Special Issue in Honor of Professor Peter Schmidt, 2023, *Empirical Economics*. Volume 64, Issue 6, 3121-3165.

Public Subsidies and Innovation: A Doubly Robust Machine Learning Approach Leveraging Deep Neural Networks

Kerda Varaku* Robin Sickles†

Abstract

Economic growth is crucial to improve standards of living, prosperity and welfare. R&D and knowledge spillovers can offset the diminishing returns to physical capital (machines and labor) and drive long-run growth. Market imperfections can bring R&D below the socially desired level thus, many governments intervene to increase the stock of knowledge, and knowledge spillovers, via subsidies for R&D. We use European firm-level data to explore the effects of public subsidies on firms' R&D input and output. Average treatment effects are estimated controlling for both observable and unobserved heterogeneity. Possible endogeneity in subsidy assignment is addressed and the local instrumental variable (LIV) curve is identified via double machine learning methods. Results indicate that public subsidies increase both R&D intensity and output with more pronounced effects on the R&D intensity of high technology and knowledge intensive firms. The effects of public support remain positive and significant even after accounting for treatment endogeneity.

JEL Classification: H25, C14, C45, C54, C55

Keywords: Double Machine Learning, Public Subsidies, Innovation, Non-parametric Estimation, Deep Neural Networks

1 Introduction

Peter Schmidt has made a number of contributions that speak to the methods and approaches that we undertake in this paper. The effectiveness of government subsidies in expanding the production possibilities of companies and ultimately the countries in which they reside speaks in general to the issue of efficiency and productivity. Peter is one of the iconic contributors to the literature on efficiency and productivity (Aigner et al. (1977)). Nonparametric approaches to addressing unobservables in panel

*Department of Economics, Rice University, Houston, TX. Email: kerdava@gmail.com

†Department of Economics, Rice University, Houston, TX. Email: rsickles@rice.edu

settings such as ours have characterized Peter's work on factor models (Ahn et al. (2013)). Interestingly, in a direct question posed to Peter in a recent Econometric Theory Interview (Sickles (2022)) on his view of the relevance of big data and machine learning techniques in assessing causal effects, which is the contribution of our paper, Peter's perspective is rather measured. He says:

"I think it's important to distinguish Big Data from Big Model. With respect to Big Data, more data can't hurt. With respect to Big Model, remember that you can have big models with little data – you just put in lots of terms..."

He goes on to say:

"...I am not a fan of the sparsity assumption and I don't necessarily understand why it's better to let the data tell us that there are six variables with non-zero coefficients than to try to pick them out ourselves. It seems to me if you want to let the computer make these choices what we need is something akin to the short memory assumption in time series – that even if there was an infinite set of variables, the sum of their coefficients is finite..."

Needless to say, the sparsity assumption is used in our work, the data does in fact tell us what variables have non-zero coefficients, and we do not use "something akin" to the short memory assumption in time series. That said, our contribution is motivated by the many contributions that Peter has made over his storied career and, we trust, he will find our paper worthy of this special issue.

The focus of this study is to evaluate the treatment effects of public subsidies on R&D intensity and R&D output. Innovation plays an important role in sustaining economic growth (Aghion and Howitt (1990), Audretsch (1995)), thus many countries invest a considerable amount of public funds in their attempts to stimulate innovative activity. This paper studies the effectiveness of such investments.

Innovation is the main factor that drives sustainable long-run economic growth. It is responsible for productivity growth such that the same or lower amounts of input can generate a greater amount of output. With higher productivity, workers enjoy higher wages and producers enjoy higher profits, thus allowing them to invest even more and employ even more workers. In addition to productivity growth, innovation leads to new and more efficient ways to deal with the main problems in society, usually by the means of new technology. The new technology benefits society in different ways, for example in fighting and treating illnesses, reducing poverty, and increasing educational accessibility, all of which improve the overall well-being and standards of living.

The relation between innovation and growth has been studied extensively, starting with Neoclassical (Solow (1956)) and Endogenous (Romer (1990)) growth theory. These theories, and the ones that followed, emphasize the importance of technological advancement for long-run growth and have motivated many governments to devote considerable funds to stimulate research and development (hereafter R&D) and innovation. The motivation behind the vast amount of funds invested in innovation lies behind the fact that firms invest less than socially desirable. First, the uncertain nature of

R&D projects and their costs discourages companies. Second, given that knowledge is non-exclusive and non-rival, innovation from one firm has positive externalities and positive spillovers to other firms (see for e.g Nelson (1959); Arrow (1962)). Firms are thus discouraged to innovate since they are not able to appropriate all the returns from innovation. Moreover, many projects could be beneficial for the whole society but unprofitable for the firm, again leading to an R&D below the socially optimal level. To correct such market failures, governments intervene with the intention of reducing firms' costs and increasing incentives for private R&D using tools such as tax credits, deductions, grants, subsidised loans, loan guarantees etc.

