Do Training Programs Boost Productivity? Evidence from Small Businesses

By Chengrui Wang*

This paper evaluates the effect of a small business training program on firm productivity level launched in January 2013 using monthly firm-level data. I construct a balanced panel from administrative sources by dealing with anomalies in datasets and imputing missing values with Gaussian Processes. Since this program is voluntary, I address selection bias using propensity-score-weighted difference-in-differences method (PSM-DID) to estimate the causal effect. I also apply nearest-neighbor matching DID, and doubly robust DID to check robustness of results. Across different specifications, the program significantly increases small firms' labor productivity by 40% on average. Among three sectors, smaller firms in hospitality experience greatest productivity increase, boosting both efficiency within firms and equity across firms.

Researchers have long utilized different ways to understand how to boost firm productivity level. Among them, the small business training program which consisted of several training sessions for firms' managers began on January 1st, 2013, targeted at increasing firm productivity level. Also, this program targeted small firms with fewer than 100 employees and firms opt to be enrolled in the program.

Due to the reporting errors in these administrative datasets, I firstly fixed firm IDs and date in firm sales and auxiliary data. Also, missing values in firm sales, employment, wage bill and revenue are further analyzed in detail. I tried both dropping the missing observations and imputing these missing values, with former referred as "*Dropped Data*" and latter referred as "*Imputed Data*".

To utilize the balanced panel data structure, I employed a univariate Gaussian Process (GP) regression to predict those missing values. (Schulz, Speekenbrink and Krause, 2018) Posterior means are obtained to fill in these missing values.

I also performed a few sanity checks for program details. Of 498 firms, 174 adopt in January, 2013, 7 adopt in other years or months and 371 didn't participate in the program. Since most firms adopt the program when the program began, we only consider the effect towards these firms and drop firms adopting this program in other times. Also, firms with over 100 employees on adopting year are also dropped as this violates the program eligibility requirement. Thus, 487 firms are left in "*Dropped Data*" while 485 firms in "*Imputed Data*". Moreover, no firms opt out of the program after adopting, so I don't need to consider that issue in

* Wang: Fudan University (vincentwcr041219@gmail.com). I declare that I have no relevant or material financial interests that relate to the research described in this paper. Note that background information, literature review and contribution are omitted in this working draft.

this paper.

Since those programs are voluntary and selection bias might exist(i.e., more promising small firms are more incentivized to take part in the training program), I build weights of each observation with propensity score method. Identification further relies on Difference-in-Difference(DID) with weighted least square estimation. Event analysis and placebo tests are also performed.

Heterogeneous effects for firms in different sectors and sizes are further analyzed. We observe greater effect for smaller size firms and those in hospitality sector, which is in line with McKenzie and Woodruff (2014).

Moreover, robustness checks are performed with different weighting methods and empirical strategies.

I. Data

Firm level data are given, including firm ids working as a primary key between datasets. Firm name, sector, sales, employment, wage bill, revenue and whether the firm adopts the small business training program are given.

A. Basic Firm Information

Inside basic firm information, a single correspondence between firm id, firm name and the sector that the firm belongs to is given. All ids and firm names are unique in this dataset. In total, 498 different firms are given in the dataset, with 209 in the retail sector, 149 in hospitality and 140 operating in fast food.

B. Firm Sales by Month

There exist extra 0's and spaces in firm ids, so I clean it such that the identification here is the same as in *Basic Firm Information*, which enables me to bring in sector information for each firm. Monthly sales timing are harmonized to start of month with dates converted to datetime format.

Also, I observe 2100 missing firms sales in the 54546 observations. For further analysis, I either drop missing entries or impute them, generating two datasets called "*Dropped Data*" and "*Imputed Data*", which requires the assumption of missing completely at random and missing at random respectively.[1]

Specifically, missing sales are imputed within firm using Gaussian Processes with linear interpolation fallback. Mathematically, Let y_t denote the target variable (here firm sales) observed at monthly time index $t \in \mathbb{Z}$. For firm i, we assume a GP prior:

$$y_t \sim \mathcal{GP}\big(0,\, k(t,t')\big)\,,$$

with covariance kernel:

$$k(t,t') \;=\; \sigma^2 \exp\!\left(-\frac{(t-t')^2}{2\ell^2}\right) \;+\; \sigma_\varepsilon^2 \mathbf{1}\{t=t'\},$$

[1]For the difference between the two assumptions, see Bhaskaran and Smeeth (2014).

which is a squared–exponential (RBF) kernel plus an independent noise term. The model is estimated by maximum likelihood with up to three restarts, normalizing the target variable within firm. The monthly date is mapped to an integer month index to form the one–dimensional input.

