NATIONAL COUNCIL OF APPLIED ECONOMIC RESEARCH

Working Paper No. : WP 129

Model-Selection Inference for Causal Impact of Clusters and Collaboration on MSMEs in India

Samarth Gupta

August 2020

National Council of Applied Economic Research

Model-Selection Inference for Causal Impact of Clusters and Collaboration on MSMEs in India*

NCAER Working Paper

Samarth Gupta[a]

Abstract

Do agglomeration-based spillovers impact firms more than the technical know-how obtained through inter-firm collaboration? Quantifying the effect of these treatments on firm performance can be valuable for policy-makers as well as managers/entrepreneurs. I observe the universe of Indian MSMEs inside an industrial cluster but with no collaboration (Treatment Group 1), those in collaboration with other firms for technical know-how but outside a cluster (Treatment Group 2) and those outside cluster with no collaboration (Control Group). Selection of firms into these treatments and sub-sequent performance of the firm may be simultaneously driven by observable factors. To address selection bias and overcome model mis-specifcation, I use two data-driven, model-selection methods, developed in Belloni et al. (2013) and Chernozhukov et al.(2015), to estimate causal impact of the treatments on GVA of ˝rms. The results suggest that ATE of cluster and collaboration is nearly equal at 30%. I conclude by offering policy implications of the results.

Keywords: Entrepreneurship, Firm Performance, SME

JEL Classification: L25, L26, 017

[a] National Council of Applied Economic Research, New Delhi, I would like to thank Dr. Shashanka Bhide and Dr. Sanjib Pohit for reviewing the paper. I would like to thank Dr. Bornali Bhandari for a fruitful discussion. Please email at sgupta@ncaer.org for any queries

Model-Selection Inference for Causal Impact of Clusters and Collaboration on MSMEs in India

Samarth Gupta*

December 2020

Abstract

Do agglomeration-based spillovers impact firms more than the technical know-how obtained through inter-firm collaboration? Quantifying the effect of these treatments on firm performance can be valuable for policy-makers as well as managers/entrepreneurs. I observe the universe of Indian MSMEs inside an industrial cluster but with no collaboration (Treatment Group 1), those in collaboration with other firms for technical know-how but outside a cluster (Treatment Group 2) and those outside cluster with no collaboration (Control Group). Selection of firms into these treatments and subsequent performance of the firm may be simultaneously driven by observable factors. To address selection bias and overcome model mis-specification, I use two data-driven, model-selection methods, developed in Belloni et al. (2013) and Chernozhukov et al. (2015), to estimate causal impact of the treatments on GVA of firms. The results suggest that ATE of cluster and collaboration is nearly equal at 30%. I conclude by offering policy implications of the results.

JEL Codes: L25, L26, 017,

*NCAER. I would like to thank Dr. Shashank Bhide and Dr. Sanjib Pohit for reviewing the paper. I would like to thank Dr. Bornali Bhandari for a fruitful discussion. Please email at sgupta@ncaer.org for any queries.

1 Introduction

Economists have devoted substantial attention to valuable knowledge transfer across firms through clusters and collaboration. Schmitz and Nadvi (1999) terms cluster's benefits as *passive* where gains *"fall into the producer's lap"*, whereas collaboration requires *"active joint effort"* between consenting parties. However, what creates a bigger impact on firm performance—positive externalities from industrial agglomerations or mutually-beneficial collaboration for technical know-how?

Government agencies and entrepreneurs/managers may find relative estimates of the two treatment valuable. Governments in the developing world are actively pursuing policies to increase firm productivity in their countries. These include promoting collaboration between businesses through reforms for enforcing contracts(WorldBank, 2020) or place-based policies such as Special Economic Zones (Neumark and Simpson, 2014). From the perspective of a manager or an entrepreneur, collaboration with other firms or relocating to a cluster are both valuable but costly decisions. Estimates of the gains of either treatment may aid in the decision making process.

I use MSME Census 2006-07 of India to estimate the individual impact of cluster and collaboration in firm performance. This dataset enumerates more than 1.2 million registered micro, small, and medium sized enterprises in India in 2006[1]. It provides detailed information on firm's outcome and productivity metrics such as gross value added (GVA), value of output and inputs used, labour size, entrepreneur's socio-demographic characteristics such as gender, religion, social group, etc. and enterprises' location, industrial and operational categories.

I classify firms in this dataset into three mutually exclusive groups. Treatment Group 1 is the set of firms which are inside a cluster but with no collaboration. Firms which are outside a cluster but collaborating with other firms are classified under Treatment Group 2. Finally, firms which are neither collaborating nor inside a cluster are classified under Control Group.

Identifying the impact of these treatment for policy prescription is difficult. These treatments are not randomly assigned. More productive firms may be more likely to select into a cluster. Likewise, the need for mutual benefits drives collaboration between firms. Higher productivity of firms in a treatment group, then, reflects the selection bias along with the

[1] A firm with initial investments in plants, machinery and equipment less than INR 2.5 millions is defined as micro, with investment between INR 2.5 and INR 50 millions is small, and with investment exceeding 50 millions is defined as medium enterprise.

effect of a treatment.

Conventionally, there are several methods to address the selection bias in a non-experimental observational study such as Least Square regression or propensity score matching (Rosenbaum and Rubin, 1983, 1985). However, these methods are susceptible to model mis-specification—either a relevant variable is excluded or an irrelevant variable is included in the estimation equation. While the former biases the remaining estimates, the latter creates efficiency losses (Rao, 1971).

To address selection bias while avoiding model mis-specification, I use two recently developed model-selection methods for causal inference in non-experimental data. These methods are post double selection lasso (Belloni et al., 2013) and post-lasso orthogonalized inference (Chernozhukov et al., 2015). Under both methods, LASSO (Tibshirani, 1996) is used to select the variables with highest predictive power for the outcome variable and treatment indicator. Belloni et al. (2013) shows that Average Treatment Effect (ATE) can be obtained from a regression of the outcome on treatment indicator if one uses the union of variables selected from the two LASSO regressions as covariates. On the other hand, under Chernozhukov et al. (2015), dependent variable and treatment indicator are first orthogonalized using the variables selected from the two LASSO regressions. ATE is obtained by regressing orthogonalized dependent variable on the orthogonalized treatment indicator.

I perform the above estimation methods using more than 2500 observable covariates available in the MSME Census 2006-07. These covariates absorb many confounders such as firm-characteristics, location identifiers and socio-demographic features of the entrepreneurs. Under either method mentioned above, ATE on GVA of a firm in an industrial cluster is 0.301 with a standard error of 0.087. Similarly, ATE of collaboration on GVA is nearly 32% using both methods. Thus, the average productivity gains for a firm are marginally higher under collaboration compared to being inside a cluster—benefits from partnership between firms are nearly the same or slightly higher than spillovers a firm receives from other firms around it.

The results are important for several reasons. Most other studies in India which look at benefits of collaboration or clusters restrict attention to few successful cases (described later). This is the first study to estimate impact of these interventions in India on the entire universe of MSMEs. The estimates produced here are, thus, more credible and have higher external validity.

There is a renewed interest in developing countries, and in India particularly, on productivity enhancing industrial policies. For example, in April 2020, government of India announced

construction of multiple industrial clusters as a long-term relief measure against COVID-19 induced recession. However, construction of industrial clusters is costly and may also create negative externalities (de Groot et al., 2009; Busso et al., 2013). On the other hand, improving legal and regulatory standards which promote inter-business partnerships do not carry such costs/risks. In this paper, I show that a firm's productivity gains under collaboration are nearly as high as being inside a cluster. Thus, government agencies can create nearly the same productivity generation for firms at a lower cost if they prioritize improving legal and regulatory environment for inter-firm partnerships.

