The Causal Effect of A Small Business Training Program on Business Outcomes

Kavina Peters

October 22, 2025

Abstract

Causal inference shows larger revenue gains on average for small businesses ($\leq$ 100 employees) who chose to participate in a business training program (pop.size approx 180) when compared to those who did not (pop.size approx 90). Backdoors involving observed variables were closed by controlling for confounding variables by subclassification of firms into 12 classes based on 3 sectors and 4 revenue levels. To discount the effect of any general economic trends - during the training and evaluation period - that similarly affect both adopters and nonadopters, a difference in differences approach (between adopters and non-adopters before and after the training program) was used to provide a regression-checked robust estimate of causation. The data did not indicate confounding effect of other pertinent observed variables such as (volatality in) number of employees, per-capita wage, wage bill to revenue ratio, service to sales revenue ratio. The subclassification ensured that these variables were not controlled, in order to close any backdoors arising from their potential (not confirmed) collider effect. The question remains open whether useful finer grained causation estimates can be obtained involving these and other pertinent variables, such as duration of adoption (instead of a Boolean variable), and higher order gradients in business outcome gains.

1 Introduction

Since small businesses are often cash-strapped, a careful evaluation of the efficacy of expensive manager training programs is valuable. Available data are generally not obtained by randomized controlled trials; rather, they document before and after business outcomes (e.g. revenue and sales estimates) for voluntary adopters and non-adopters of training programs. For such voluntary adoption scenarios, the tested causal inference approach is difference in differences, testing for robustness using regression. This factors out the effect of general economic trends affecting both adopters and non-adopters during the period of the training program -

However, the real challenge in proving the causal relationship between adoption and improvement in business outcomes is to close backdoors, i.e., control the effect of several potential confounding variables that may influence both the choice to adopt and the business business outcome. Furthermore, controlling for these variables should not inadvertently open other backdoors, for example by controlling any potential collider variables. Such collider variables are affected both by adoption and improved business outcomes: and controlling them could erroneously indicate causation where not even correlation exists.

In this paper, we show, using careful causal inference, that larger revenue gains on average for small businesses ($\leq$ 100 employee) who chose to participate in a business training program (pop.size approx 180) when compared to those who did not (pop.size approx 90). Backdoors involving confounding variables were closed by controlling these variables by subclassification of firms into 12 classes based on 3 sectors and 4 revenue levels, while not opening any other backdoors involving potential collider variables.

2 Data

The first step of any data analysis project is data cleaning. In this case, there were many reporting errors in the data set that would have clouded the empirical results. The first issue arose with the firm IDs, where some firms seemed to be identical except for an extra '00' at the end of their ID. Therefore

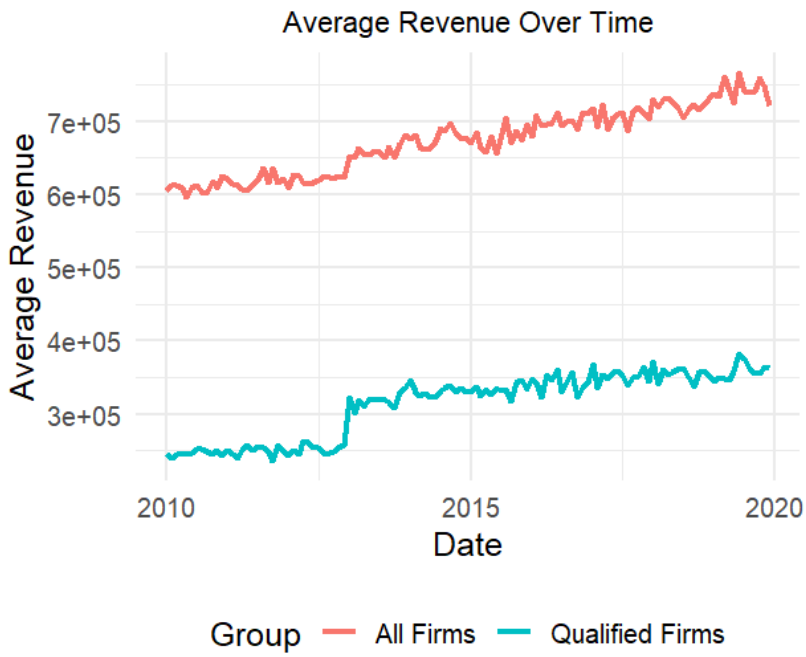

Figure 1: This Figure shows the difference in Average Revenue when we include unqualified firms, and when we remove them from our dataset.