For each firm, we sort observations by date, split them into observed indices $\mathcal{T}_{\text{obs}}$ and missing indices $\mathcal{T}_{\text{mis}}$, estimate kernel hyperparameters on $\mathcal{T}_{\text{obs}}$, and use the GP posterior mean to impute y_t for $t \in \mathcal{T}_{\text{mis}}$. When a firm has fewer than two observed points or the GP optimization fails to converge, we fall back to linear interpolation from both directions to ensure a complete series.

This approach exploits within–firm temporal smoothness while allowing flexible, data–driven trends and firm–specific noise levels. It does not pool information across firms, thereby preserving heterogeneity. Identification relies on a missing-at-random assumption conditional on time for each firm, and the RBF kernel encodes the belief that nearby months have more similar values.

C. Firm Auxiliary Data by Month

In the auxiliary data, I firstly clean the firm id issue same as in firm sales data. Next, I observe that date variables are sometimes misspecified, so I change the date in one file all to the first date of that year_month indicated by the file name.

Additionally, employment, wage bill, and revenue are also missing sometimes, so I apply the same two strategy as in firm sales data after appending all entries from separate files.

Finally, all information for a firm in a single month are merged using the firm id, generating cleaned version of "*Dropped Data*" and "*Imputed Data*".

D. Sanity Checks

With the cleaned version of datasets, I want to verify details listed in the small business training program with real data.

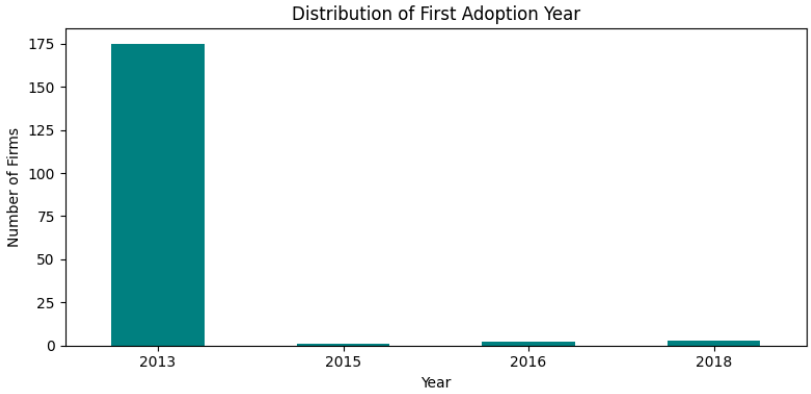
First, I want to check which year adopting firms adopt this program. Figure 1 shows adoption year of adopting firms. Since below 10 firms out of over 180 adopting firms adopt this program on dates other than January 1st, 2013, I drop those special cases in the datasets.

Second, I check that no firms drop out of the program after adopting this program.

Third, I analyze whether adopting firms vary across sectors. Figure 2 show that the shares of firms adopting this program in different sectors are relatively similar, making heterogeneity analysis in different sectors available.

Finally, I test if all firms adopting this program have employment number below 100 at the time of adopting this program. Figure 3 show that a few firms with firm size over 100 adopt this programs, so I also drop them. Also, from comparison of dropped data and imputed data we could observe that the distribution remains quite similar, demonstrating validity of our imputation process.

Figure 1. : Adoption year of firms

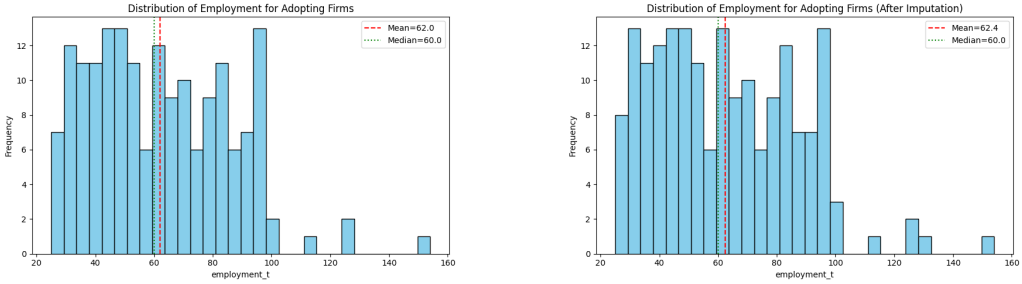


Note: Adoption year distribution for both "*Dropped Data*" and "*Imputed Data*" since adoption year is never missing. Of 181 firms that adopt this program, only 7 adopt on date other than January 1st, 2013.