The paper also contributes toward recent research on productivity dispersion across firms. Syverson (2011) reviews the literature on productivity differences across firms to classify various factors into two categories—firm's internal decision making and its external environment. Gibbons and Henderson (2014) reports the average TFP ratio between 90th and 10th percentile of firms in India is nearly 5:1. Hsieh and Klenow (2009), Tybout (2000) and Hsieh and Olken (2014) have extensively explored the impact of misallocation on size and productivity of firms in India. An important question in this field is whether productivity differences remain persistent or can they dissipate. Further, if it is the latter, then how can one aid the dispersion? In this paper, I record the relative advantages of one internal (collaboration) and one external factor (cluster) in generating performance differences. The results have two implications for this literature. First, direct networking and contractual arrangements between firms for technical knowledge may be as valuable as accidental spillovers it obtains from its environment. Second, productivity gains in a cluster are local to the region whereas collaboration benefits are more mobile, indicating the potential for performance dispersion.

There are several other methods which can be used for estimation in non-experimental observational studies. Most common of these are propensity score matching methods (Dehejia and Wahba, 1998)[2]. However, recent studies have demonstrated the shortcomings of matching methods (Shortreed and Ertefaie, 2017; King and Nielsen, 2019). Specifically, the issues stem from these methods demanding researchers to impose model restrictions through selection of variables. However, Heckman et al. (1998) shows that adding extraneous covariates may increase the variance of all the other estimates. Lately, machine learning methods such as LASSO are increasingly supplanting the conventional econometric techniques. Few recent papers have used machine learning methods to predict performance of firms, instead of estimating causality of treatments on firm performance (Coad and Srhoj, 2019; Bargagli-stoffi

[2]See Heckman et al. (1997, 1998) for a review of application of propensity score for evaluation of labor market policies.

et al., 2020). To the best of my knowledge, this is the first paper to use model selection methods for estimating causal impacts of interventions on firm performance.

The rest of the paper is organized as follows. In section 2, I provide an overview of the literature on industrial clusters and collaboration between firms. In section 3, I provide a brief description of the identification strategy used in the paper. Those who are well versed the propensity score matching methods may skip this section. Section 4 describes the MSME Census in more details along with some summary statistics. Section 5 provides the main results and section 6 concludes.

2 Clusters and Collaboration: Literature Review

Knowledge spillover theory of entrepreneurship identifies human capital, its acquisition and its application as the determinant of entrepreneur's success (Acs et al., 2013; Ghio et al., 2015). There are several modes of knowledge acquisition. In this paper, I focus on two of these modes—spillovers in clusters and technical know-how from collaboration—both of which can augment human capital of the entrepreneur[3]. Firms present inside a cluster experience positive spillovers from surrounding firms. Firms enter into a collaboration to benefit from economies of scale or technology transfer. Thus, both these modes of knowledge transfer can have considerable impact on firm performance. My paper contributes by quantifying the value of these two treatments of knowledge transfer for micro, small and medium sized firms in India.

An extensive literature exists on the origins and types of positive externalities of agglomerations. Combes and Gobillon (2014) provides a framework to identify determinants of various types of agglomeration spillovers. Given the productivity generation of clusters, they have received substantial policy attention as well (Neumark and Simpson, 2014; Brakman and van Marrewijk, 2012; Lucena-Piquero and Vicente, 2019). The MSME Census 2006-07 defines a cluster as a special zone, where at least 100 firms of the same industry are present. As such, spillover due to local specialization, also known as the Marshall-Arrow-Romer spillover, are most likely to be present in this context (Neffke et al., 2009)[4]. These local specialization

[3]There are other forms of knowledge transfer and human capital acquisition. Rasiah and Govindaraju (2009) records value of research and development for firms for industry-university partnerships. Unger et al. (2013) finds that application of knowledge and skills have a stronger relation with enterprise performance than formal education. Lafontaine and Shaw (2014) and Konstadakopulos (2000) find evidence for learning-by-doing in performance of an entrepreneur.

[4]The other two types of agglomeration spillovers are Jacobs' externalities (Jacobs, 1969), which arise due to industrial diversity, and urbanization externalities, which arise when firms are present in or proximate to key infrastructural points such as large cities (de Groot et al., 2009).

may create gains for firms through various channels. For example, Storper and Venables (2004) finds how proximity with one's competitors advances skill transfer through imitation. Duranton and Puga (2004) finds how local specialization attracts highly skilled labour, and reduces matching and search costs for employers.

Industrial clusters have received substantial attention in developing economies as a means to spur industrial growth (Schmitz, 1995; Nadvi, 1996; Schmitz and Nadvi, 1999; de Groot et al., 2009). For country-specific reviews, see Palit and Bhattacharjee (2008) for India and Ge (1999) for China. Leong (2013) compares the macro-economic impact of Special Economic Zones in India and China to show SEZs improve regional growth. For India, in particular, evidence on the micro-economic impact of industrial clusters comes mostly from case studies of successful and selective auto-mobile clusters (Venkataramanaiah and Parashar, 2007; Okada and Siddharthan, 2008). In contrast, my paper conducts a more holistic review of all clusters recorded in MSME Census 2006-07 to find causal impact on firms which choose to join a cluster. Thus, the implications of the results in the paper are far wider.

The typology of collaborations has also been well recorded. An inter-disciplinary survey of the literature suggests the role of inter-firm coordination as overcoming market failures and costly hierarchies (Grandori and Soda, 1995; Caglio and Ditillo, 2009). In MSME Census 2006-07, enterprises are asked if they collaborate with other firms for technical know-how. As such, benefits of inter-firm coordination recorded in this paper reflect the economies of scope which reduce costs of technology acquisition (Teece, 1980, 1986; Nieto Jesus and Santamaria, 2006; Schilling and Phelps, 2007)[5].

The literature on collaboration among Indian firms is sparse. Scott-kemmis and Bell (1985) and Levina and Vaast (2008) have analyzed Indo-British and Indo-American collaborations to show the innovation and performance effect are diminished due to inter-firm boundaries. Biswas et al. (2007) conducts a case study of collaboration among textile Indian SMEs. However, my paper contributes by providing a more extensive study of all Indian MSMEs which collaborate with a domestic partner.

[5]Other benefits of collaboration include quasi-integration between firms (Blois, 1972), hybrid organizational arrangements which reduce governance costs (Williamson, 1981, 1991; Dyer, 1998) and overcoming market failures (Powell, 1987, 1990; Bradach and Eccles, 1989), and, *inter-alia*, complementarity of resources which lead to innovation (Richardson, 2003). See also Evan (1966); Whitley (1991); Granovetter (1983, 1985) for views on inter-firm collaboration from outside the economics literature.

3 Identification Strategy

Consider a firm i with the following production function:

$$Y_i = e^{\gamma T} e^{X_i \beta} e^{\mu}$$

where, Y_i is a performance-measure of firm i, such as productivity, and X are covariates of the firm/entrepreneur which may influence output. T is a binary variable that takes value 1 if the firm received a treatment, and 0 otherwise. μ is a random error which is i.i.d over firms. $\mu \sim N(0, \sigma)$.

Applying log transformation of the production function and denoting $\log Y_i$ by y_i provides the following linear expression:

$$y_i = \gamma T + \beta X_i + \mu_i$$

If firms receive treatment randomly, i.e. $E(T.X_i) = 0$, then we obtain $E(y_i|T = 1) - E(y_i|T = 0) = \gamma$. Under random assignment of treatment, difference between the mean output of firms with and without treatment provides the treatment effect. However, in non-experimental observational setting, selecting into a treatment is seldom random, and may be a function of observable characteristics of the firm/entrepreneur , X_i. For example, an entrepreneur with access to a particular network may be more likely to form collaboration with other firms. Similarly, entrepreneurs in certain geographic zones are more likely to join a cluster. In such cases,

$$T = \alpha X_i + \kappa_i$$

where κ_i is the error term in the selection process.

