Research Report: Simulating Ito 2005 AER paper "Asymmetric Incentives in Subsidies: Evidence from a Large-Scale Electricity Rebate Program".

Youenn Mandon - Manon Vallat

October 2024

1 Introduction

1.1 Motivation

In 2021, electricity and heat generation remained the world's largest greenhouse gas-emitting sector, accounting for 39% of total emissions from energy combustion. Decreasing these emissions is a priority to mitigate climate change. One way to tackle this problem is to improve the energy efficiency of housing across the globe, allowing inhabitants to reduce their energy consumption and cutting their emissions.

However, designing policies with this goal can prove quite hard. How to nudge individuals to invest in the efficiency of their housing? How to ensure there is no rebound effect nor free riding? All these issues are made even worse in a context of tight public spending where the efficiency of such policies is questioned. Many papers have tried to solve this conundrum by studying the cost-benefit effect of energy efficiency programs. [Ito15] is one of those studies. Drawing into the rich theoretical literature on the matter, it tries to evaluate the effective efficiency of a large-scale electricity rebate program in California. His idea is to test whether some subsidy can prove helpful in reducing the energy consumption of households.

1.2 Koichiro Ito's approach

1.2.1 Context

Ito's paper evaluates the effectiveness of California's "20/20" electricity rebate program implemented during the summer of 2005. This program was simple: every customer enrolled in the program could benefit from a 20% price discount on their monthly electricity bills if they reduced their summer 2024 electricity consumption by 20% compared to last year. In order to be eligible, customers had to open an account before a cutoff date in 2004. All the customers were automatically enrolled.

This program's effectiveness was quite controversial. Many argued that it was inefficient as it did not consider the changes in weather between summer 2004 and summer 2005 or some other random fluctuations. Some people benefited from the discount on electricity bills without making an effort to reduce their consumption but were lucky because 2004 was for them a year of high electricity consumption for various reasons. We should however keep in mind that this program was not announced in 2004, before the cutoff date. Thus, customers could not anticipate by increasing their electricity consumption in 2004 in order to get a discount the year after.

1.2.2 Research design

In his paper, Ito wants to measure the effect of the 20/20 electricity program on electricity consumption for consumers enrolled. To do so, he needs to isolate the effect of the program by controlling for individual and geographical characteristics.

To do so, he applies a RDD to the 20/20 electricity rebate program, exploiting the eligibility conditions. Indeed, being eligible after a cutoff date creates a discontinuity and this program was not possible to anticipate before this date, avoiding self-selection bias. Very importantly, all the customers who opened an account before the cutoff date were automatically enrolled. Thus, people who opened an account before and after this cutoff date can be considered very similar and the treatment is considered randomly assigned.

The first RD design explains y_{it}, the customer i's natural log of electricity consumption for month t, by a treatment dummy D_{it} and individual fixed-effects (θ_i) and time fixed-effects (λ_t) as in equation 1.

$$y_{it} = \alpha \times D_{it} + f_t(x_i) + \theta_i + \lambda_t + u_{it} \tag{1}$$

Here, x_i is the number of days between the date a customer i opened an account and the cutoff date. D_{it} is the treatment dummy made of two other dummies (equation 2)

$$D_{it} = D_i \times D_t \tag{2}$$

D_i and D_t are constructed as precised in equation 3.

$$D_i = \begin{cases} 1, & \text{if } x_i \leq 0, \\ 0, & \text{otherwise} \end{cases} \qquad D_t = \begin{cases} 1, & \text{if } t \in \text{treatment period}, \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

We will only focus on this regression design but there are some other regressions to explain the heterogeneous effects he found after running this first regression.

1.2.3 Data

This paper uses panel data of customer-level monthly electricity billing records for California's three largest electric utilities. It contains the information for customers who opened accounts only about a year before the treatment period began and includes geographical variables, income variables and weather variables.