Figure 2. : Adoption year of firms

Note: Adoption sector distribution for both "*Dropped Data*" and "*Imputed Data*" since firm sector is never missing.

Figure 3. : Distribution of firm size over adopting firms



Note: The left figure shows distribution for dropped data, while the right shows distribution for imputed data. Note that they are quite similar, demonstrating good imputing quality.

II. Specification

I estimate the causal effect of the program on firm productivity using a difference-in-differences (DID) design weighted by stabilized inverse propensity scores.

Let i index firms and t index months. The outcome is $y_{it} = \log(\text{productivity}_{it})$, proxied by the log of sales per employee. Since firms could choose to participate in the program or not, the selection bias might be a severe problem.

A. Propensity Score Model

To mitigate selection into treatment, I estimate firm-level propensity scores using only pre-treatment information. For each firm, we construct pre-2013-01 covariates: means and standard deviations of employment, revenue and wage bills pre-2013, the 2012 means of those variables, and sector indicators. Denote this vector by X_i. We fit a logistic regression

$$p_i \equiv \Pr(\text{treat}_i = 1 \mid X_i) = \text{logit}^{-1}\big(\beta_0 + X_i'\beta\big),$$

with numeric covariates standardized and sector entered via one-hot dummies. I trim extreme scores to the common support $p_i \in (0.05, 0.95)$.

Then, I construct stabilized weights at the firm level,

$$w_i \;=\; \begin{cases} \dfrac{\bar{p}}{p_i}, & \text{if treat}_i = 1, \\[2ex] \dfrac{1-\bar{p}}{1-p_i}, & \text{if treat}_i = 0, \end{cases}$$

where $\bar{p} = \mathbb{E}[\text{treat}_i]$ in the trimmed sample. These weights are then merged back to the panel so that each observation (i, t) receives w_i.

B. DID Specifications

The main weighted DID specification is:

$$y_{it} \;=\; \alpha_i \;+\; \gamma_t \;+\; \tau\,\mathrm{treat}_{it} \times \mathrm{post}_{it} \;+\; \mathrm{Sector}_i \times t \;+\; \varepsilon_{it},$$

where α_i denotes the firm fixed effects, γ_t denotes the time fixed effects, and $\mathrm{Sector}_i \times t$ is the sector trend.

This equation is estimated by weighted least squares (WLS) using w_i and firm-clustered standard errors.

C. Identification and Inference

Under the assumptions below, the coefficient $\hat{\tau}$ identifies the average causal effect of the program on log productivity for the propensity-score-weighted population within common support. Inference is based on heteroskedasticity-robust standard errors clustered at the firm level.

(i) Overlap: $0 < p_i < 1$ for firms used in estimation (enforced by trimming). (ii) Conditional parallel trends: in the absence of treatment, treated and control firms would have followed the same average outcome trends conditional on firm fixed effects, time fixed effects and sector trends. (iii) Stable unit treatment value (no interference) and consistent measurement. (iv) Correct specification of the propensity score sufficient to achieve covariate balance relevant for selection on observables.

III. Main Results

For both dropped and imputed data, I apply the PS-DID specification. Since the results for both data are similar, with imputed data showing better convergence, I will show tables and figures of the imputed data here and leave those of dropped data into the appendix.

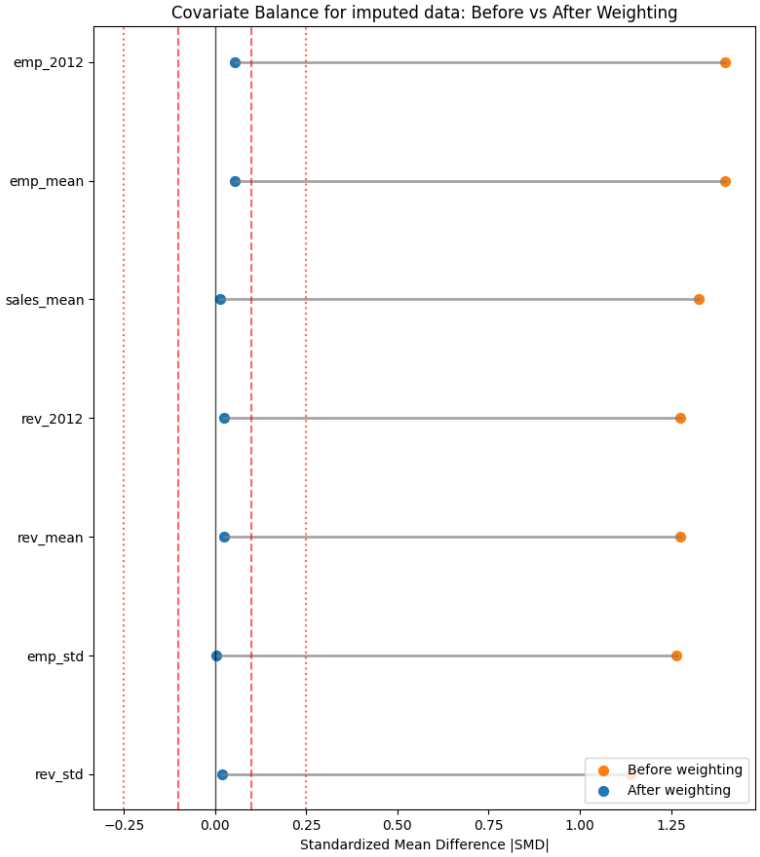
After reweighting on samples, we get the following love plot.(Figure 4) This graph shows that this propensity score balancing is significantly driving down the standardized mean difference.(See Haukoos and Lewis (2015) for more details on propensity score diagnostics.)