While there is a consensus that increasing innovation activity has a positive impact on economic competitiveness and growth (see Aghion and Howitt (1990) and Audretsch (1995)), when it comes to government intervention, there are still disagreements between economists and policymakers about the desirability of subsidizing private R&D activities as the empirical evidence still remains mixed. Direct interventions might distort the entrepreneurs' incentives. All firms have incentives to apply for a subsidy given that the cost of applying is relatively low and thus firms who could afford to conduct R&D by private means could also apply and be granted a subsidy. In this case, firms might simply substitute private investment for public funds. Stiglitz (1988) provides an overview of the literature related to the incentives created by such programs that can distort the government intended benefits. Even if public subsidies do increase innovation activities and intensity, an important question is whether or not this translates to an increase in innovation output.

Such questions call for an empirical evaluation that can determine the effects of the public subsidies on private R&D investment and output. On one hand, previous studies have concluded that subsidized firms invest more in R&D activities compared with other firms in the US (see Audretsch et al. (2002); Lach (2002) and Görg and Strobl (2007)). Other studies such as Heshmati et al. (2005) find that government subsidies have a positive effect only on research expenditures of small firms. On the other hand, Wallsten (2000) finds out that public R&D financing completely crowds out US private firms' R&D inputs using instrumental variables to account for endogeneity. Crowding out is also found in the study of Marino et al. (2016) for French firms or David et al. (2000). Crowding out private R&D inputs implies that without the public support, firms could have had the same or higher amount of R&D input, thus negatively influencing welfare and growth in the long-run. While the heterogeneity of these results could come as a result of particular sample characteristics, it also signals problems such as misspecification bias and/or endogeneity which are common in these non-experimental settings and need to be accounted for.

In this paper, a fully data-driven empirical evaluation is presented in order to answer the question: What is the effect of public subsidies on firms' R&D intensity and R&D output? The motivation behind the empirical method utilized in this paper is that most of the methods used in the literature could be misspecified, thus yielding inaccurate results and incorrect conclusions. Instead of making parametric assumptions when modeling the nuisance parameters (i.e., the outcome (R&D input or output) and treatment (subsidy) assignment mechanism), non-parametric methods are used to learn them from the data. Double robust estimation is used to obtain treatment effect estimators that achieve $\sqrt{n}$ convergence rate even when the nuisance functions these estimators are based on are estimated non-parametrically at slower rates. In addition to misspecification, we also tackle the existence of unobserved characteristics of firms, such as management expertise and competence, that affect innovation and drive the nonrandom subsidy assignment, in a semi-parametric approach using instrumental variables.

2 Literature and Motivation

The literature evaluating the effects of public support on firms' R&D input and output is vast. Most of the empirical papers related to this literature make strong assumptions and/or impose restrictions on the functional form of the model (outcome or treatment models). The most commonly used methods in the literature are parametric methods such as ordinary least squares (OLS) and non-parametric methods such as matching methods (see for e.g. Hamberg (1966); Carmichael (1981); Lach (2002); Lichtenberg (1987, 1988); Wallsten (2000); Toivanen and Niininen (2000)).

Parametric models are a convenient estimation method because the parameters of interest and their asymptotic properties are relatively easy to derive, but of course, if misspecified, they could lead to severe bias. Non-parametric methods such as matching methods are appealing as they do not require any functional form assumptions, but can be problematic with multiple covariates as it becomes increasingly difficult to find matches due to the curse of dimensionality and, at the same time, there exists the risk of matching on irrelevant features, which may distort results. In addition, this method not only ignores potential unobservable effects but it also relies on the existence of common support (Heckman et al. (1996, 1998)).

The curse of dimensionality can be solved by propensity score matching (see Rosenbaum and Rubin (1983)), which summarizes the information from the covariates that influences treatment by the propensity score. Propensity score matching is commonly used to estimate the average effects of government funding on innovation such as in Czarnitzki and Fier (2001); Fier (2002), and Czarnitzki and Fier (2002). These studies have concluded that complete crowding out of private R&D efforts through public subsidies should be rejected. They match on the propensity scores $p(x)$, a function of the covariates x estimated via probit. If this function is misspecified, the sample of matched subjects and thus the distribution of observed baseline covariates will not be similar between treated and untreated subjects conditional on the propensity score. Such matching has also been shown to increases imbalance, inefficiency, model dependence, research discretion, and statistical bias (see King and Nielsen (2019); Austin (2007); Austin et al. (2007)).

Complete crowding out has been found by Wallsten (2000) using data from the Small Business Innovation Research (SBIR) program in the USA. However, the assumed linearity of the outcome equation in their IV approach could lead to a potential misspecification bias. Hussinger (2008) studies the German manufacturing sector using a semi-parametric approach but he again assumes a parametric linear model for the outcome equation. The treatment equation is estimated semi-parametrically, nevertheless, the parametric components are still sensitive to specification errors. While they find positive treatment effects of public subsidies on innovation, the magnitude changes significantly depending on the method with which the selection equation is estimated.

If one instead assumes that either the treatment equation or the outcome equation is correctly specified (but not necessarily both), then double robust methods can be used. These methods combine outcome and treatment modeling and were derived originally to improve the efficiency of the inverse probability weight estimators (see Robins et al. (1995)). Double robust estimators are $\sqrt{n}$ consistent as long as one of the models is correctly specified. Nevertheless, in the more likely case in which both models are misspecified, the performance of double robust estimators can be poor (see Kang et al. (2007)).