these differences were consolidated and a total of 181 unique adopting firms were found from 2013 January to 2019 December. The second problematic reporting error was that many dates were not reported in clear date time format, which would cause problems when classifying groups as 'before and after' treatment. For these errors, it was determined that these dates could be replaced as the first day of the month (ie 2010-01-01) for each month.

The second data cleaning task was removing all rows of companies that had more than 100 employees. These companies were not qualified for the Small Business program, and keeping them in the non-adopters group would be a confounding factor as companies with larger number of employees are also likely to have a larger revenue, sales, or other variable that would remove credibility from a conclusion of the causal effect of the program.

The third data cleaning task was to consider missing values, deciding which to remove and which to replace by interpolation.

3 Empirical Strategy

The Directed Acyclic Graph in Figure 2 shows the underlying model of backdoor paths resulting from confounding observed variables (indicated by the data), and potential collider variables: the data did not indicate these as confounding variables, but also did not confirm them as colliders.

3.1 Controlling for Confounding Variables by Subclassification

Subclassification of firms based on sector and revenue class (at the end of the evaluation period) is used to control for these confounding variables. 12 classes (3 sectors and 4 revenue classes) of firms are used.

An algorithm to visualize the causation processes the panel data of firm monthly revenues to assess the impact of adoption (Boolean) of the training program (at any time, for any nonzero length of time) on firm revenue. Firms with more than 100 employees in Jan 2013 at start of training program are excluded. Each months revenue data are split into the above mentioned 12 classes. Average revenue across firms in a class is tracked over time and adopter and non-adopter groups are compared.

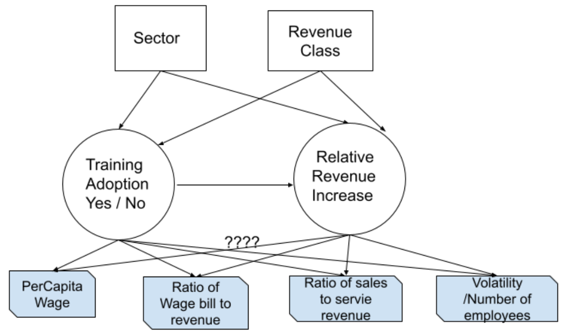

Figure 2: Causal Model showing Confounding Variables and Potential Colliders

3.2 The Difference in Differences Analysis

In order to prove causation between the intervention, i.e., adoption (for any nonzero period) of the Small Business Training Program, and the outcome variable, we use a Difference-in-Differences (DD) model. The Outcome Variable I chose is Average Revenue. We represent our DD model as:

$$Y_{it} = \beta_0 + \beta_1 T_i + \beta_2 P_t + \beta_3 (T_i \times P_t) + \epsilon_{it}$$

Where:

- Y_{it} is the outcome (Revenue) for individual i in time period t.
- T_i is a dummy variable indicating the adopting group ($T_i = 1$ if adopt.t=1, $T_i = 0$ otherwise).
- P_t is a dummy variable indicating the post-treatment period ($P_t = 1$ if post-treatment, $P_t = 0$ if pre-treatment).
- β_0 is the intercept, representing the average outcome for the control group in the pre-treatment period.β_1 is the difference between the treatment and control groups in the pre-treatment period.β_2 is the change in the control group's outcome from the pre- to post-treatment period.β_3 is the DiD estimator, which measures the treatment effect.
- ϵ_{it} is the error term.

3.3 Robustness Check

In order to ensure robustness, we conduct a sensitivity assessment by testing confounding factors. The DD model assumes 'Parallel Trends', meaning that had the treatment not occurred, the difference between the two groups over time would have stayed the same. Therefore we must ensure that there are no unobserved variables in the treatment group, that change over time and affect the outcome.

