (at least in theory, with less sparse solutions).

This is how we obtain two databases (one for SD and one for HCF) with 400 clusters for MEX and 75 for CAM. Before proceeding with the estimation, the data were standardized at the household level.

To deal with the problem of multiple units treated, the methodology [3] proposed in [Xu16] or what in [AL18] was followed called *the standard synthetic control estimate*, provided a synthetic control is constructed for each treated unit and the treatment effect is estimated for the post-treatment periods and these effects are averaged to have an average treatment effect, and estimated the confidence region using a *Jackknife* estimator for the variance.

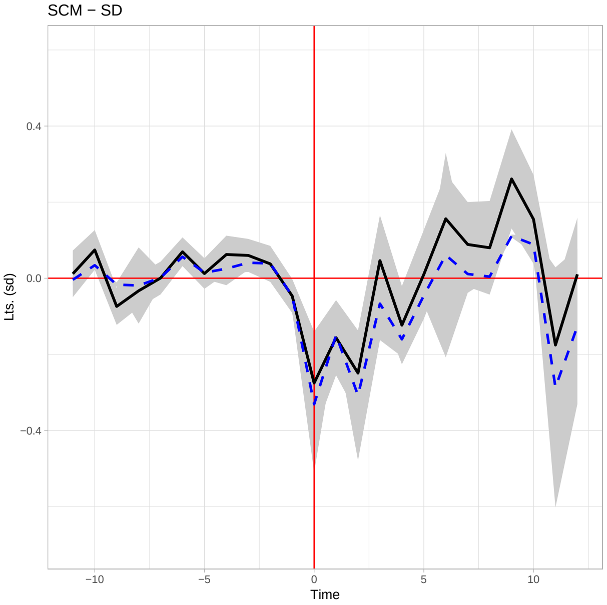
[2] If the mean is used instead of the median, the problem is equivalent to

$$\text{argmin}_S \sum_{i=1}^{k} |S_i| Var S_i$$

[3] Using the libraries Synth and bootstrap of R

Figure 3: Aggregation level : Household

(a) SD (std)



# References

[AL18] Alberto Abadie and Jeremy L'Hour, *A penalized synthetic control estimator for disaggregated data*, Preliminary (2018).

[Xu16] Yiqing Xu, *Generalized synthetic control method: Causal inference with interactive fixed effects models*, Political Analysis, Forthcoming (2016).