Machine learning techniques can mitigate such misspecification biases by learning both the outcome and treatment model from the data, and can also eliminate the need for covariate selection as they can handle high dimensional data. Even with relatively small datasets, machine learning tools may be incorporated as a way to offer better functional approximations rather than assuming particular parametric functional forms (see Breiman et al. (2001)).

Machine learning methods are typically designed for prediction and direct use of these methods for causal inference may generate biased estimators, which also converge slower than $\sqrt{n}$ (Mullainathan and Spiess (2017)). Most of the machine learning estimators rely on regularization which allows for a trade-off between variance and bias. With regularization, one intentionally imposes some bias in order to decrease variance, but this bias converges slower than $\sqrt{n}$. In addition, most machine learning methods have unknown asymptotic properties. The recent literature has developed statistical models and methods that aim to remove such bias and be able to make causal inference using a variety of machine learning methods in a semi-parametric setting (for e.g., Wager and Athey (2018); Chernozhukov et al. (2018)). The empirical methods of this paper rely on the theoretical results of Chernozhukov et al. (2018) and Kennedy et al. (2019). Chernozhukov et al. (2018) prove that treatment effects estimators derived from influence functions and cross fitting, are unbiased and $\sqrt{n}$ consistent, even though they are based on nuisance parameters which converge at a slower rate. The nuisance parameters can be estimated using different machine learning methods such as boosting, random forests, neural networks, deep neural networks, etc. This work allows for identification of many treatment effect parameters such as average treatment effects (ATE), average treatment of the treated (ATT), local average treatment effects (LATE) and from the results of Kennedy et al. (2019), also marginal treatment effects or the local instrumental variable (LIV) curve.

In this paper, we estimate the outcome, treatment and instrument mechanism with a neural network and a deep neural network. The use of neural networks is motivated by the Universal approximation theorem, proved by Cybenko (1989) and generalized by various authors in the 1990s. The theorem states that a neural network can approximate any function to any degree of accuracy, provided that some mild conditions hold. Deep neural networks are shallow neural networks composed of more than one hidden layer and they have been shown to perform at least as well as neural networks, especially when dealing with highly nonlinear functions. Deep learning is one of the methods that has gained in popularity due to its state of the art performance when dealing with high-level abstractions in the data. It is being massively used in such disparate ares as medicine, finance, and quantum chemistry (see for e.g., Gulshan et al. (2016); Heaton et al. (2017); Gilmer et al. (2017); Varaku (2020)) and displays distinct advantages over competing methods (see Krizhevsky et al. (2012); Simonyan and Zisserman (2014); Hinton et al. (2012); Wang and Yeung (2013)). Deep neural networks also have been shown (experimentally) to outperform other methods in terms of causal inference (e.g., Westreich et al. (2010); Johansson et al. (2016); Shalit et al. (2017); Hartford et al. (2017)). We are unaware of previous studies that have applied double/debiased machine methods using deep neural networks. Chernozhukov et al. (2018) give theoretical conditions related to the rate of convergence of the non-parametric method which need to hold, and Farrell et al. (2018) show that such rates are achieved for a class of deep neural networks with a rectified linear unit activation function. To our knowledge ours is the first application of these methods related to public subsidies and innovation and to take a fully data-driven approach to measure the effects of government subsidies on innovation.

3 The Neoclassical and Endogenous Growth Model

This section briefly discusses the theoretical reasons and motivations behind government interventions in R&D as well as gives an overview of the R&D subsidy programs in Europe.

The shortcomings of neoclassical growth theory (Solow (1956)), wherein long-run growth is determined by exogenous factors, are overcome by endogenous growth theory (Arrow (1971); Romer (1990); Aghion and Howitt (1990)). Here the economic growth is determined by endogenous forces, specifically by investment decisions made by profit-maximizing agents. Similar to the tenants of the neoclassical growth model, endogenous growth theory states that technological change lies at the heart of economic growth. However, it is not exogenous. It is determined by factors such as investment in human capital, knowledge and R&D expenditure (for example Romer (1987), Romer (1990) and Lucas (1998)). Investing in R&D involves uncertainty and thereby risks as to whether the R&D will generate innovative output. At the same time, given that innovation is a key element in generating economic growth, the social return to R&D investment is higher than the private one (see

Helpman (2009)). Nevertheless, not being able to appropriate all the returns to innovation (because of spillovers) and because of the risks and uncertainty related to R&D investment, firms underinvest in R&D.

An implication of endogenous growth theory is that policies that promote innovation also promote growth. In 2000, the Lisbon Strategy was drafted which indicated several growth strategies to make the EU "the most competitive and dynamic knowledge-based economy in the world, capable of sustainable economic growth with more and better jobs and greater social cohesion". The main area for priory action was set to be "investing more in knowledge and innovation". The European Council set a target of 3% of GDP to be invested in R&D, where at least 2/3 would come by private means. The same target was passed along in the Europe 2020 strategy, the latest growth strategy in EU, which encourages cooperation between research teams across countries and disciplines. Framework Program Horizon 2020, one of the instruments of Europe 2020, has a fund of more than EUR 80 billion available to R&D and around EUR 100 billion is expected to be available for the next Framework Program.