The Confounding Variables that were indicated by this dataset are:

1. The industry Sector (Fast Food, Retail, Hospitality),
2. The Revenue class (four classes, firms with revenue from 0-500k, 500k-1m, from 1m-1.5m, and from 1.5m-10m).

Controlling for each of these confounding factors, 12 DD estimates were computed.

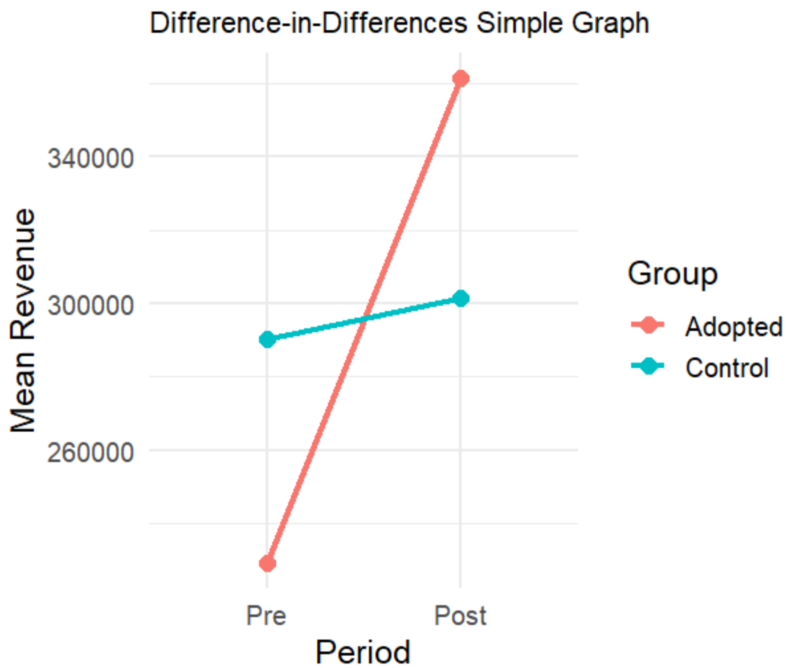

Figure 3: Simple Diff in Diff representation combining all sectors and over all Revenue Classes; y axis shows average monthly revenue, while the x axis shows the timepoints at the beginning of the training program and at the end of the evaluation period

Industry Sector	Diff in Diff estimate
Retail	132589
Hospitality	164959.5
Fast Food	58083.84

Table 1: Table shows DD estimates for Three scenarios

4 Empirical Results

Figure 4 visualizes causation across the 12 class subclassification described in the previous section to control the confounding variables.

Figure 5 shows the same visualization as Figure 5 except that it also includes the nonadopters who were ineligible (i.e. had more than 100 employees). We included this figure to show an more pronounced causation.

When we estimate the Difference-in-Differences value over all sectors, and over all Revenue classes, we find a positive β_3 value, meaning that higher revenue gains are observed in adopting firms compared to non-adopters. This value is also statistically significant when we conduct a one-tailed t-test.

5 Conclusions

There are two potential concepts that were considered but discarded. The first is integrating Two-Way fixed effects with differential timing. Because there are late-adopters of the program (ie after Jan 2013), the start date for each adopting firm could have affected the outcome (revenue). However, out of the 181 adopting firms, 174 adopted in January 2013, meaning that only 7 adopted later, a negligible outlier group.

The second concept was broadening the number of outcome variables that could be even more affected by the training program such as ratio of sales to service revenue, ratio of wage bill to revenue, Wage Per Capita, or volatility in number of employees. In fact, although not confirmed by the data,