To account for this, one can ideally include all the relevant covariates in a regression of outcome on treatment indicator. Alternatively, one could explicitly model the selection equation above through a logit specification, and match treated and untreated firms on the propensity of selecting into the treatment. However, under either method, model mis-specification poses a risk—exclusion of some variables can lead to omitted variable bias, whereas inclusion of unnecessary controls can lead higher mean squared error on the treatment effect estimator. While one can rely on prior theoretical literature to select only some variables as controls, such practices are susceptible to charges of p-hacking. Owing to this criticism, I use two data driven method which help select the list of covariates to be included as controls in a linear regression.

3.1 Post-Double Selection LASSO

First, I use the post double selection lasso (PDS LASSO from hereon) technique described in Belloni et al. (2013). Following algorithm details how PDS LASSO obtains the Average Treatment Effect:

1. First, find the set of covariates which best predict the outcome variable, y_i, using LASSO (Tibshirani, 1996). Denote this set as $\mathcal{X}_1$.
2. Second, find the set of covariates which best predict the treatment variable, T, using LASSO. Denote this set as $\mathcal{X}_2$.
3. Regress y_i on T with $X_1 \cup X_2$ as the set of controls. The coefficient on T is the Average Treatment Effect

Belloni et al. (2013) also provides regularity conditions under which the finite sample distribution of the estimator is close to the normal distribution and the estimator of the standard error is also consistent. This allows inference of the estimator.

3.2 Post-LASSO Orthogonalized Regression

Second, I use the post-lasso orthogonalized regression method (Chernozhukov et al., 2015). Following algorithm describes this process.

1. First, run LASSO (Tibshirani, 1996) of the outcome variable on $\mathcal{X}$ to obtain coefficients, $\hat{\beta}$, on covariates.
2. Second, run LASSO of treatment indicator, T, on $\mathcal{X}$ to obtain coefficients, $\hat{\kappa}$, on covariates[6].
3. Obtain orthogonalized values, $\mu_y = y_i - \hat{\beta}\mathcal{X}$ and $\nu = T - \hat{\theta}\mathcal{X}$.
4. Regress μ_y on ν. The coefficient on ν is the Average Treatment Effect.

Asymptotic properties for the estimator are provided in Chernozhukov et al. (2015), which allows inference.

The central assumption in the above two methods is conditional independence of treatment, which states that conditional on observables, treatment is exogenous. Such an assumption remains untested (Angrist and Pischke, 2008). While it is a strong requirement, it should be noted that the large number of observations (> 1.2 Millions) allows using more than 2500

[6]Since LASSO performs feature selection, some values of $\hat{\beta}$ and $\hat{\theta}$ may be zero. See Hastie et al. (2009) for a more detailed discussion.

high-dimensional covariates. In section 4.1, I provide a detailed, evidence-backed description of the variables included for selection used in the two methods. Inclusion of these many controls may help in approaching this assumption.

3.3 Treatment and Control Groups

In this paper, I consider two treatments. First treatment is whether the firm is inside a cluster developed by the government. The second treatment is whether the entrepreneur obtained technical know-how from a domestic collaboration firm. In order to isolate the treatment effects for the two separately, I construct mutually exclusive treatment groups. This is done by considering those firms in a treatment group which selected into only one treatment. Therefore, treatment group 1 consists of firms which were present inside a government cluster but were not in any collaboration for technical know-how. Similarly, treatment group 2 consists of firms which are present outside the cluster but received technical know-how from a collaborating firm. Firms present outside a cluster and which are not collaborating with other domestic firm for technical know-how are part of the control group.

Figure 1 shows a schematic to describe treatment groups 1 and 2, and control group. Total observations in the MSME Census 2006-07 are 1,421,003. Out of this, 23,978 are inside some government cluster. 20,633 enterprises inside government clusters do not have partnership with any other domestic collaborator. These observations constitute Treatment Group 1. Of the firms outside government clusters, 62,416 have obtained technical know-how from collaboration with other domestic firms. These observations constitute Treatment Group 2. The control group is the set of firms outside government clusters and which do not have collaboration. There are 1,243,149 of such enterprises. I identify ATE of treatment 1 and 2 against control group using the above mentioned model-selection methods.

In addition to the above defined treatment and control groups, there is another group of 3,345 firms which received both treatments. Comparing this group of firms with any other treatment or control group may be misleading. The concern arises if propensity of one treatment increases conditional on the other. Methods applied in this paper (Belloni et al., 2013; Chernozhukov et al., 2015) do not account for multiple, overlapping treatments such as collaborating while inside a cluster. Thus, causal effects for firms receiving both treatments are not presented here. I leave that for future work[7].

[7] Lechner (2005) demonstrates a method where causal effects of multiple, overlapping treatments can be identified and estimated. Lopez and Gutman (2017) provides a review of estimation methods for causal effects of multiple treatments. Calia et al. (2016) applied these methods to study the twin effects of innovation and offshoring on the performance of European firms.

While causal effects of multi-overlapping treatments are interesting, ATE estimates provided here of each treatment group against a control group are still valuable. ATE of being inside a cluster may help government agencies in assessing the benefits of an industrial cluster. ATE of a collaboration will also help entrepreneurs and managers in understanding the value of partnership for technical know-how. Finally, from an academic perspective, these estimates can be used to understand the value of spillovers for all parties and the value of shared, private knowledge transfer between two mutually consenting parties.

4 Data and Summary Statistics

I use MSME Census 2006-07, which records detailed information on all formal micro, small and medium sized enterprises in India. In this data set I can observe various entrepreneur and enterprise-specific characteristics.

Panels A, B and C of Table 1 provide summary statistics on socio-demographic of the owner and industrial characteristics for all firms, firms inside government clusters, and firms with technical know-how from a collaborating firm, respectively. Around 14% of all firms are owned by women. This is slightly higher than in treatment groups 1 and 2, where nearly 12% and 11% firms are owned by women. Similarly, 10% of all firms are managed by females, as opposed to 9% for firms in treatment groups 1 and 2. Overall, nearly 65% of all firms are in the manufacturing sector. Share of manufacturing firms is substantially higher in clusters (79%) and those with collaboration (71%).

MSME Census also records output produced and input used by the firm. Table 2 provides summary statistics on these variables in this dataset. The average output of all firms is INR 4.2 Millions and average number of workers employed is 5.66 (Panel A). For firms inside a government cluster (Panel B), the average output are INR 6.8 Millions and average number of workers employed is 11.30, both higher compared to all firms. Average output and average workforce for firms with collaboration are INR 11.6 Millions and 9.8, respectively (Panel C). In figures 2 and 3, I plot the log output and log inputs distributions, respectively, of firms in the control and treatment groups. Log output distributions for firms in either treatment groups stochastically dominates the distribution of log output for control group firms. However, one finds that firms in treatment groups also use more inputs, as can be observed from the distributions of their log inputs.

Table 2 also provides mean value of Gross Value Added (GVA), which MSME 2006-07 defines as the difference between value of total output and inputs. Average GVA for all firms is INR 1.6 Millions. The corresponding figures for firms in treatment groups 1 and 2 are INR 2.5

Millions and INR 4.4 Millions, respectively. Figure 4 plots the density of log of GVA in control group, treatment group 1 and treatment group 2. For the two treatment groups, distribution of log GVA has a thick right tail indicating higher GVA.

Summary statistics suggest considerably large differences in the size and productivity of firms which received a treatment, as opposed to all firms. However, these firms also exhibit significant differences in their observable features. As such, average performance difference between treated and untreated firms may only be partially due to treatment effect, with endogenous selection driving the remaining difference. Further, firms in and outside treatment groups also vary in the amount of inputs used, which may influence total output but not productivity.