1.2.4 Results

Many important results are presented in this article. First, the basic RDD shows an important heterogeneity in the evolution of electricity consumption across regions. In inland areas, where summer temperatures are relatively high and incomes relatively low, consumers enrolled in the program consumed 4% less electricity. However, there is no statistically significant effect for coastal areas where there are moderate summer temperatures and relatively high incomes.

To explain this heterogeneity in the treatment effects, he estimates the interaction effect between the treatment variable and income levels, air conditioner saturation rates, and climate conditions. The results suggest that the treatment effect increases with higher temperatures and air conditioner saturation rates and decreases with income.

Then, he measures the "giving-up effect", i.e. the potential differences in treatment effect on consumers far from the target of 20% electricity reduction. Indeed, he retrieves that the treatment effect is not significantly different from zero for consumers who are far from the target level.

Last, he explores to which extent the results still hold for consumers who opened accounts earlier and that are not accounted in his sample.

1.3 Our approach

Ito's data were obtained under confidentiality agreements and unavailable to us. Thus, we will not be able to observe whether our fake dataset is similar to the true one, which will limitate our potential analysis. We implement a research design based on Ito's design but adapted to our specific situation where we generate a fake dataset following a precise data-generating process. We will detail our choices in the second part of this report.

2 Simulation

2.1 Goal

Our first objective is to have a better understanding of the type of data used by Ito in his article in order to understand how he finds his results, even if we cannot replicate the paper. With this simulation, we also want to understand the importance of bandwidth selection in the context of an RD design. Specifically, the idea is to visualize the trade-off between bias and statistical power inherent in RD designs. Choosing a narrower bandwidth can improve the quality of our estimators but significantly reduces the sample size, which in turn diminishes statistical power. Conversely, increasing the bandwidth provides more observations but may introduce bias. It is worth noting, however, that in Ito's 2005 article, the author examines the impact of bandwidth size on these estimators and demonstrates that the very large number of observations in his dataset mitigates this trade-off.

Thus, we also aim to measure the impact of the number of observations on the results of our RDD and determine at what sample size an RDD yields reliable estimators. This is particularly relevant given that, even with a large initial sample, our regression ultimately involves a much smaller subset of observations.

A last goal would be to visualize how heterogeneity in the data could affect the estimates we find by artificially generating heterogeneous effects.

2.2 The overall structure of our simulations

Since the real data is inaccessible to us, we choose to generate data based on the information provided in the article about the various variables. We also decide to start by generating data using a very simplified DGP, which we can progressively make more complex to better understand the challenges associated with generating such data. We make the following simplifications:

- We simulate data for a single electricity company (whereas the 20/20 rebate program involves three). Specifically, we generate a simulation for SCE's customers, whose number is approximately 3,500,000 as of June 2004, according to the article.
- We assume that 30,000 accounts are opened each day at SCE in California. This corresponds to the account opening rate for this company reported in the article. However, we simplify by assuming that it is impossible to close an account once opened, although this is, of course, possible in reality. Consequently, the total number of customers will be 3,680,000 by December 2004, where our database ends (calculated as 3,500,000 + 30,000 * 6).
- We assign account opening dates by month rather than by day.
- We allocate account opening dates only to individuals potentially within a maximum bandwidth of 6 months. With the cutoff date in June 2004, we focus on a much smaller dataset containing customers who opened accounts with SCE between January 2004 and December 2004. Based on the previous figures, the number of individuals within the bandwidth is approximately 360,000. This represents a very small subset compared to the initial base of 3,680,000 customers.

We simulate two DGPs and consider a third potential DGP at the end of our document.

2.3 First simple simulation

We begin with a very simple DGP, which will allow us to understand the various challenges associated with more complex and realistic data generation processes.

2.3.1 Generated variables description

In this very simple DGP, we generate the following variables:

- i (or H_{id} in our code): customer's ID;
- t (or *month* in our code): number between 1 and 24 to represent the month between January 2004 and December 2005;
- kWh_{it}: electricity consumption of customer i for month t;
- $date\_account_i$: the date at which customer i opened an electricity account (only attributed to a subsample);
- $income_i$: customer's mean income over a year;
- $monthly\_income_{it}$: the income of customer i for month t;
- D_i: dummy for having an account opened before the cutoff date (thus when $date\_account_i$ is between 1 and 6);
- D_t: dummy for the treatment period, which starts in May 2005 ($t = 15$);
- D_{it}: dummy for treatment.