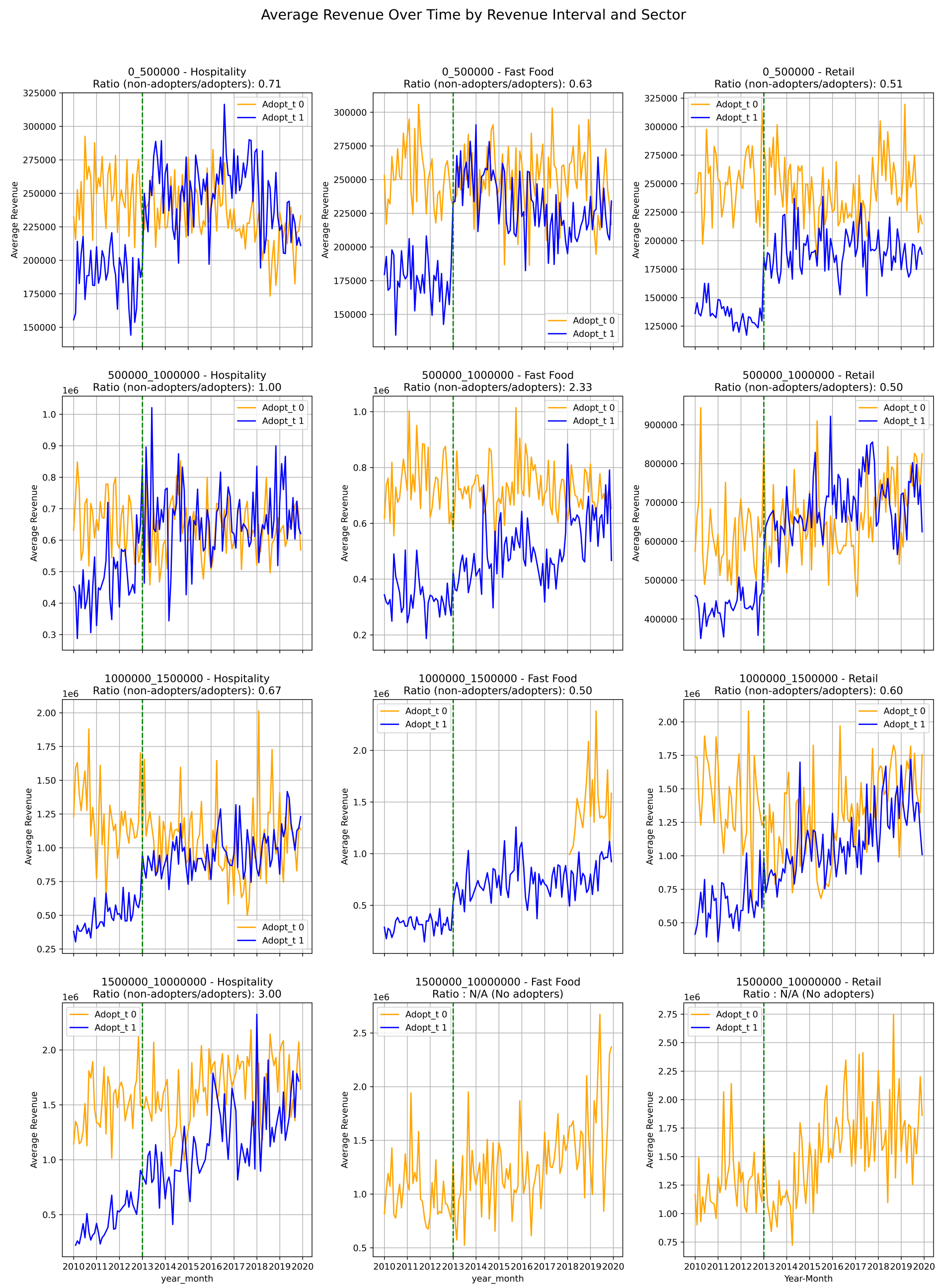

Figure 4: After the introduction of the training program (intervention, Jan 2013, green vertical line), adopting firms show higher revenue growth across sectors and revenue brackets. **Obs 0:** Firms with more than 100 employees in Jan 2013 are excluded since the training is meant for small firms. Firms with missing employee data are treated as eligible. **Obs 1:** Revenue intervals were based on data from `2019_11.csv`, which had the highest number of firm IDs. In some cases (e.g., Fast Food in top revenue bracket), meaningful comparisons were not possible. **Obs 2:** Adopters generally had lower revenue at baseline (i.e., pre-2013). **Obs 3:** Plot titles include the ratio of non-adopters to adopters, determined using 2019 data. **Obs 4:** No non-adopters were found in the Fast Food sector in the \$1M–\$1.5M revenue class before 2017.