In the subsequent analysis, I use log of gross value added as the performance measure of firms. Gross Value Added is the difference between output and inputs, and indicates the amount of output a firm can generate for the same level of input. Interpreted in this manner, GVA is closest to reflecting productivity of firms—for a given value of inputs, they inform how much output a firm can generate[8].

4.1 Covariates

Although the algorithms used in the paper select the variables, the researcher still has to feed the algorithms with the first set of covariates. Thus, we require certain enterprise- or entrepreneur-specific variables which may potentially determine decision to choose a treatment as opposed to being in the control group. Motivated by prior literature, I use the following set of variables as covariates for estimating propensity scores[9].

- Socio-demographics of the entrepreneur: Entrepreneur's characteristics may determine selection into treatment. For example, entrepreneur of certain communities may be more informed about government policies or may have networks to form collaboration. Qian and Acs (2013) demonstrates how absorptive capacity of entrepreneurs may determine the realization of gains from knowledge spillovers. Thus, I use the following socio-demographic variables as covariates.
  - Religion: MSME 2006-07 identifies seven categories of religion of the entrepreneur. These are Hindu, Muslim, Sikh, Christian, Jain, Buddhist, and Others. I use a dummy for each religion as a covariate.

[8]Ideally, one would want to estimate TFP measures as described in Olley and Pakes (1992) and Levinsohn and Petrin (2003). MSME Census 2006-07 does not provide that detailed information.

[9]Note that the reliance on prior literature is only to feed the algorithms with superset of controls. Selection of these covariates for subsequent impact estimation is driven by LASSO.

- Social Category: I can observe four categories of social category depending on caste of the entrepreneur. These are SC, ST, OBC and others.
- Gender of the entrepreneur
- In addition to this, I also include interactions between gender and social category of the owner.

- Firm Characteristics: Firms across and within industries vary substantially in there productivity (Gibbons and Henderson, 2014). Many papers have shown the impact of trade exposure on performance of firms (Bloom et al., 2016). Organization structure and ownership of the firm can also affect productivity (Hortacsu and Syverson, 2009). Following firm characteristics are included in the model selection equations.
  - Dummy variable for 2-digit industry code of the firm as a covariate.
  - Dummy for the type of organization. These are proprietary, partnership, private company, public limited company, cooperative and others.
  - I use original purchase value of plants, machinery and equipment as a covariate for computing propensity score for being in clusters. Cluster developments are usually targeted on the size of the firm. Thus, initial capital of the entrepreneur may determine the choice of entry in a cluster. However, collaboration with other firms may influence initial capital of the firm by direct investment. Thus, instead of being a determinant of treatment group 2, initial capital may be an outcome of being in collaboration. Thus, this variable is not used for estimating propensity to select into collaboration.

- External Characteristics: Local market conditions can have a substantial impact on productivity of firms. Specifically, intra-market competition among firms can force firms to become more productive (Syverson, 2004). Local level regulations and status of input markets can have similar effects (Greenstone et al., 2012; Petrin and Sivadasan, 2013). For India, in particular, Besley and Burgess (2006) have demonstrated how variation in labor market regulations across states in India have created productivity dispersion. Aghion et al. (2008) also records the unequal effect of liberalization on Indian economy. Bhatt et al. (2011) finds strong correlation in establishment of SEZs and areas with higher inflow of foreign investments in India. Therefore, I include the following indicators
  - A dummy for district of the enterprise.

- A dummy if the enterprise is in rural and urban area.
- Interaction dummy nature of the enterprise (manufacturing, services and repair) and district of the enterprise.

Overall, there are 2,520 covariates.

5 Results

As described in section 3, under both methods, LASSO regression is run twice—first on the outcome variable and then on the treatment indicator. This helps select the variables which are either used as controls (in post-double selection lasso) or for orthogonalizing the outcome and treatment indicators (in post-lasso orthogonalized regression).

Covariates selected by LASSO regressions, and used as controls, are available in the log file. For the log of GVA of firms, 52 covariates are selected. Most of these (28) are interactions between indicators for districts and nature of the enterprise. This reflects the regional and industrial productivity variation of firms across India. LASSO selects 6 covariates for treatment indicator of a cluster. Three of these are interactions between district and nature of enterprise indicators, while two are district identifiers and only one pertains to a specific industry. Similarly, for collaboration, LASSO selects only four covariates, while the coefficients on others are shrunk to zero.

5.1 ATE of Clusters

Table 3 provides the estimate of Average Treatment Effect for a firm on being inside a cluster. Column 1 reports the differences of means between firms in treatment group 1 and control group. The difference in average log of GVA is 0.701 with a standard error of 0.007; the average GVA for a firm inside a cluster is nearly 70% higher. However, as discussed above, average difference may be biased due to selection of more productive firms into a cluster. Column 2 regresses log GVA of all firms on the indicator for cluster along with all the controls described in Section 4.1. Now, the coefficient on treatment indicator reduces to 0.18 with a standard error of 0.070. Thus, conditional on all the observable covariates, firms inside a cluster are nearly 18% more productive.

Controls used in column 2 are motivated by prior empirical and theoretical literature. However, many of these could be strongly correlated with each other and, thus, may lead to overfitting. Data driven techniques described in Belloni et al. (2013) and Chernozhukov et al. (2015) help to get around this problem by selecting the variables using LASSO. In

column 3, I report the post-double selection LASSO estimator. The coefficient on treatment indicator now is 0.306 with a standard error of 0.087. Post-LASSO orthogonalized estimator, reported in column 4, is 0.301 with a standard error of 0.087. The Average Treatment Effect on GVA of a firm locating in a cluster is nearly 30%.

5.2 ATE of Collaboration

Table 4 reports the estimates of ATE on GVA of being in a collaboration. Average difference in the mean level of log GVA between firms in treatment group 2 and control group is 0.57 with a standard error of 0.006; firms with collaboration have a 57% higher GVA. Column 2 provides the results of a regression of log GVA on indicator for treatment group 2 along with all the controls. Now, the coefficient on treatment group indicator reduces to 0.14 with a standard error of 0.039. Thus, controlling for the above mentioned controls, a firm with collaboration for technical know-how has nearly 14% higher GVA.

Estimates in column 2 of table 4 stem out of a naive model selection method where all the observable covariates are included in the regression. Columns 3 and 4 report ATEs after employing model selection methods proposed under Belloni et al. (2013) and Chernozhukov et al. (2015), respectively. The ATE in column 3 is 0.324 and is significant at 1% level. Similarly, the ATE in column 4 is 0.319 with a standard error of 0.059. Thus, a data-driven model-selection exercise suggests that a firm observes a 32% increase in its GVA after entering in a collaboration.

6 Conclusion

In this paper, I provide an estimate of the gains for a firm in India if it moves to a cluster or if it enters a collaboration for technical know-how. The results suggest a firm inside an industrial cluster observes an increase in GVA of nearly 30% Similarly, a firm with collaboration observes a growth of nearly 32% in its GVA. These estimates are higher compared to the conventional OLS estimates obtained from a regression controlling for all observable covariates.

While the ATE estimates are nearly equal for both treatments, overall assessment of a treatment would require not just estimating the value but the costs of the treatment. Building a cluster can be a costly enterprise for governments. Clusters may also impose negative externalities on surrounding areas such as congestion or pollution (de Groot et al., 2009; Busso et al., 2013). In India's context, specifically, Sampat (2008) records the displacement

and violations of democratic procedures in establishment of SEZs. Inter-firm collaboration, on the other hand, can have several added advantage other than shared, private knowledge between firms. For eg; Nieto Jesus and Santamaria (2006) shows how technical collaboration can push innovation frontier for smaller firms vis-a-vis larger partners. Thus, the overall benefit of clusters, accounting for negative externalities, may be lower compared to inter-firm collaboration.