Grounded on the theoretical justifications above, these policies attempt to stimulate perpetual growth, increase welfare and raise prosperity by increasing the public funding for R&D and promoting cooperation and innovation. A summary of the recent role of R&D subsidies and expenditures in Europe is provided in the supplement A to this manuscript. It is a history that sets the stage for the empirical study of such an extensive program undertaken over the last decades.

4 Data

Our data come from the Community Innovation Survey (CIS) and contains EU science and technology statistics. The CIS is conducted for all European countries, including Norway and Iceland, and is carried out every two years. The harmonized questionnaire offers firm-level data beginning in 1991. Survey response is voluntary and thus the countries participating in different survey years may change.[1]

The CIS collects information on new and significantly improved products or processes of firms in the manufacturing or service industry. Even though not all firms are innovators, the CIS survey requests that all questions be completed by both enterprises with innovative activities and those without such activities. Newly constructed microdata is available after approximately two and a half years after the survey is conducted and each survey covers data related to innovation during the last three years. We use data on three waives of the surveys: CIS2010 (covers the period from 2008-2010), CIS2012 (covers the period from 2010-2012) and CIS2014 (covers the period from 2012-2014). The microdata is a proprietary dataset of Eurostat, available as scientific-use files (SUF), which are partially

[1] Given that it is impossible to construct a balanced panel dataset without losing a large number of firms, in this paper we use a pooled regression approach.

anonymized, and as secure-use files (SC) are only accessible in the Safe Centre at Eurostat's premises in Luxembourg. We use the partially anonymized dataset for 16 European countries[2].

The R&D subsidy variable is divided into 3 main categories: subsidies from local or regional authorities, subsidies from the central government (including central government agencies or ministries), and subsidies from the EU. For this last category firms must identify whether they participated in the Framework Program for Research and Technical Development. Since many firms receive more than one treatment it is not possible to identify the unique treatment effect of each of those sources of subsidies without reducing the sample to very few firms that received only one type of support. Nevertheless, this dataset gives us an advantage compared to many other studies that examine the effects of one particular program, but are unable to control for the effects of other different programs in which the firm may have participated. With our data, we can identify firms that received no form of support from any public scheme or program. Thus, the treatment we are examining is whether or not the firm received any type of subsidy. The subsidies include financial support via tax credits or deductions, grants, subsidized loans, and loan guarantees.

We investigate the effects of government funding on (1) R&D intensity and (2) R&D output. R&D intensity is measured as total R&D spending over turnover, where turnover is based on its value at the end of the survey reference period. R&D spending includes expenditures on intramural R&D, expenditures on extramural R&D, and expenditures incurred in the acquisition of machinery, external knowledge, or other activities related to R&D. With this ratio we form the first outcome variable $Y \in \mathbb{R}^+$. Only the total R&D expenditure is available from the CIS survey (public plus private spending) and thus the portion of private R&D spending is not observable. With this information we can only test the total crowding out hypothesis, i.e., whether the firm would have had the same total R&D intensity even in the absence of the public support. A strict interpretation of full crowding out would suggest the existence of a one-for-one substitution of public for private money. The partial crowding out hypothesis, wherein total R&D intensity increases after public support but private effort falls, cannot be tested with our data.

While many studies in this literature use the number of patents as a proxy for R&D or innovative output, we use an alternative definition of innovative output for the following reasons. First, not all innovation output can be patented. Second, patenting an innovative output could be undesirable. If the benefits from applying/obtaining a patent are less than the costs, a firm would choose not to disclose its innovation, and thus not patent it. As also argued by Griliches (1998), the firm will only choose to apply for a patent if the economic value of the patent right exceeds the cost of patenting. In addition, firms located in different countries can differ in their propensity to patent depending on the effectiveness of institutions to guarantee property rights protection. For these reasons we use an

[2]Data availability as scientific-use files (SUF) and as secure-use files in the Safe Centre (SC) in Luxembourg are reported for each country in Section B of the supplementary material for the last three waives of the survey.

alternative proxy for R&D output based on the Oecd (2005) manual, where innovation is defined as "the implementation of a new or significantly improved product (good or service), or process, a new marketing method, or a new organizational method in business practices, workplace organization or external relations". Given that the subsidies are not usually granted for organizational and marketing innovation, we only use the implementation of new or significantly improved products or processes to account for innovation. As explained in Grifell-Tatjé et al. (2018) it is important to include both product and process innovation since during the life-cycle of technological change it is often the case that an increase in efficiency occurs after a period during which the firm focuses on process innovation followed by a period of product innovation. In addition, we also include attempted innovation, i.e., projects which are either abandoned or which are ongoing. With this definition of innovation, the outcome variable Y is 1 if firms innovate or 0 otherwise. We view this as a more complete measure of innovation and one that is comparable measure both through time and across countries.