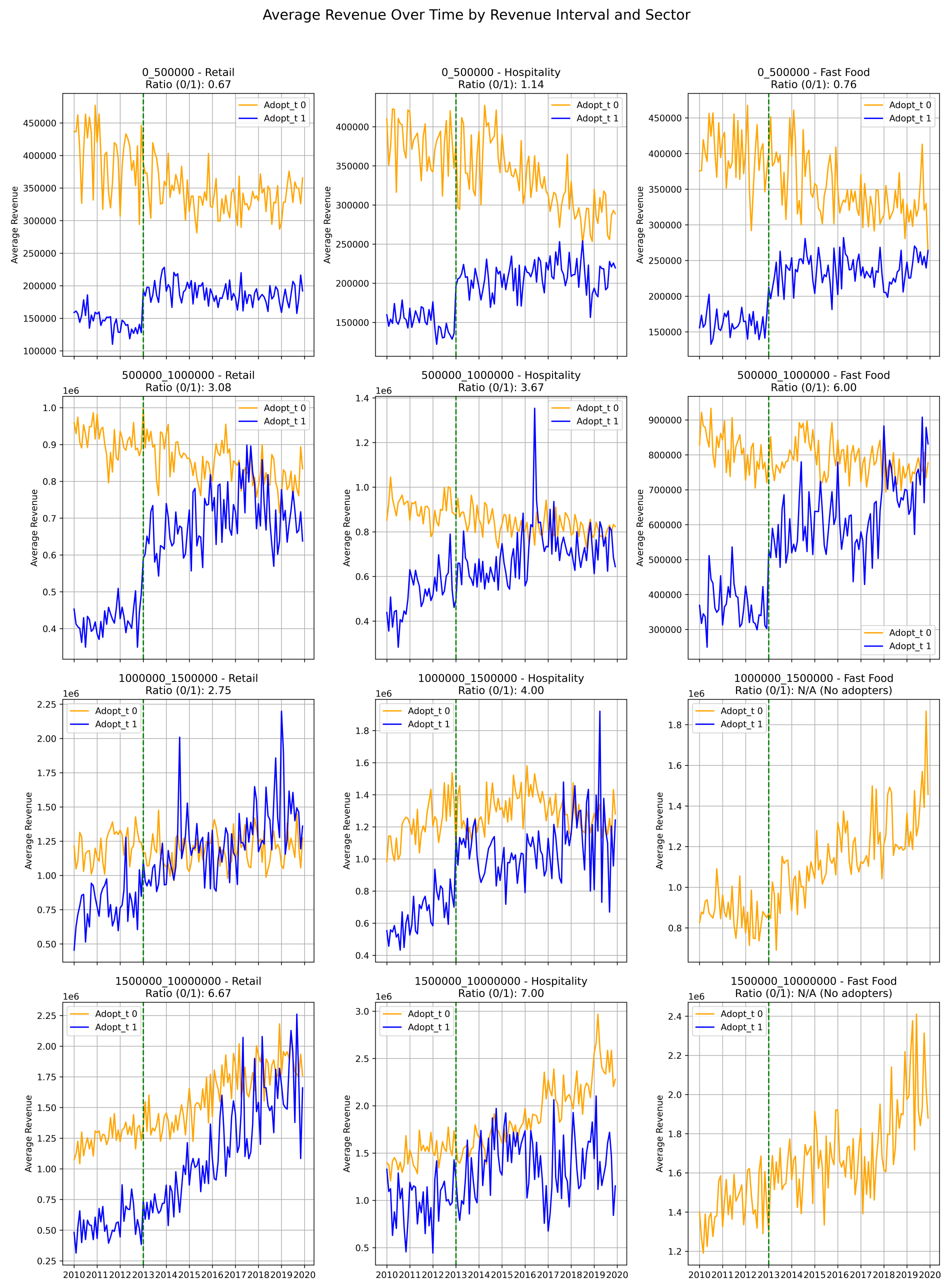

Figure 5: Firms with more than 100 employees in Jan 2013 are included in the nonadopter set. After the introduction of the intervention (Jan 2013, green vertical line), adopting firms show higher revenue growth across sectors and revenue brackets. **Obs 1:** Revenue intervals were based on data from `2019_11.csv`, which had the highest number of firm IDs. In some cases (e.g., Fast Food in top revenue bracket), meaningful comparisons were not possible. **Obs 2:** Adopters generally had lower revenue at baseline (i.e., pre-2013). **Obs 3:** Plot titles include the ratio of non-adopters to adopters, determined using 2019 data.

these variables were ultimately treated as potential collider variables in the model. The question remains open whether useful finer grained causation estimates can be obtained involving these and other pertinent variables, such as duration of adoption (instead of a Boolean variable), and higher order gradients in business outcome gains.