Table 1 shows that after the program, firms' productivity level increases by 39.4% standard deviation.

This is also clear shown from Figure 5. Firms adopting the program experience a significant positive productivity shock while those not participating in the program didn't. Note that control groups have higher productivity level than treated groups as bigger firms are usually more productive, which is in line with previous literature.

Furthermore, I check the parallel trend with event analysis and placebo tests, shown in Figure 6 and Figure 7. These figures further show the unconfoundedness of our estimates.

Figure 4. : Love plot for dropped data



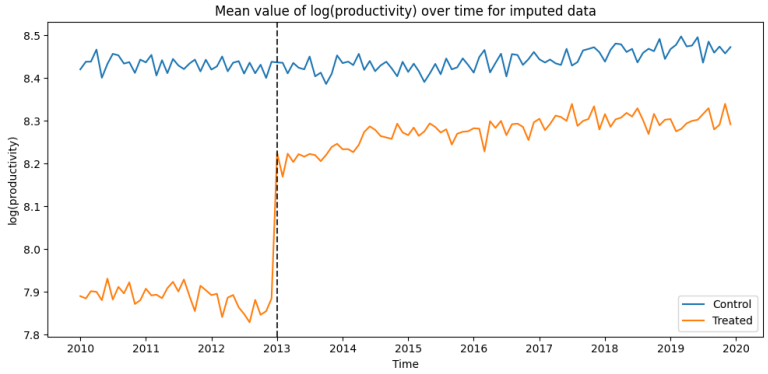
Note: After reweighting, the standardized mean difference between treated and untreated sample drops significantly, below the bar 0.1 shown in the red dashed line.

Table 1—: PSM-DID Estimation Results for imputed data

	(1)	(2)	(3)	(4)
Post×Treat	0.461***	0.373***	0.394***	0.394***
	(0.090)	(0.026)	(0.037)	(0.036)
Firm fixed effects	✗	✓	✓	✓
Time fixed effects	✗	✗	✓	✓
Sector × Time trend	✗	✗	✗	✓
N	41,840	41,840	41,840	41,840
R^2	0.076	0.855	0.856	0.857

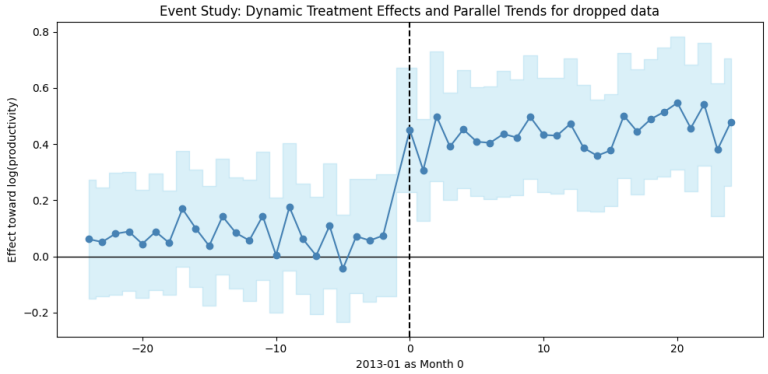
Note: Dependent variable is the logarithm of productivity, and all models use propensity score stabilized weights (WLS), with standard errors clustered at the firm level. Robust standard errors are reported in parentheses.
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Figure 5. : Log-productivity over time