The independent variables we considered that can potentially affect both treatment D and outcome Y are not preselected. However, variables that are themselves affected by the treatment D are not included as this would cause endogeneity (see Wooldridge (2010)). The ability to control for a large number of firm characteristics allows us to minimize selection bias, which is more likely to be present using traditional methods with fewer covariates. We control for firm size, employment growth, the education level of employees, the firm's geographical location and the industry in which they compete, among other variables. Firm size is an important feature that affects both innovation and also the probability that the firm will be assigned a subsidy. It has been shown that R&D intensity often rises disproportionately with firm size (see for e.g., Levin et al. (1985); Acs and Audretsch (2003)) since small firms may face higher costs of capital and are more likely to be cash-constrained. In addition, larger firms are more likely to receive public subsidies (see Hussinger (2008); Czarnitzki and Fier (2002); Almus and Czarnitzki (2003)). We use employment in the last calendar year of the reference period as an indicator of firm size. Firms belong to one of the following groups: (1) firms with less than 50 employees, (2) firms with 50 – 249 employees and (3) firms with more than 250 employees. Each of the groups is referred to as small, medium and large respectively.

The dataset provides different NACE codes for the industry the firm belongs to. This is important to control for since it is a proxy for different technological potentials and appropriability conditions (see Klette et al. (2000)) in different industries. In addition, a categorical variable that is equal to 1 if the firm is part of an enterprise group[3] and 0 otherwise is included. Public agencies may prefer to subsidize firms that are part of a group since they have a higher chance of internalizing and benefiting from the spillovers of their parent companies. Categorical variables for the location of the head office are also included. The location is categorized as (1) the same as the country of operation, (2) located in another EU, EFTA (European Free Trade Association) or candidate country and (3) located in the

[3] For example, a multinational company or a holding company.

rest of the world. Firms with a parent group located in national territory can have better network linkages and are more informed about national subsidy schemes. As a result they are more likely to apply for them and thus the propensity to receive a national subsidy can be higher. This is also shown for many European firms empirically by Czarnitzki and Fier (2002); Busom (2000); Hussinger (2008) who conclude that affiliates of foreign firms are less likely to get public subsidies.

As for the missing responses, for categorical variables, we create dummy variables that are inputed into the neural network. Missing responses could contain useful information that should not be disregarded and given our ability to handle many covariates, this method is not only feasible but also preferable in terms of making use of all relevant information in the dataset. For continuous variables, we drop the missing observations.

4.1 Summary statistics

In this dataset 11% of the firms have received some type of government funding during the years 2008 - 2014, and 33% of all firms have generated some innovation output (see definition above). A noticeable difference can be seen between the funded versus non funded firms in terms of innovation output (Table 3). 26% of non funded firms have innovated, while more than 94% of funded firms have innovated. This might suggest that those firms that are funded are more likely to innovate, as the direct comparisons across groups suggest a substantial difference in mean output. Nevertheless, it also could be the case that the government selects exactly those firms that are more likely to innovate to begin with. Thus in the absence of funding, those firms would still be more likely to innovate compared with other firms.

In addition, the average expenditure on R&D as a percentage of turnover is around 8% (Table 2). For firms that are not funded, the average is 2.4% while for those that are funded it is 15%. The main question that this paper tries to answer is whether these observed differences can be attributed to the public subsidy. As mentioned before, such a direct comparison might be misleading because of the non-randomness of government funding and thus this problem requires a causal analysis.

	Mean	Std.	Min.	Max
Funding	0.1129	0.3165	0	1
R&D intensity	0.0834	0.2991	0	4.988
Innovation output	0.3388	0.4733	0	1

Table 1: R&D intensity and output summary statistics for CIS 2010-CIS 2014

Funding	R&D Intensity				
	Count	Mean	Std.	Min.	Max
0	130381	0.0241	0.1521	0	4.9817
1	28944	0.1535	0.4010	0	4.9885

Table 2: R&D intensity summary statistics by funding group for CIS 2010-CIS 2014.

Funding	Innovation output				
	Count	Mean	Std.	Min.	Max
0	241298	0.2612	0.4393	0	1
1	30721	0.9486	0.2206	0	1

Table 3: Innovation summary statistics by funding group for CIS 2010-CIS 2014.

Most of the sample is composed of small firms (less than 50 employees) followed by medium firms ($50-250$ employees) and large firms (more than 250 employees). The small firms constitute of 60% of the sample, medium-sized firms constitute around 30% of the sample and the large ones constitute 10% of the sample with 164797, 78770 and 28452 firms respectively in each group (see Figure 1).

