Synthetic Control Method : Obesity

Isaac Meza

November 5, 2021

1 Metodología

Los datos corresponden a los años 2013 y 2014 a nivel hogar-semana del consumo de diversos productos que se catalogaron en alimentos y bebidas de forma manual de acuerdo a (criterio de Arturo). Con el fin de contar con la misma clasificación de productos se realizó una homologación de productos entre México (MEX) y Centro América (CAM), esto con el propósito de tener un mejor control sintético para los hogares de MEX; quedándonos con 11 bebidas y 17 alimentos. En esta homologación se revisaron de forma manual y para cada uno, las unidades de consumo de los productos así como el precio expresado en la misma moneda, cuidando que hubiera consistencia entre el volumen y precio para los productos entre las dos regiones. Para los valores faltantes, se realizó una imputación de acuerdo a una regresión lineal entre volumen y precio con efectos fijos de región y tendencia lineal de tiempo, para cada uno de los productos.

Asimismo se etiquetaron dos productos como los sujetos a impuesto: bebidas gaseosas y galletas pues son estos los productos que se identificaron como aquellos sujetos sin ambigüedad a un gravamen, además de que la tasa impositiva era la más grande para estos productos.

Se sumó el gasto semanal por hogar para cada uno de los productos y con el fin de suavizar las series, se colapsó (en media) a nivel mensual por hogar, de modo que la unidad de observación quedó como el promedio de consumo y gasto a la semana por un mes a nivel hogar.

Se balanceó el panel de modo que tuviéramos hogares con datos para la fecha pre y post tratamiento y se completaron las brechas entre los datos con una interpolación poligonal, es decir que se tomo el promedio entre la observación $t-1$ y $t+1$ para imputar el valor de la observación en t, cuando esta estuviera faltante. Finalmente se usó una media móvil con 3 lags, 2 términos forward y la observación actual, de modo que el suavizador aplicado es

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

Esto resultó en un panel fuertemente balanceado consistiendo de 7401 hogares de los cuales 5447 eran unidades de tratamiento, es decir hogares de MEX.

El siguiente paso consistió en encontrar clusters de hogares para tratamiento y control por separado, es decir MEX y CAM:

Dado el conjunto $\{x_1, x_2, \ldots, x_n\}$ donde $x_i = (x_i^{-12}, x_i^{-11}, \ldots, x_i^{11})$ es un vector de consumo para los periodos pre y post tratamiento, y n corresponde al número de hogares. El objetivo es particionar las n observaciones en k clusters $S := \{S_1, S_2, \ldots S_k\}$, de forma que se minimiza la suma de cuadrados dentro de los clusters, formalmente:

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

donde μ_i es la mediana[1] de los puntos en S_i

Esto con el propósito de reducir el espacio de búsqueda para el control sintético, explorando los pesos óptimos entre los clusters de los hogares de CAM y así agilizar el cómputo del proceso de optimización. Además que cada cluster reduciría, por construcción, la varianza o dispersión de las variables de interés entre los hogares que la componen y así nos quedaríamos (al menos en teoría, con soluciones menos ralas).

Así es como obtenemos dos bases de datos (una para SD y otra para HCF) con 400 clusters para MEX y 75 para CAM. Antes de proceder con la estimación, se estandarizaron los datos a nivel hogar.

Para lidiar con el problema de la múltiples unidades tratadas se siguió la metodología[2] propuesta en [Xu16] o lo que en [AL18] llama *the standard synthetic control estimation*, esencialmente se construye un control sintético para cada unidad tratada y se estima el efecto de tratamiento para los periodos post-tratamiento y se promedian estos efectos para tener un efecto de tratamiento promedio, y se estimó la región de confianza usando un estimador *Jackknife* para la varianza.

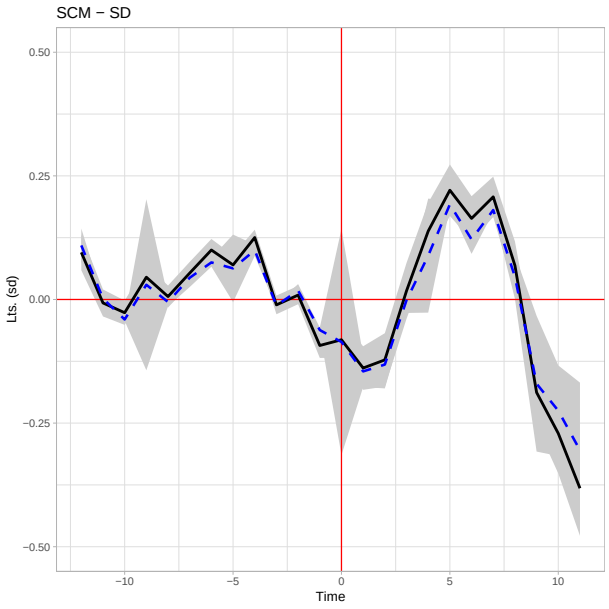
[1] Si en vez de la mediana se usa la media, el problema es equivalente a

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

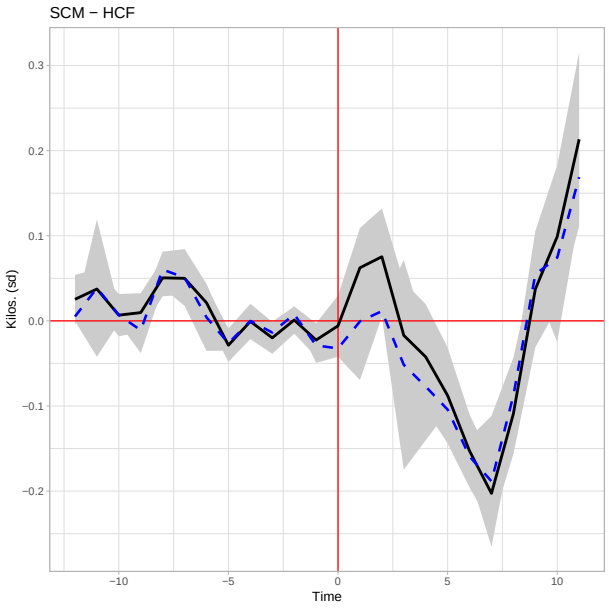
[2] Usando las librerias SYNTH y BOOTSTRAP de R

Figure 1: Aggregation level : Household

(a) SD (std)



(b) HCF (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).