2.3.2 Data generation

The causal diagram representing the DGP is presented bellow.

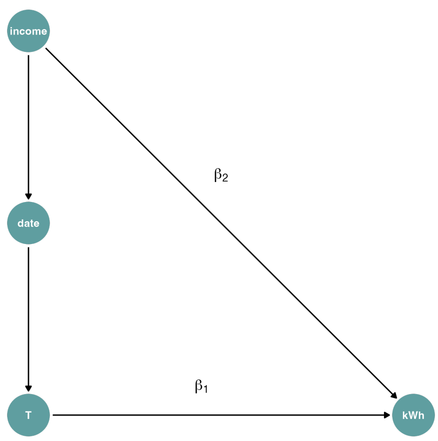

Figure 1: Very simple DGP

We generate our three variables kWh, income and date_account the following way:

- **Income**: it is drawn from a log-normal distribution for every household. We then add a noise for every observation of the time period to obtain the *monthly_income* variable.

- **Date_account**: as the opening dates are the variable behind the treatment take up, and we want the treatment not to be random in the more complicated model (in this case we could use a simple DiD framework rather than a RDD), we correlate the opening date of electricity account with the income variable. In this first simulation, only the top 10% richest households (q = 0.9) are selected to open an account in 2004, others have a 0 value.
- **kWh**: it is a function of the treatment dummy and of income, such that:

$$kWh = \beta_0 + \beta_1 * D_{it} + \beta_2 * monthly\_income + model\_noise$$

Calibration choices were made using available data on median household income in California and average monthly electricity consumption in California in 2005.

2.3.3 Results and conclusion

After plotting distributions to observe that we indeed correctly generated our data, we manage to retrieve the effect of the treatment with a very simple regression. We also show that restricting the database with different bandwidths doesn't change significantly the estimator value. As expected in this first simulation, the RDD is not particularly relevant in the context of the simple DGP (see Figure 2).

	bandwidth	intercept	estimate	slope_significance
1	1	-0.34580510	-59.98008	***
2	2	-0.18569210	-60.02680	***
3	3	-0.06303594	-60.03003	***
4	4	-0.09378751	-60.04845	***
5	5	-0.11215572	-60.03124	***

Figure 2: Trying different bandwidths

Applying fixed effect, we find exactly the same results. This is normal as fixed effects is tantamount to some sort of intersect here where it represents the average consumption without any treatment.

Finally, as shown in Figure 3, we managed to iterate 30 times our model on a smaller sample with estimators always centered on the true effect.

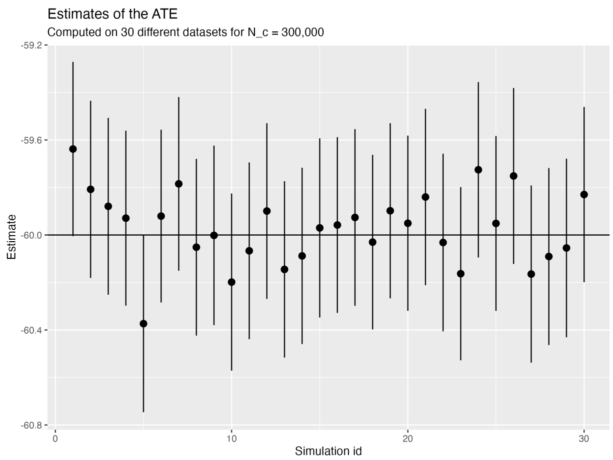

Figure 3: Iterating our model 30 times

3 Second more complex simulation

The second model is aimed to provide a better background for an RDD. Indeed, the treatment was still random among the richest households. We need to create a non-random allocation of treatment to test different sizes of bandwidth so the RDD will be irrelevant past a certain bandwidth as individuals won't be comparable. Unfortunately, we did not have time to implement these features. Here are some insights at what we would have liked to try.