This paper's 'primary' estimate of the program's causal affect is that over the evaluation period Average Monthly Revenue of adopters increased by over 130,000 USD while Average Monthly Revenue of nonadopters increased by less than 10,000 USD. This difference in differences is sustained over all sectors and over all revenue classes. I.e., after controlling for these confounding factors, every subclass where comparison was possible showed the same trend of increasing revenue due to the program, and for a policymaker, it is likely to be the most useful pair of numbers, with which to recommend that the program should be offered to firms in the future.

6 Appendix: Visualization Algorithm Pseudocode

Algorithm 1: Pipeline to analyze revenue trends by firm adoption status within revenue intervals and sectors

```
Data: CSV data files for monthly firm data, firm information, and aggregate firm sales
Result: Plots of average revenue over time by adoption status and sector, within revenue intervals
Step 1: Filter Ineligible Firms
Read 2013_1.csv as DataFrame df2013_1
Identify firms with employees > 100, store their firm_id in list in_eligible_firms
Print: number of such firms and average number of employees in this group
Step 2: Read and Filter Monthly Data
foreach CSV file except aggregate_firm_sales.csv and firm_information.csv do
    Read file as DataFrame moDataAll[csv_file]
    Remove rows with firm_id in in_eligible_firms, store as moData[csv_file]
end
Step 3: Split Monthly Data by Sector
foreach moData[csv_file] do
    Split into three DataFrames by sector: Hospitality, Fast Food, Retail
    Store as sectorMoData[csv_file, sector]
    Print 3 random headers from each
end
Step 4: Process Aggregate Sales Data
Read aggregate_firm_sales.csv into aggregate_firm_salesAll
Filter out rows with firm_id in in_eligible_firms, store as aggregate_firm_sales
Split by sector into sector_aggregate_firm_sales[sector]
Print 1 random header per sector
Step 5: Define Revenue Intervals and Prepare for Analysis
Define revenue intervals: [0, 500,000], [500,001, 1,000,000], [1,000,001, 1,500,000], [1,500,001, 10,000,000]

Identify the moData[csv_file] with the most rows (call it maxRowFile)
foreach sector, revenue interval do
    Split sectorMoData[maxRowFile, sector] by revenue into revSectorMoData[maxRowFile,
      sector, interval]
    Further split by adopt_t ∈ {0, 1, NaN}
end
Step 6: Extract Adoption Groups
foreach sector, interval, adopt_t do
    Extract list of firm_id as revIntervalAdopt[sector, interval, adopt_t]
end
Step 7: Compute Average Revenue
foreach csv_file, sector, interval, adopt_t do
    Select rows from revSectorMoData[csv_file, sector, interval] with firm_id ∈
      revIntervalAdopt[sector, interval, adopt_t]
    Store as everadopt[csv_file, sector, interval, adopt_t]
end
foreach csv_file, sector, interval, adopt_t ∈ {0, 1} do
    Compute average of revenue_t, store as avgRev[csv_file, sector, interval, adopt_t]
end
Print: avgRev[2010_1.csv, Hospitality, 0_500000, 1] and avgRev[2010_1.csv, Hospitality,
  0_500000, 0]
Step 8: Plot Results
foreach sector do
    foreach revenue interval do
        Create subplots: x-axis = year.month (chronologically), y-axis = average revenue
        Orange line for non-adopters, Blue line for adopters
        Add vertical green line at 2013_1 (training start)
        Title each plot with adoption ratio: #non-adopters / #adopters for corresponding sector and interval
    end
    Arrange plots in grid: rows = revenue intervals, columns = sectors
end
```