How can policy-makers encourage collaboration between firms? Recently, government agencies in developing countries have encouraged collaboration between foreign and domestic firms through offset agreements (Taylor, 2004). In such arrangements, a multi-national firm commits to either transfer knowledge to a local partner or procure some inputs from the domestic market (Johnson, 1999). These agreements are mostly limited to defense equipment procurement. However, fruitful collaboration requires an incentive compatible contract to ensure the needs of both parties. In environments with poor contract enforcement mechanism and slow judicial process, such as India, contracting imposes a high cost of uncertainty. Evidence for this comes from Rao (2019) which finds the detrimental impact of slow judicial process on formal sector firm growth in India. Scott-kemmis and Bell (1985) also demonstrates this in Indo-British collaborations where suppliers were unwilling to provide full technology transfer. Thus, the benefits of collaboration can be potentially higher with effective legal and regulatory safeguards in place, with credible punishment on contract reneging. This will provide an assurance to larger firms to collaborate with smaller partners.

References

Acs, Z. J., Audretsch, D. B., and Lehmann, E. E. (2013). The knowledge spillover theory of entrepreneurship. *Small Business Economics*, 41(4):757–774.

Aghion, P., Burgess, R., Redding, S. J., and Zilibotti, F. (2008). The unequal effects of liberalization: Evidence from dismantling the license raj in India. *American Economic Review*, 98(4):1397–1412.

Angrist, J. D. and Pischke, J. S. (2008). Mostly harmless econometrics: An empiricist's companion. *Mostly Harmless Econometrics: An Empiricist's Companion*, (March).

Bargagli-stoffi, F. J., Niederreiter, J., and Riccaboni, M. (2020). Supervised Learning for the Prediction of Firm Dynamics. In *Data Science for Economics and Finance: Methodologies and Applications*, number September. Springer.

Belloni, A., Chernozhukov, V., and Hansen, C. (2013). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81(2):608–650.

Besley, T. and Burgess, R. (2006). Can Labor Regulation Hinder Economic Performance? Evidence from India. *The Quarterly Journal of Economics*, 119(1):91–134.

Bhatt, A., Puri, M., and Appolloni, A. (2011). SEZ proliferation in India : are the objectives being realized ?

Biswas, S., Roy, S., and Seshagiri, S. (2007). Collaboration in Indian SME Clusters: A Case Study. *The 3rd International Conference on Communities and Technologies*, pages 28–30.

Blois, K. (1972). Vertical Quasi-Integration. *The Journal of Industrial Economics*, 20(3):253–272.

Bloom, N., Draca, M., and Van Reenen, J. (2016). Trade induced technical change? The impact of Chinese imports on innovation, IT and productivity. *The Review of Economic Studies*, 83(1):87–117.

Bradach, J. and Eccles, R. (1989). Markets versus hierarchies: from ideal types to plural forms. *Annual Review of Sociology*, 15(1).

Brakman, S. and van Marrewijk, C. (2012). Reflections on cluster policies.

Busso, M., Gregory, J., and Kline, P. (2013). Assessing the incidence and efficiency of a prominent place based policy. *American Economic Review*, 103(2):897–947.

Caglio, A. and Ditillo, A. (2009). *Controlling Collaboration Between Firms.*

Calia, P., D'Attoma, I., and Pacei, S. (2016). Do Innovation and Offshoring Make a Difference? An Empirical Exploration of the Effects on the Performance of European Firms. *International Business Research*, 9(7):98.

Chernozhukov, V., Hansen, C., and Spindler, M. (2015). Post selection and post regularization inference in linear models with many controls and instruments. *American Economic Review*, 105(5):486–490.

Coad, A. and Srhoj, S. (2019). Catching Gazelles with a Lasso: Big data techniques for the prediction of high-growth firms. *Small Business Economics*, 55:541–565.

Combes, P. P. and Gobillon, L. (2014). The Empirics of Agglomeration Economies.

de Groot, H. L., Poot, J., and Smit, J. M. (2009). New Zealand Agglomeration Externalities , Innovation and Regional Growth : Theoretical Perspectives and Meta-Analysis. *Handbook of Regional Growth and Development Theories*, pages 256–281.

Dehejia, R. and Wahba, S. (1998). Propensity Score Matching Methods For Non-Experimental Causal Studies.

Duranton, G. and Puga, D. (2004). Micro-Foundations of Urban Agglomeration Economies. *Handbook of Urban and Regional Economics*.

Dyer, J. H. (1998). Effective Interfirm Collaboration: How Firms Minimize Transaction Costs and Maximize Transaction Value. *Strategic Management Journal*, 18(May):1998.

Evan, W. M. (1966). The organization-set: Toward a theory of interorganizational relations. *Approaches to organizational design*, pages 173–191.

Ge, W. (1999). *Special economic zones and the economic transition in China*. World Scientific.

Ghio, N., Guerini, M., Lehmann, E. E., and Rossi-Lamastra, C. (2015). The emergence of the knowledge spillover theory of entrepreneurship. *Small Business Economics*, 44(1).

Gibbons, R. and Henderson, R. (2014). What do managers do? *Handbook of Economics*, (November).

Grandori, A. and Soda, G. (1995). Inter-firm Networks: Antecedents, Mechanisms and Forms. *Organization Studies*, 16(2):183–214.

Granovetter, M. (1983). The Strength of Weak Ties: A Network Theory Revisited. *Sociological Theory*, 1(1983):201–233.

Granovetter, M. (1985). Economic Action and Social Structure: the Problem of Embeddedness. *Journal of Economic Sociology*, 3(3):44–58.

Greenstone, M., List, J., and Syverson, C. (2012). THE EFFECTS OF ENVIRONMENTAL REGULATION ON THE COMPETITIVENESS OF U.S. MANUFACTURING.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*, volume 27.

Heckman, J. J., Ichimura, H., and Todd, P. (1998). Matching As An Econometric Evaluation Estimator. *Review of Economic Studies*, 65(2):261–294.

Heckman, J. J., Ichimura, H., and Todd, P. E. (1997). Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. *The Review of Economic Studies*, 64(4):605–654.

Hortacsu, A. and Syverson, C. (2009). Why do firms own production chains? *US Census Bureau Center for Economic Studies*, pages 9–31.

Hsieh, C.-T. and Klenow, P. J. (2009). Misallocation and Manufacturing TFP in China and India. *Quarterly Journal of Economics*, 12(4):1403–1448.

Hsieh, C.-t. and Olken, B. A. (2014). The Missing "Missing Middle". 28(3):89–108.

Jacobs, J. (1969). *The Economy of Cities*. Vintage, New York.

Johnson, J. (1999). Offsets in the international marketplace: an aerospace industry view. *Trends and challenges in aerospace offsets*.

King, G. and Nielsen, R. (2019). Why Propensity Scores Should Not Be Used for Matching. *Political Analysis*, 27(4):435–454.

Konstadakopulos, D. (2000). Learning Behaviour and Co-operation of Small High Technology Firms in the ASEAN Region: Some Evidence from the Singapore-Johor Agglomeration. *Asean Economic Bulletin*, 17(1):48–59.

Lafontaine, F. and Shaw, K. (2014). Serial Entrepreneurship: Learning By Doing?

Lechner, M. (2005). Identification and Estimation of Causal Effects of Multiple Treatments Under the Conditional Independence Assumption. *SSRN Electronic Journal*, pages 43–58.

Leong, C. K. (2013). Special economic zones and growth in China and India: An empirical investigation. *International Economics and Economic Policy*, 10(4):549–567.

Levina, N. and Vaast, E. (2008). Innovating or doing as Told? Status Differences and Overlapping Boundaries in Offshore Collaboration. *MIS Quarterly*, 32(2):307–332.

Levinsohn, J. and Petrin, A. (2003). Estimating production functions using inputs to control for unobservables. *Review of Economic Studies*.