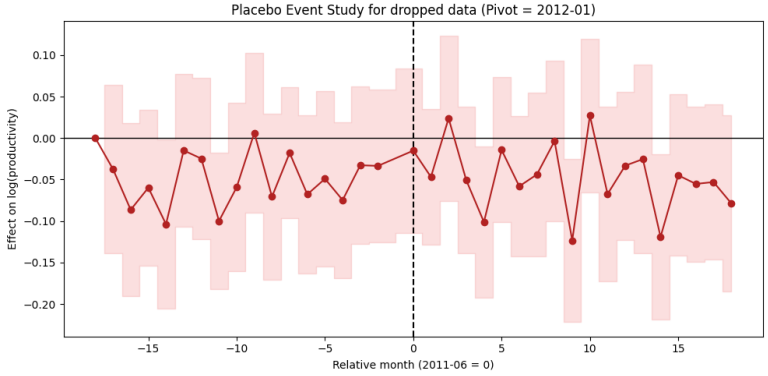
Note: Productivity level of treated firms increases dramatically from the plot, while firms not adopting the program didn't experience a significant positive productivity shock.

Figure 6. : Event analysis for casual effect



Note: Before the shock happens the program effect is not significant from zero, while afterwards it shows a significant positive trend. Blue shaded area is the 95% confidence interval.

Figure 7. : Event analysis for placebo tests



Note: Assume a fake program on 2011-06 and apply the same analyzing strategy gives all effect insignificant, further demonstrating validity of results. Red shaded area is the 95% confidence interval.

A. Heterogeneity

I check the heterogeneous effect of this program on firms in different sectors and size bins.

Table 2 shows estimates by sector. Effects are consistently positive across sectors, with hospitality showing greatest effect. This might be due to relative importance of managing skills in this sector.

Table 2—: PSM-DID Estimation Results by Sector

	Fast Food	Hospitality	Retail
Post×Treat	0.395***	0.422***	0.371***
	(0.071)	(0.063)	(0.056)
Firm fixed effects	✓	✓	✓
Time fixed effects	✓	✓	✓
Sector × Time trend	✓	✓	✓
N	12,060	11,993	17,787
R^2	0.852	0.828	0.869

Note: dependent variable is specified as log(productivity). Each column corresponds to a regression in the respective sector sample weighted by PS stabilized weights, with standard errors clustered at the firm level.
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Next, we separate firms into four employment size bins and estimate effects accordingly. Table 3 shows that effects are strictly positive and significant, with smaller firms showing greater impact. Since smaller firms usually have lower productivity level, this program boosts not only firm efficiency but also equality between firms.

B. Robustness

To check for robustness of our estimates, I change the matching to derive a new weighting variable and utilize the doubly robust DID framework.

Matching score DID

Given the logistic regression I constructed in Section II.A, instead of deriving stabilized weights directly, I then match each treated firm to its nearest control by the absolute distance in $\text{logit}(p_i)$ subject to a caliper equal to 0.2 times the standard deviation of $\text{logit}(p_i)$. Let $\mathcal{M}$ denote the set of matched firms and m_i the number of times firm i is used as a match. (For more details, see Caliendo and Kopeinig (2008))

Table 3—: PS-DID Estimation Results by Firm Size Bins

	[0, 25)	[25, 50)	[50, 75)	[75, 100)
Post×Treat		0.438***	0.410***	0.406***
		(0.074)	(0.074)	(0.060)
Firm fixed effects		✓	✓	✓
Time fixed effects		✓	✓	✓
Sector × Time trend		✓	✓	✓
N	0	9,474	9,155	10,550
R^2		0.747	0.661	0.686

Note: Firm size bins are based on pre-treatment average employment; dependent variable is specified as log(productivity). Each column corresponds to a regression in the respective size bin sample weighted by PS stabilized weights, with standard errors clustered at the firm level. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

On the matched panel $\{(i,t) : i \in \mathcal{M}\}$, we run weighted least squares with the same DID specification as before but using matching weights $w_i^{\text{match}} = 1$ for treated firms and $w_i^{\text{match}} = m_i$ for controls.

Love plot is shown below.(See Figure 8) From the plot, both weighting methods give stable standard mean different with advantage in different variables.