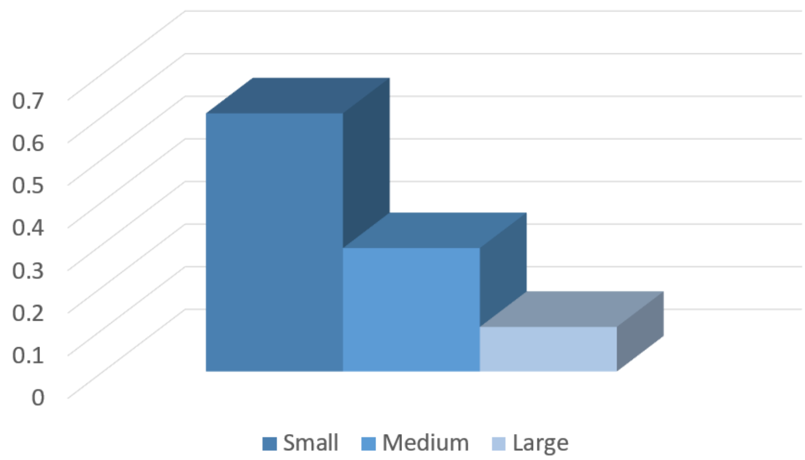

Figure 1: Number of firms by size. Small firms are defined as firms with less than 50 employees, medium firms are defined as those firms with $50-250$ employees, and large firms are defined as those firms with more than 250 employees.

The subsidies are assigned mostly to small and medium firms with 14766 and 10482 subsidies to small and medium firms out of 30721 total subsidized firms (this corresponds to 48% and 34% of total given subsidies). Only 5473 subsidies (18% of total assigned subsidies) are assigned for large firms, but this is natural given the small number of large firms in the sample. Looking at the portion of the firms receiving subsidies by group, it can be see that a higher portion of large firms receives subsidies, followed by medium and then small firms. 19% of the large firms, 13% of medium firms and 9% of small firms in the sample were subsidized (see Figure 2). This could indicate that public agencies favor larger enterprises compared to small ones even though it is the small enterprises that could lack access to credit and a well-established market, thus needing more incentives through public support. Nevertheless, it also could be that instead of focusing on the size, the subsidies are focused on particular areas and technologies that can generate higher benefits and have a larger impact on the economy and the society such as, for example, research and development in defense, health or other technological or knowledge-intensive firms. Moreover, the gap between private and social returns is higher for these particular areas. Indeed, when looking at Figure 3 more than 16% of high-tech and knowledge-intensive firms receive subsidies compared to only 7% of the other firms. This could signal a preference for high-tech and knowledge-intensive industries such as pharmaceutical, manufacturers of computers, aircrafts and electronics, chemicals and drugs, or educational, scientific and health services[4]. It is indeed these firms that spend more in R&D and generate the majority of the patents, thus contributing more to the long-run growth.

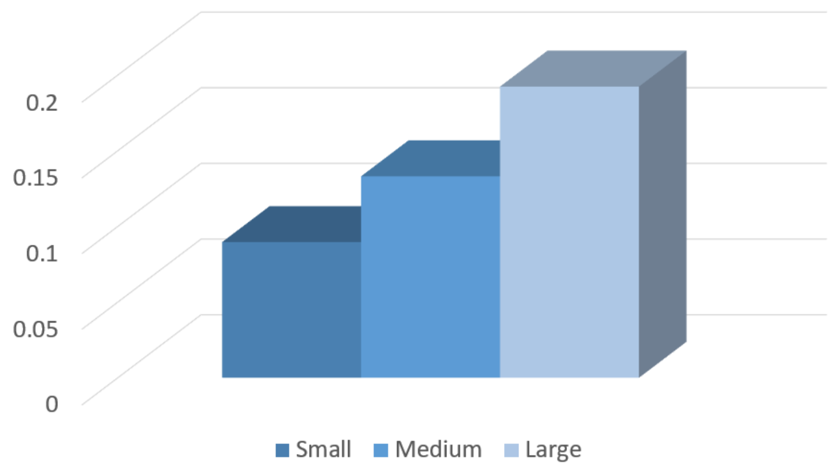

Figure 2: Proportion of firms receiving subsidies by firm's size.

[4]For details about technology-based and knowledge-based classification of industries see Section C in the supplementary material.

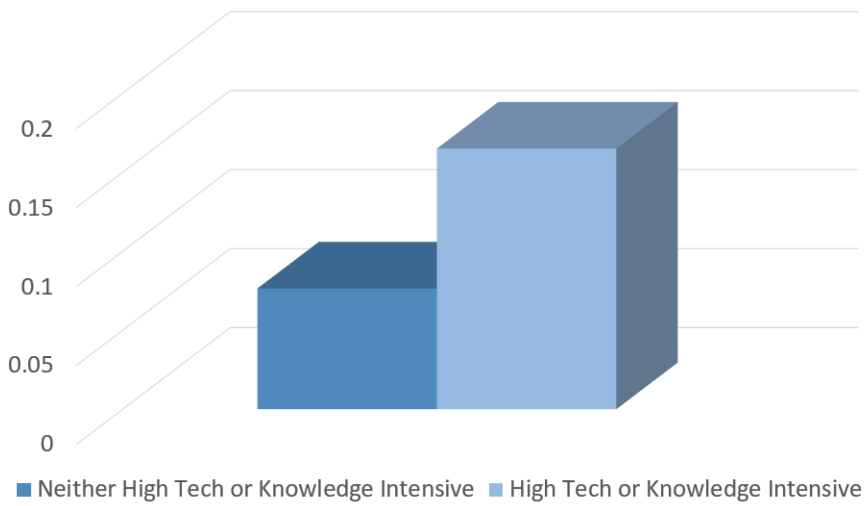

Figure 3: Proportion of firms receiving subsidies firm's type.