Lopez, M. J. and Gutman, R. (2017). Estimation of causal effects with multiple treatments: A review and new ideas. *Statistical Science*, 32(3):432–454.

Lucena-Piquero, D. and Vicente, J. (2019). The visible hand of cluster policy makers: An analysis of Aerospace Valley (2006-2015) using a place-based network methodology. *Research Policy*, 48(3):830–842.

Nadvi, K. (1996). *Small firm industrial districts in Pakistan.* PhD thesis, University of Sussex.

Neffke, F., Svensson Henning, M., Boschma, R., Lundquist, K.-J., and Olander, L.-O. (2009). The Dynamics of Agglomeration Externalities along the Life Cycle of Industries. *Regional Studies*, 12:1537–1546.

Neumark, D. and Simpson, H. (2014). Place-Based Policies.

Nieto Jesus, M. and Santamaria, L. (2006). TECHNOLOGICAL COLLABORATION: BRIDGING THE INNOVATION GAP BETWEEN SMALL AND LARGE FIRMS.

Okada, A. and Siddharthan, N. S. (2008). Automobile Clusters in India: Evidence from Chennai and the National Capital Region. *The Flowchart Approach to Industrial Cluster Policy*, pages 109–144.

Olley, G. and Pakes, A. (1992). The dynamics productivity in the telecommunications equipment industry.

Palit, A. and Bhattacharjee, S. (2008). *Special economic zones in India: myths and realities.* Anthem Press.

Petrin, A. and Sivadasan, J. (2013). ESTIMATING LOST OUTPUT FROM ALLOCATIVE INEFFICIENCY, WITH AN APPLICATION TO CHILE AND FIRING COSTS. *Review of Economics and Statistics*, 95(1):286–301.

Powell, W. W. (1987). Hybrid Organizational Arrangements: New Form or Transitional Development. *California Management Review*, 30(1):67–87.

Powell, W. W. (1990). Neither Market Nor Hiearchy: Network forms of organization. *Reseach in Organizational Behavior*, 12(January 1990):295–336.

Qian, H. and Acs, Z. J. (2013). An absorptive capacity theory of knowledge spillover entrepreneurship. *Small Business Economics*, 40(2):185–197.

Rao, M. (2019). Judges, Lenders, and the Bottom Line: Court-ing Firm Growth in India.

Rao, P. (1971). Some Notes on Misspecification in Multiple Regressions. *The American Statistician*, 25(5):37–39.

Rasiah, R. and Govindaraju, C. (2009). University-Industry R&D Collaboration in the Automotive, Biotechnology and Electronics Firm in Malaysia. *7th Globelics Conference*, (October):6–8.

Richardson, G. (2003). The Organization of Industry Re-visited.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, (70):41–55.

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score.

Sampat, P. (2008). Special Economic Zones in India. *Economic and Political Weekly*, 43(28):25–27.

Schilling, M. A. and Phelps, C. C. (2007). Interfirm Collaboration Networks: The Impact of Large-Scale Network Structure on Firm Innovation. *Management Science*, 53(7):1113–1126.

Schmitz, H. (1995). Collective efficiency: Growth path for small-scale industry. *The Journal of Development Studies*, 31(4):529–566.

Schmitz, H. and Nadvi, K. (1999). Clustering and industrialization: Introduction. *World Development*, 27(9):1503–1514.

Scott-kemmis, D. and Bell, M. (1985). Technological Dynamism and Technological Content of Collaboration : Are Indian Firms and TechnolAre Indian Firms Missing Opportunities ? *Economic and Political Weekly*, 20(45).

Shortreed, S. M. and Ertefaie, A. (2017). Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*, 73(4):1111–1122.

Storper, M. and Venables, A. J. (2004). Buzz : Face-To-Face Contact and the Urban Economy Michael Storper and Anthony J . Venables. *Journal of Economic Geography*, 4(4):351–370.

Syverson, C. (2004). Market structure and productivity: A concrete example. *Journal of Political Economy*, 112(6):1181–1222.

Syverson, C. (2011). What Determines Productivity? *Journal of Economic Literature*, 49(2):326–365.

Taylor, T. (2004). Using procurement offsets as an economic development strategy. In Brauer, J. and Dunne, P., editors, *Arms Trade and Economic Development: Theory, Policy and Cases in Arms Trade Offsets*, chapter 2, pages 34–47. Routledge.

Teece, D. (1980). Economies of scope and the scope of the enterprise. *Journal of Economic Behavior and Organization*, 1(3).

Teece, D. J. (1986). Profiting from technological innovation: Implications for integration, collaboration, licensing and public policy. *Research Policy*, 15(6):11–46.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Tybout, J. (2000). Firms Manufacturing in Developing Countries : How Well They Do, and Why. *Journal of Economic Literature*, 38(1):11–44.

Unger, J. M., Rauch, A., Frese, M., and Rosenbusch, N. (2013). Human Capital and Entrepreneurial Success: A Meta-Analytical Review. *Journal of Business Venturing*.

Venkataramanaiah, S. and Parashar, S. (2007). Enhancing the competitiveness of SMEs through industrial clusters: The Indian experience. *International Journal of Technology Management and Sustainable Development*, 6(3):164–174.

Whitley, R. D. (1991). The social construction of business systems in East Asia. *Organization Studies*, 12(1):1–28.

Williamson, O. (1981). The Economics of Organization: The Transaction Cost Approach. *The American Journal of Sociology*, 87(3):548–577.

Williamson, O. E. (1991). Comparative Economic Organization : The Analysis of Discrete Structural Alternatives. *Administrative Science Quarterly*, 36(2):269–296.

WorldBank (2020). Doing Business 2020. Technical report, World Bank, Washington DC.

7 Figures

Figure 1: Selection of Treatment Groups 1 and 2, and Control Group

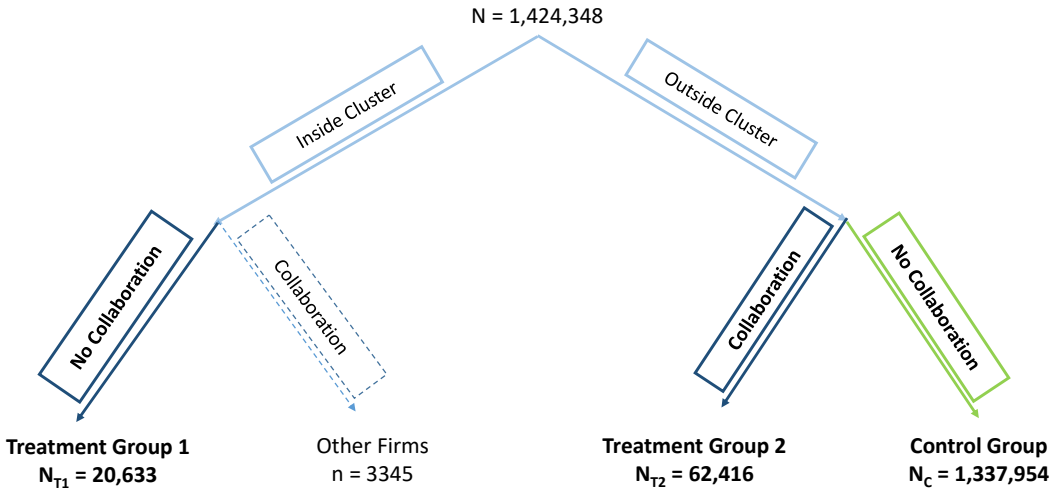

Figure shows how treatment group 1 and 2, and control group firms were selected from the universe of all firms.