3.1 Allocating richer people around the cutoff

The first attempt consisted in adapting the DGP of the first model by allocating the richest people around the cutoff to a certain point. We tried to reach this goal modifying the previous code, as explained in the Quarto document. Unfortunately, it did not produce any difference with the first model and we had no time left to investigate further.

3.2 Using propensity scores

In fact, the best way would be to construct a categorical variable to reach this goal. Each month every household could have a certain chance of opening an account. The opening date of the account being a dummy (1 if the household opened the account this month, 0 if not) we might imagine a model which given the vector of characteristics, attributes a probability of opening an account. Households with the same characteristics would open an account at the same time, avoiding the selection into treatment issues.

Unfortunately, we did not have time to explore this issue and try implementing a categorical variable.

4 Third simulation

The third simulation was meant to be the most comprehensive, using much of the information contained in the article. It is a more complex approach to the issue, using many variables.

In this model, we split customers between a coastal and an inland area following a binomial distribution configured using the distribution shown in the article (300,000 customers living in inland areas, 3,200,000 living in coastal areas).

As California has roughly 35,000,000 billion inhabitants, our sample represents (very) roughly 10% of the population. This way, customers are then allocated between 260 zip codes (10% of California's zip codes). Those that were coastal are allocated the first 52 zip codes (20% of the area), while the inland zip codes occupy the rest of the zip codes, from 53 to 260.

Weather is generated with two vectors of monthly temperature, one from Ridgequest, an inland city, and the other from Los Angeles airport. Both are monthly average temperatures. These temperatures are measured by 37 of California's 370 weather stations. The weather stations are randomly matched to zip codes following a random distribution. On average, there shall be a weather station for 7 zip codes. Every month, the weather station draws a temperature from a normal distribution centered on the average temperature vector. This temperature is then associated with all the zip codes within the area of this weather station.

Figure 4 shows the DGP we tried to implement.

However, we did not have time to generate our dataset with this complex DGP and have thus not any results to analyze.

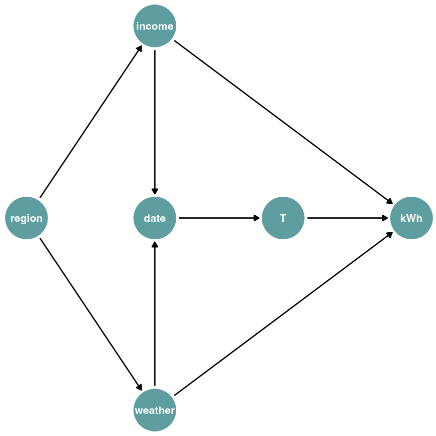

Figure 4: More complex DGP

5 Conclusion

To conclude, we were able to simulate a huge dataset representing the 3,700,000 customers of the article from which we started. From this dataset, we were able to understand the potential limitations of an RD design, even if the article clearly considered the possibility of having the wrong size of bandwidth. In this case, the large number of observations is a clear argument in favor of an RDD. Moreover, testing simulations with way fewer customers (we divided the number by more than 100) still gave us pretty good estimates over multiple iterations in the context of our first simple DGP, which reinforces the credibility of the RD design.

However, we were not able to implement the second simulation, where the RD approach would have made much more sense. One of the method we could use to do so would be to use categorical variables. For instance, we could compute a certain propensity score for every month which, when activated, defines the opening date of the account.

Finally, the third model was only partially implemented, with only its structure. We had not enough time to try to add our findings from the two first models in its data generating process. It would be interesting to continue this project to see if the RD and the fixed effects still works in a more complex framework.

References

[Ito15] Koichiro Ito. "Asymmetric Incentives in Subsidies: Evidence from a Large-Scale Electricity Rebate Program". In: *American Economic Journal: Economic Policy* 7.3 (Aug. 2015), pp. 209–37. DOI: 10.1257/pol.20130397. URL: https://www.aeaweb.org/articles?id=10.1257/pol.20130397.