5 Empirical Method

This section describes the empirical method for estimating the treatment effects of government funding on R&D intensity and output.

Suppose that we are interested in estimating the ATE, i.e., how the expected value of the outcome Y (R&D intensity or output) would have changed if all the units had taken the treatment ($D = 1$, all firms receive public funding) versus if none of the units takes the treatment ($D = 0$, no firm is funded). In practice, we can only observe each unit either as treated or as a control. Thus the parameter of interest is formulated in terms of a counterfactual, i.e., what would the outcome have been, had the treated (control) unit been in the control (treated) group.

Y denotes the observed outcome while Y_1 and Y_0 denote the counterfactual outcomes, i.e, the outcome had the unit been treated and the outcome had the unit not been treated respectively. The ATE is defined as $E[Y_1 - Y_0]$, the expected effect of being treated for a randomly drawn firm from the population. While ATE is of interest if the policy-makers are considering a full economy shift from a baseline into an alternative state (an intervention that could be universally expanded) or if the subsidy assignment is random, another treatment effect could be of more interest, the ATT. The ATT is defined as $E[Y_1 - Y_0|D = 1]$ and it is the mean effect for firms that received the subsidy. This is more relevant if the policy-makers are questioning or considering the elimination of a program already in place and it shows the average gain from the program for a treated firm randomly drawn from the treated population. Simply comparing the firms who are subsidized versus those who are not is misleading because $E[Y_1 - Y_0] \neq E[Y|D = 1] - E[Y|D = 0]$, unless D is randomly assigned across firms (see Wooldridge (2010); Angrist and Pischke (2008)). The randomness of D is a very strong assumption since either the decision to apply for a subsidy, nor the probability of receiving it, is independent of firm's characteristics. A common way to address this strong assumption is to condition on covariates X that determine Y_1 and Y_0 and to assume that the treatment is random only after controlling for these covariates. In other words, we assume $(Y_1, Y_0) \perp\!\!\!\perp D|X$, which is the well-known conditional independence assumption. In addition, it is assumed that for a set of covariates X, we observe both the control and treated firms and can construct a propensity score that satisfies $0 < P(D = 1|X) < 1$, referred to as the overlap assumption. While it is plausible to believe that the gains from the subsidy $Y_1 - Y_0$ are correlated with the subsidy assignment, it is possible that after conditioning on a wide range of observables X this correlation disappears. Then we can write the ATE as: $\theta^{ATE}(X) = E[Y_1 - Y_0|X]$ and the unconditional effect $\theta^{ATE} = E[g(X,1) - g(X,0)] = \mu_1 - \mu_0$ where the conditional means are defined as follows:

$$E[Y|X,D] = g(X,D),$$

for some unknown function $g(\cdot)$ or

$$Y = g(X,D) + U, \qquad E[U|X,D] = 0.$$

The propensity score equation on the other hand can be written as:

$$D = m(X) + V \qquad E[V|X] = 0.$$

The conditional expectations are unknown, but in general, some parametric functional form is specified and then the parameters are learned (for example by minimizing mean squared errors) to obtain $\hat{\mu}_1$ and $\hat{\mu}_0$. Then

$$\hat{\theta}^{ATE} = \hat{\mu}_1 - \hat{\mu}_0,$$

is consistent only under correct specification of the outcome equation. Alternatively, the inverse propensity score weighting allows the estimation of ATE and ATT assuming the propensity score model $m(X)$ is correctly specified. With this assumption,

$$\hat{\theta}^{ATE} = \frac{1}{n}\sum_i^n \frac{(D - \hat{m}(X_i))Y_i}{\hat{m}(X_i)(1 - \hat{m}(X_i))}$$

and

$$\hat{\theta}^{ATT} = \frac{1}{n}\sum_i^n \frac{(D - \hat{m}(X_i))Y_i}{\hat{p}(1 - \hat{m}(X_i))}$$

with $\hat{p} = \frac{1}{n}\sum_i^n D_i$.

If we assume that either the outcome model or the propensity score model is correctly specified, we can combine these two estimators, and use double robust regression methods. The advantage of the doubly robust regression is that it relies on the consistency of either the outcome equation or the propensity score, but not necessarily both. This method gives some robustness to model specification.

