Introduction to R Programming

Slide Set 7: Application 2 - Applied Study

Maria Ptashkina

Barcelona GSE ITFD

September 2021

Learning Objectives

- Content
  - Applied study using difference-in-differences method
  - Linear regression
- R-specific learning objectives
  - Data wrangling
  - Piping
  - Looping
  - Introduction to regression syntax in R

Applied Study Design

- We will estimate the effects of the 2014 sugar tax in the US: in November 2014, the city of Berkeley in California implemented a tax on the distributors of sugar-sweetened beverages (which are considered unhealthy)
- The tax of one cent per fluid ounce meant that if retailers raised their prices to exactly counter the effects of the tax, a $1 can of soda (12 oz) would now cost $1.12. But did sellers actually respond this way?
- In order to answer the question we compare the outcomes of two groups, both before and after the policy took effect: the treatment group (those who were affected by the policy), and the control group (those who were not affected by the policy)
- Specifically, we take the difference in outcomes of the treatment and control group before the policy was implemented, and compare it with the difference in outcomes after the policy was implemented (Difference-in-Differences method)

Housekeeping and Study Setup

- Create an R project, clean your workspace, load the necessary packages
- In this exercise you will be working with the data from the Global Food Research Program, the ‘Berkeley Store Price Survey’ dataset
- The dataset in Excel is provided to you (‘sps_public.xlsx’)
- Open the Excel file and look at the ‘Data Dictionary’ to understand the variables and information the dataset contains
- Import the Excel sheet ‘Data Dictionary’ into a `var_info` data frame
- Import the Excel sheet ‘Data’ into a `dat` data frame, inspect the object

Task 1: Cleaning and Inspecting the Data

- Check the classes of the variables
- `R` classified all the variables containing numbers as numerical (`num`). However, for some of these variables (specifically, `type`, `taxed`, `supp`, `store_id`, `store_type`, `type2` and `product_id`), the numbers actually represent categories (factors). Use the factor function to convert the variables to factor variables (you can also use labels to specify the names of the categories)
- Check the categories of the variable `time`
- The third price survey was in March 2016, not in March 2015, so the data has been labelled incorrectly. Change all the values `MAR2015` to `MAR2016`
- Check the class of the `time` variable
- Change `time` into a factor variable

Task 1: Cleaning and Inspecting the Data (cont.)

- Count the number of unique stores, store the number in an object `no_stores`
- Count the number of unique products, store the number in an object `no_products`

Task 2: Creating Frequency Tables

- Create a frequency table showing the number (count) of store observations (store type) in December 2014 and June 2015, with 'store type' as the row variable and 'time period' as the column variable. For each store type, is the number of observations similar in each time period? (use `tally()` function)
- Repeat the same exercise using `group_by()` and `count()` from `tidyverse`
- Organize the table in the same way as `tally()` function by adding `spread()` to the previous command

Task 2: Creating Frequency Tables (cont.)

- Create a frequency table showing the number of taxed and non-taxed beverages in December 2014 and June 2015, with ‘store type’ as the row variable and ‘taxed’ as the column variable (‘Taxed’ equals 1 if the sugar tax applied to that product, and 0 if the tax did not apply). For each store type, is the number of taxed and non-taxed beverages similar?
- Create a frequency table showing the number of each product type (type), with ‘product type’ as the row variable and ‘time period’ as the column variable. Which product types have the highest number of observations and which have the lowest number of observations? Why might some products have more observations than others?

Task 3: Calculate and Compare Conditional Means

- We want to calculate the average price per ounce (in cents) for taxed and untaxed beverages separately
- But we should only include products that are present in all time periods, and non-supplementary products (`supp = 0`)
- We first need to identify products (`product_id`) that have observations for all three periods
- Create a new variable called `period_test`, which takes the value 1 (`TRUE`) if we have observations in all periods for a product in a particular store, and 0 (`FALSE`) otherwise
- How many products were not observed in all three periods?
- Use the `period_test` variable to remove all products that have not been observed in all three periods, and call the new data frame `dat_c`

Task 3: Calculate and Compare Conditional Means (cont.)

- We are interested in comparing the mean price of taxed and untaxed beverages, before and after the tax (the conditional mean)
- Calculate the means of `price_per_oz` by grouping the data according to `store_type`, `taxed`, and `time`, store the results in table `table_res`

Task 4: Comparing the Change in Price

- Calculate the the two price differences (June 2015 minus December 2014 and March 2016 minus December 2014) for taxed and untaxed beverages, by store type; add them as columns to `table_res`
- Plot a column chart for average price change from December 2014 to June 2015
- Plot a column chart for average price change from from Dec 2014 to Mar 2016

Task 5: Statistical Significance

- To assess whether the difference in mean prices before and after the tax could have happened by chance due to the samples chosen (and there are no differences in the population means), we could calculate the p-value
- Use price difference between June 2015 and December 2014 in Large Supermarkets for taxed beverages: extract the prices for both periods (call them vectors `p1` and `p2`) and then calculate the difference, element by element (store as `d_t`)
- For `d_t` to correctly represent the price difference for a particular product in a particular store, we need to be certain that each element in both vectors corresponds to the same product in the same store. To check that the elements match, we will extract the store and product IDs along with the prices, and compare the ordering in `p1_alt` and `p2_alt`

Task 5: Statistical Significance (cont.)

- Calculate the Student's t-test statistic manually
- Do the same by using `R`'s built-in function `t.test` (forst on `p1` and `p2`; and then directly on the differenced vector `d_t`)
- Repeat the exercise for the price difference between June 2015 and December 2014 in Large Supermarkets for untaxed beverages
- Finally, test the statistical significance of the difference between the price changes in taxed and untaxed products (the difference in differences)

Task 6: Linear Regression

- Select the variables `store_type`, `time`, `taxed`, and `price_per_oz`, and choose only time periods corresponding to December 2014 and June 2015; assign the selection to a data frame called `reg`
- Take a look at the summary table which calculates the number of observations in each time and tax status
- Create dummy variable that takes value 1 if time period corresponds to June 2015 and zero otherwise; and a dummy variable for taxed that takes a value 1 if the observation is taxed
- Run a difference-in-differences regression controlling for a store type
- Interpret the results

References and Resources

- Doing Economics ▸ Tutorial
- R Tutorial: Difference-in-Differences (DiD) ▸ Tutorial