Figure 2: Distribution of Log Output in 2006-07

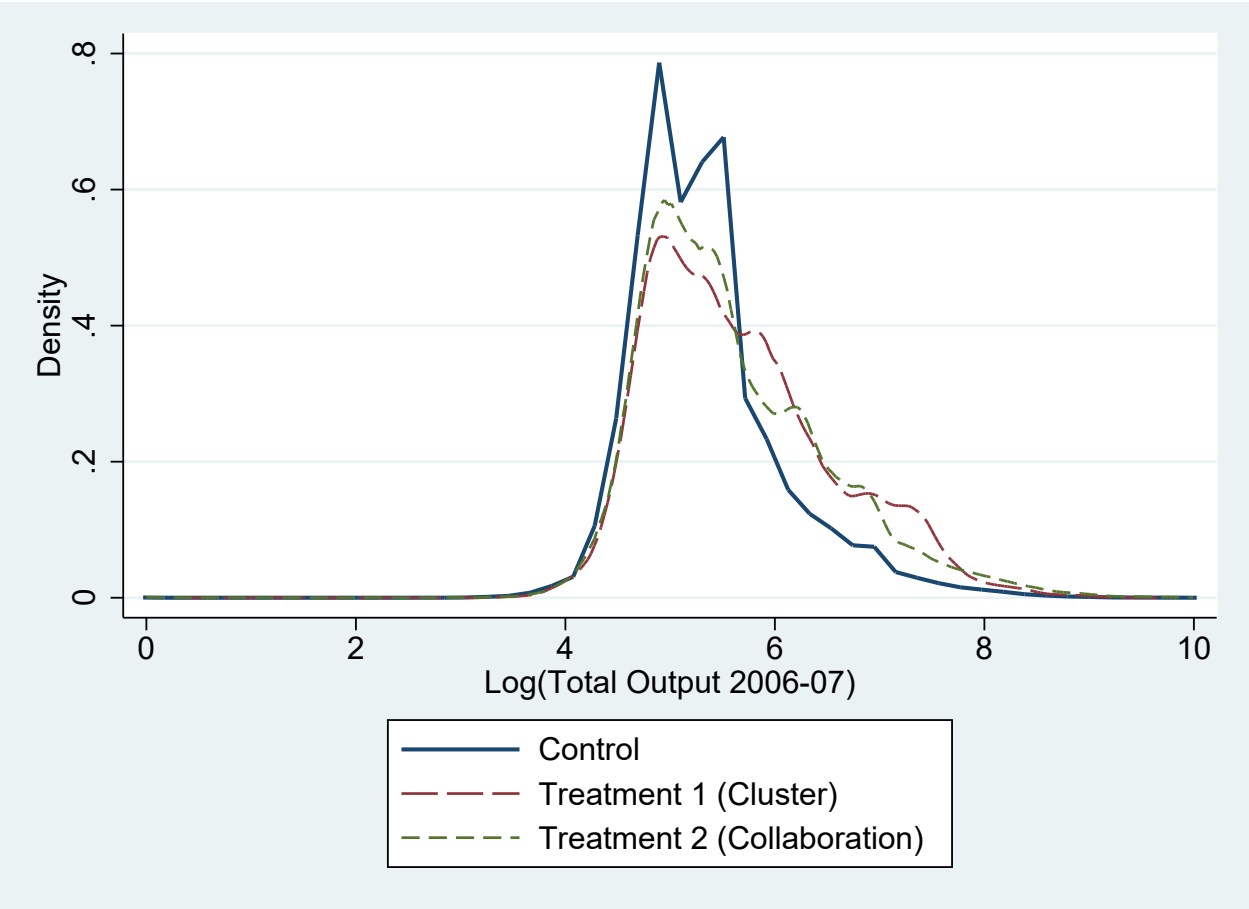

Figure shows the distributions of log output of firms in control, treatment group 1 and treatment group 2.

Figure 3: Distribution of Log Inputs in 2006-07

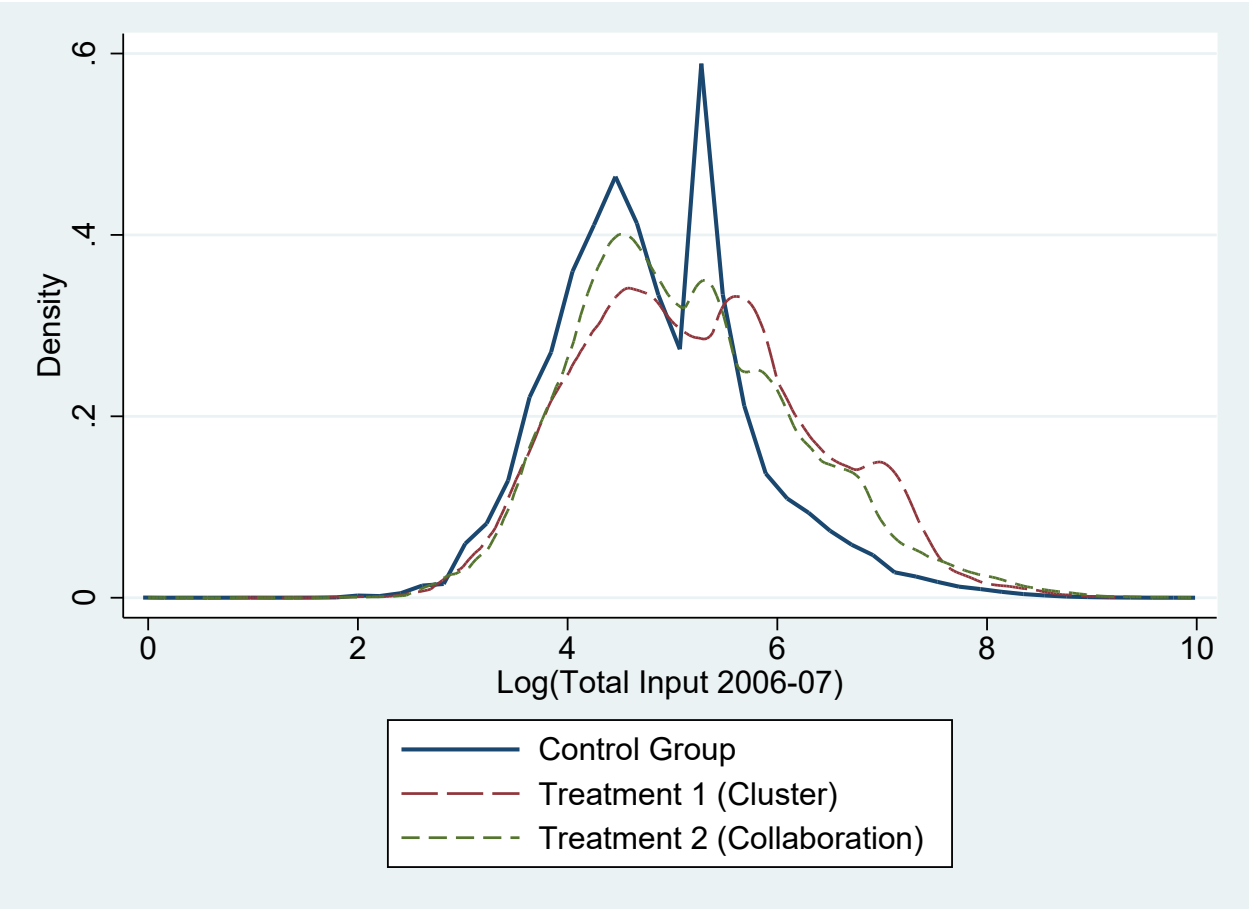

Figure shows the distributions of log inputs of firms in control, treatment group 1 and treatment group 2.