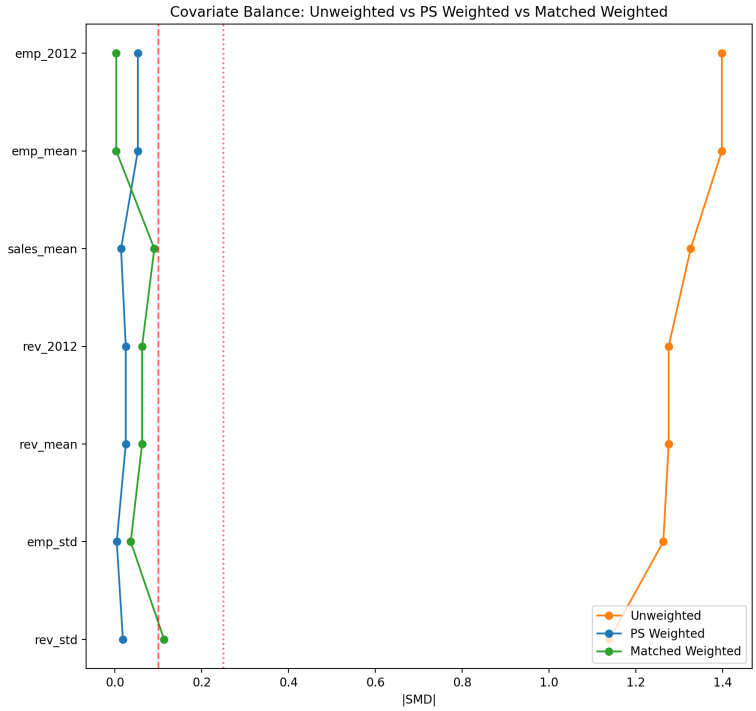
Result is given by Table 4, where the effects are still significant around 40%, but slightly larger than PS-DID estimates.

Table 4—: Matching score DID Estimation Results

	(1)	(2)	(3)	(4)
Post×Treat	0.436***	0.380***	0.435***	0.438***
	(0.083)	(0.023)	(0.048)	(0.047)
Firm fixed effects	✗	✓	✓	✓
Time fixed effects	✗	✗	✓	✓
Sector * Time trend	✗	✗	✗	✓
N	26,897	26,897	26,897	26,897
R^2	0.088	0.832	0.834	0.834

Note: Matching sample is estimated with WLS using matching weights, firm-clustered robust standard errors.
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Figure 8. : Love plot for original uniform weight and two weighting methods



Note: Orange line is the original unweighted observations, with PS and matched weighting in blue and green line respectively.

Doubly robust DID

I also estimate a doubly robust DID at the firm level. This estimators are consistent if either (but not necessarily both) a propensity score or outcome regression working models are correctly specified.Sant'Anna and Zhao (2020)

Specifically, for each firm i, define the pre- and post-period averages $y_i^{\text{pre}} = \mathbb{E}_t[y_{it} \mid \text{post}_t = 0]$ and $y_i^{\text{post}} = \mathbb{E}_t[y_{it} \mid \text{post}_t = 1]$, and the change $dY_i = y_i^{\text{post}} - y_i^{\text{pre}}$. Let $T_i = \text{treat}_i$ and $p_i = \Pr(T_i = 1 \mid X_i)$ be the firm-level propensity score estimated from pre-treatment covariates X_i. With propensity score method in section II.A, weights w_i are gievn.

The average treatment effect(ATT) is then estimated as

$$\widehat{\text{ATT}}_{\text{DR}} = \underbrace{\frac{1}{\#\{i : T_i = 1\}} \sum_{i:T_i=1} r_i}_{\text{mean residual among treated}} - \underbrace{\sum_{i:T_i=0} \tilde{w}_i^{\text{ctrl}} r_i}_{\text{weighted mean residual among controls}} ,$$

where $\tilde{w}_i^{\text{ctrl}} = w_i^{\text{ctrl}} / \sum_{j:T_j=0} w_j^{\text{ctrl}}$.

Under assumptions that $0 < p_i < 1$ and parallel trends assumption on dY_i conditioning on X_i.

Inference is obtained by firm-level bootstrap with 200 resamples in implementation. Since at firm level, firm fixed effects and time fixed effects are not applicable, so we only consider sector dummy in implementation.

Table 5 shows the result of doubly robust DID estimates. From this method, firms adopting the program experience approximately 42% increase in labor productivity level.