Figure 4: Distribution of Log GVA in 2006-07

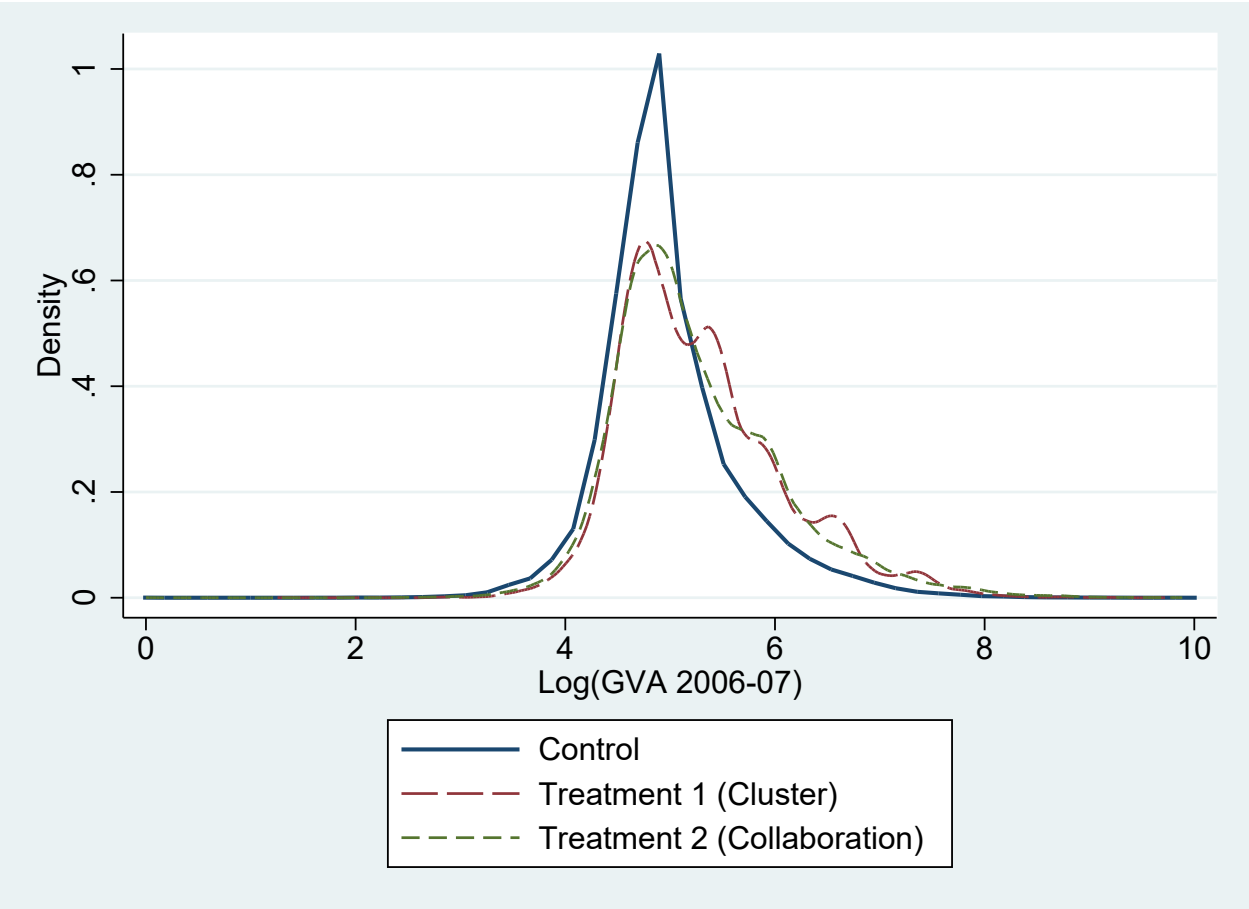

Figure shows the distributions of log GVA of firms in control, treatment group 1 and treatment group 2.

8 Tables

Table 1: Summary Statistics on Covariates of Firms/Entrepreneurs

Panel A: All Firms			
	Mean	Standard Deviation	Observations
Female Owner	0.140	0.346	1420955
Female Manager	0.105	0.306	1421001
Rural	0.468	0.499	1420999
Manufacturing	0.655	0.475	1421001
Panel B: Inside Cluster (Treatment Group 1)			
	Mean	Standard Deviation	Observations
Female Owner	0.126	0.331	20633
Female Manager	0.090	0.286	20633
Rural	0.457	0.498	20633
Manufacturing	0.791	0.407	20633
Panel C: Domestic Collaboration(Treatment Group 2)			
	Mean	Standard Deviation	Observations
Female Owner	0.114	0.318	62394
Female Manager	0.092	0.290	62416
Rural	0.448	0.497	62416
Manufacturing	0.719	0.449	62416

Panel A provides summary statistics on major covariates for all firms/entrepreneurs in the MSME Census. Panel B and Panel C provide summary statistics on firms in Treatment Group 1 and Treatment Group 2, respectively.

Table 2: Summary Statistics on Outcomes for Firms

Panel A: All Firms			
	Mean	Standard Deviation	Observations
Gross Output (Rs.)	4199185.265	74425988.314	1420220
Total Inputs (Rs.)	2621626.759	46370050.446	1418607
Gross Value Added (Rs.)	1605551.532	48922632.260	1418438
Value of Plant and Machinery	566184.690	3728200.409	1421001
Panel B: Inside Cluster (Treatment Group 1)			
	Mean	Standard Deviation	Observations
Gross Output	6866374.276	63282485.496	20602
Total Inputs	4365914.925	31872206.150	20535
Gross Value Added	2557722.797	46668283.276	20551
Value of Plant and Machinery	833228.505	3599051.612	20633
Panel C: Domestic Collaboration(Treatment Group 2)			
	Mean	Standard Deviation	Observations
Gross Output	11602325.337	1.433e+08	62329
Total Inputs	7093805.223	96833354.519	62303
Gross Value Added	4440024.846	79475803.388	62204
Value of Plant and Machinery	1132168.485	6577548.263	62416

Panel A provides summary statistics on total output and input of all firms in the MSME Census. Panel B and Panel C provide summary statistics on firms in Treatment Group 1 and Treatment Group 2, respectively.

Table 3: ATE of Clusters on Log GVA of Firms

	(1)	(2)	(3)	(4)
$\hat{\gamma}$	0.701***	0.183***	0.301***	0.306***
	(0.010)	(0.070)	(0.087)	(0.087)
Estimation	NA	OLS	PDS LASSO	Post-Orthogonalized LASSO

$\hat{\gamma}$ in Column 1 provides the unmatched average difference between log GVA of firms inside and outside cluster. $\hat{\gamma}$ in Column 2 provides regression estimates of the impact of clusters, after including all control variables mentioned in Section 4.1. $\hat{\gamma}$ in Column 3 provides ATE where controls are selected as per Belloni et al. (2013). $\hat{\gamma}$ in Column 4 provides ATE as suggested by Chernozhukov et al. (2015). *,**, *** denote significance at 10%, 5% and 1% level, respectively. Standard errors are clustered at the district level.

Table 4: ATE of Collaboration on Log GVA of Firms

	(1)	(2)	(3)	(4)
$\hat{\gamma}$	0.570***	0.146***	0.319***	0.324***
	(0.006)	(0.039)	(0.059)	(0.059)
Estimation	NA	OLS	PDS LASSO	Post-Orthogonalized LASSO

$\hat{\gamma}$ in Column 1 provides the unmatched average difference between log GVA of firms with and without collaboration. $\hat{\gamma}$ in Column 2 provides regression estimates of the impact of clusters, after including all control variables mentioned in Section 4.1. $\hat{\gamma}$ in Column 3 provides ATE where controls are selected as per Belloni et al. (2013). $\hat{\gamma}$ in Column 4 provides ATE as suggested by Chernozhukov et al. (2015). *,**, *** denote significance at 10%, 5% and 1% level, respectively. Standard errors are clustered at the district level.

NATIONAL COUNCIL OF APPLIED ECONOMIC RESEARCH
NCAER India Centre, 11 Indraprastha Estate, New Delhi 110 002, India
Tel: +91-11-23452698, Fax: +91-11-23370164 info@ncaer.org www.ncaer.org

NCAER | Quality . Relevance . Impact