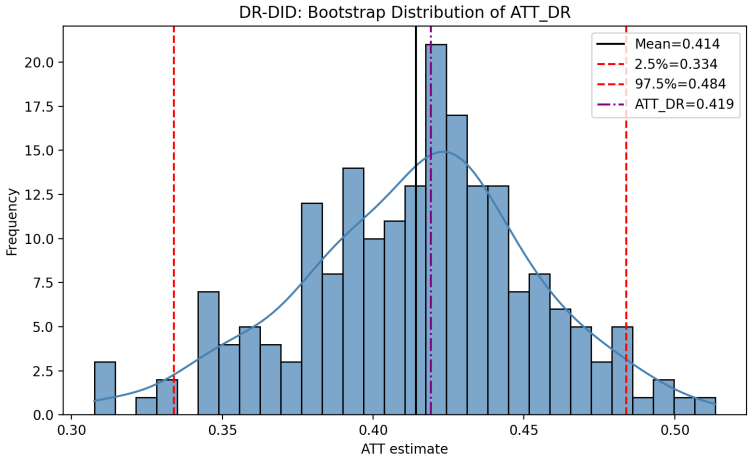
Table 5—: Doubly robust DID Estimation Results

	(1)	(2)
ATT	0.418***	0.419***
	(0.039)	(0.039)
Sector Dummy	✗	✓
N (firms)	351	351

Note: DR-DID is calculated on the firm-level $dY = Y_{\text{post}} - Y_{\text{pre}}$ cross-section; "Firm FE/Time FE" is not applicable. Industry dummies are included in specification 2 to capture industry differences. SE is based on firm-level bootstrap (B=200).
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Also, the bootstrapped result is shown in Figure 9, where we get the standard error of this estimator.

Figure 9. : Bootstrap result for doubly rebost DID estimators



Note: 200 resampling is applied to obtain the estimators. Estimator value, mean, 2.5% quantile and 97.5% quantile are also labeled as vertical lines in the graph.

IV. Conclusion and Discussion

Using monthly microdata and multiple identification strategies, I find that the small business training program increases firm productivity by approximately 0.4 standard deviation on average. Evidence on different sectors and firm size bins is uniformly positive, with greater effect on smaller firms in hospitality sector. Results are robust to alternative control construction (matching) and to doubly robust DID estimators.

Thus, this programs targeted at smaller businesses boosts not only efficiency of small firms, but also bridge the productivity gap between small and big firms, driving equality.

Future studies could focus more on better measurement of productivity by incorporating more information of firms (e.g., expenditure on capital). With more information, a structural estimate might be available to help us give more accurate inference estimates.

Besides, mechanism behind programs like this could be further investigated. Hypotheses include increasing managers' efficiency, signaling effect among others. With data on managers' wage change as a representative of their performance, mechanism could be further investigated.

REFERENCES

Bhaskaran, Krishnan, and Liam Smeeth. 2014. "What is the difference between missing completely at random and missing at random?" *International journal of epidemiology*, 43(4): 1336–1339.

Caliendo, Marco, and Sabine Kopeinig. 2008. "Some practical guidance for the implementation of propensity score matching." *Journal of economic surveys*, 22(1): 31–72.

Haukoos, Jason S, and Roger J Lewis. 2015. "The propensity score." *Jama*, 314(15): 1637–1638.

McKenzie, David, and Christopher Woodruff. 2014. "What are we learning from business training and entrepreneurship evaluations around the developing world?" *The World Bank Research Observer*, 29(1): 48–82.

Sant'Anna, Pedro HC, and Jun Zhao. 2020. "Doubly robust difference-in-differences estimators." *Journal of econometrics*, 219(1): 101–122.

Schulz, Eric, Maarten Speekenbrink, and Andreas Krause. 2018. "A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions." *Journal of mathematical psychology*, 85: 1–16.

Appendix

A1. Results for dropped data

Table A1 reports the dropped sample. Across columns with progressively richer fixed effects and sector-specific trends, estimates are stable around 0.4 log points.

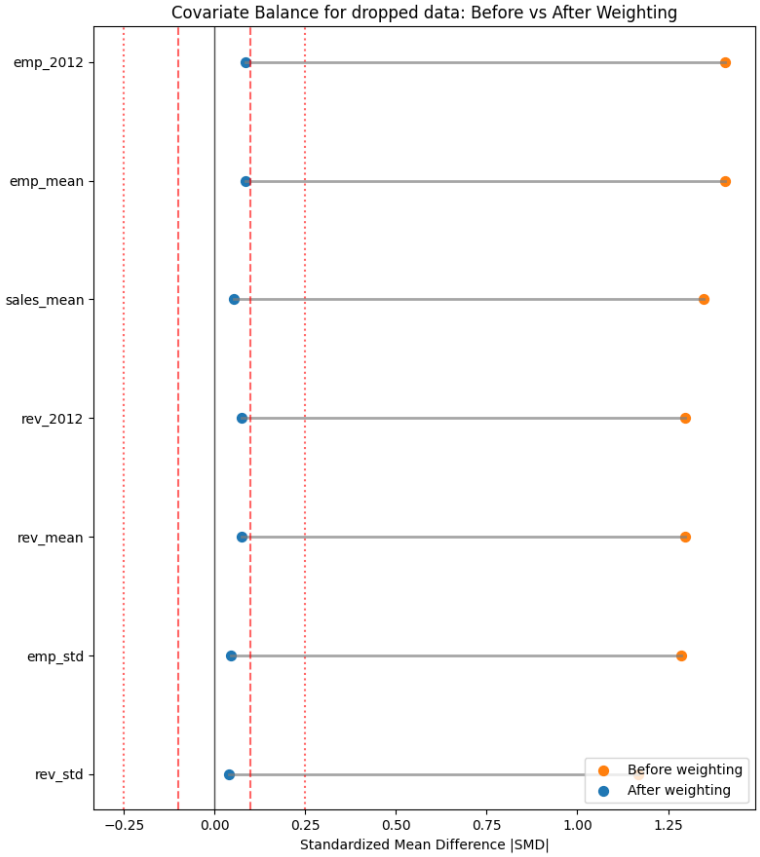
Table A1—: PS-DID Estimation Results for dropped data

	(1)	(2)	(3)	(4)
Post*Treat	0.441***	0.375***	0.396***	0.396***
	(0.088)	(0.026)	(0.037)	(0.036)
Firm fixed effects	✗	✓	✓	✓
Time fixed effects	✗	✗	✓	✓
Sector × Time trend	✗	✗	✗	✓
N	39,270	39,270	39,270	39,270
R^2	0.070	0.851	0.852	0.852

Note: Dependent variable is the logarithm of productivity, and all models use propensity score stabilized weights (WLS), with standard errors clustered at the firm level. Robust standard errors are reported in parentheses.
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

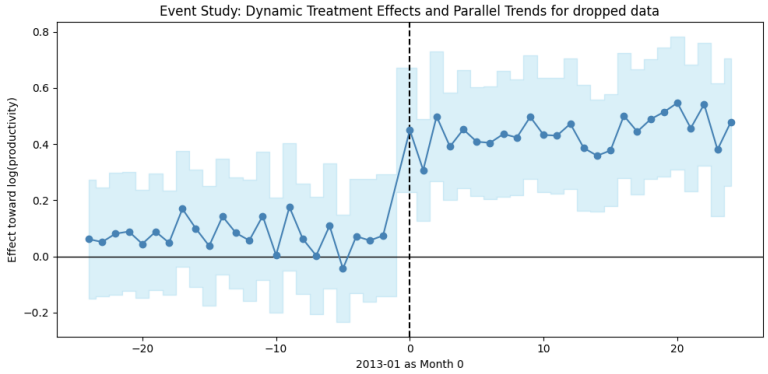
Love plots A1, event analysis A2 and placebo plots A3 are also given for dropped data.

Figure A1. : Love plot for dropped data



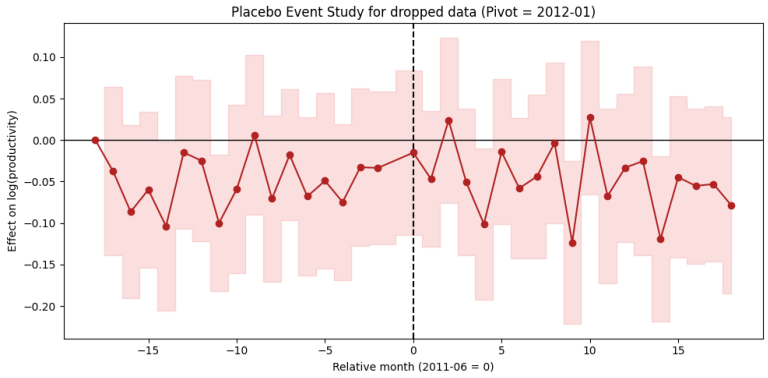
Note: After reweighting, the standardized mean difference between treated and untreated sample drops significantly, mostly below the bar 0.1 shown in the red dashed line.

Figure A2. : Event analysis for causal effect(dropped data)



Note: Before the shock happens the program effect is not significant from zero, while afterwards it shows a significant positive trend. Blue shaded area is the 95

Figure A3. : Event analysis for placebo tests(dropped data)



Note: Assume a fake program on 2011-06 and apply the same analyzing strategy gives all effect insignificant, further demonstrating validity of results. Red shaded area is the 95% confidence interval.