The doubly robust method is based on augmented probability weight estimators. These estimators are shown to be doubly robust, a notion first introduced by Scharfstein et al. (1999). Double robust estimators were also studied by Lipsitz et al. (1999); Robins (2000); Lunceford and Davidian (2004); Neugebauer and van der Laan (2005); Bang and Robins (2005); Robins and Rotnitzky (2001) and Van der Laan et al. (2003). Assuming an asymptotically linear estimator, its difference with the true parameter, $(\hat{\theta}-\theta_0)$, can be expressed as $\frac{1}{n}\sum_{i=1}^{n}\phi(W_i)+\omega$, where W_i is the data for observation i, the first term is the empirical average of the influence function $\phi(\cdot)$ and the last term is an error term converging to 0 with a rate $\sqrt{n}$, i.e., $o_P(1/\sqrt{n})$. The influence functions for a specific model lie in the orthogonal complement of the nuisance tangent space and the efficient influence function is the influence function with smallest variance. With the availability of an efficient influence function, one can construct double robust estimators and these estimators are also semi-parametric efficient. In the next sections we discuss such influence functions for estimating ATE, ATT and MTE.

5.1 Double robustness under exogeneity - ATE and ATT of public subsidies on R&D intensity and R&D output

In this section the public subsidies' effects on R&D intensity and R&D output are evaluated using a double robust machine learning approach under the assumption of unconfoundedness, i.e., the assumption that after controlling for firm's characteristics, the subsidy assignment can be treated as random. Our basic framework from the last section is:

$$Y = g(X,D) + U, \quad E[U|X,D] = 0,$$

$$D = m(X) + V, \quad E[V|X] = 0.$$

Suppose that we are interested in the average treatment effect $\theta^{ATE} = E[g(X,1) - g(X,0)]$. Double robust methods provide consistent estimators if either the outcome or the treatment equation is correctly specified. Using the influence function for ATE (as in Chernozhukov et al. (2018)), the score that allows the identification of θ^{ATE} is:

$$\psi^{ATE}(W;\theta,\eta) = g(X,1) - g(X,0) + \frac{D(Y-g(X,1))}{m(X)} - \frac{(1-D)(Y-g(X,0))}{(1-m(X))} - \theta \tag{1}$$

where $\eta = (g(\cdot), m(\cdot))$ are the nuisance functions and the parameter θ^{ATE} is obtained by solving $E[\psi^{ATE}(W;\theta,\eta)] = 0$. Let $\mathbb{P}_n$ denote the empirical measure so that empirical averages can be written as $\mathbb{P}_n\{W_i\} = n^{-1}\sum_i^n f(W_i)$. Then, the estimator $\hat{\theta}^{ATE}$ can be found as the solution to $\mathbb{P}_n\{\psi^{ATE}(W;\theta,\hat{\eta})\} = 0$, where the nuisance functions $\hat{\eta}$ are estimated non-parametrically in the first step. In a similar manner one can obtain the $\theta^{ATT} = E[g(X,1) - g(X,0)|D = 1]$, with a score written as:

$$\psi^{ATT}(W;\theta,\eta) = \frac{D(Y - g(X,0))}{p} - \frac{m(X)(1-D)(Y-g(X,0))}{p(1-m(X))} - \frac{D\theta}{p} \tag{2}$$

where $p = P_n(D)$ (see for e.g. Van der Laan et al. (2003); Chernozhukov et al. (2018)).

Nevertheless the identification of these two effects relies on the exogeneity assumption. Presence of endogeneity violates one of the main assumptions for the identification of ATE and ATT and the parameters obtain from the scores above would not converge to the true coefficient values, thus they would be inconsistent for ATE and ATT. These effects would then refer to the average predictive effect (APE) and average predictive effect for the exposed ($APEX$) respectively. Without exogeneity, ATE and ATT can not be identified, nevertheless other quantities of interest are identifiable, as discussed in the next section.

5.2 Double robustness under endogeneity - LATE and LIV curve

In this section, violations of the unconfoudedness assumptions are considered. It is possible that governments select firms that are more likely to generate innovation output to begin with, based on unobserved characteristics that can not be controlled for. This is also called "selection on unobservables", i.e., when unobserved characteristics of firms that affect innovation drive the nonrandom subsidy assignment. It is possible that public agencies select firms with a more competent and experienced management. Management characteristics are important determinants of the probability to generate R&D output (Bloom et al. (2009); Nallari and Bayraktar (2010); Ali et al. (2017)), but this information is not observed in our dataset and thus the conditional treatments are no longer randomly assigned. With endogeneity present, the ATE and ATT can not be identified without further restrictions. To address this issue an instrumental variable (hereafter IV) method is proposed and a double robust machine learning approach is used to estimate the parameters of interest.

If there exist IVs satisfying the assumptions of Angrist and Imbens (1995) and Frölich (2007), then it is possible to identify average effects for subpopulations that are induced by the instrument to change the value of the endogenous regressor D. This subpopulation is referred to as the compliers, and the average treatment effect that can be identified is known as the local average treatment effect ($LATE$) (if the IVs are discrete) or marginal treatment effect (MTE) (when the IVs are continuous) (see Heckman and Vytlacil (2005)). The instruments Z must satisfy the following assumptions and regularity conditions:

- $D = D^Z$ and $Y = Y^D$ with probability 1 - This means that the potential treatments and potential outcomes are uniquely determined by the instrument and by the treatment. There is no interference by other units and no difference depending on how the instrument or treatment is administered.

- $(z, x) \in \sup(Z, X)$ if $x \in \sup(X)$ - The instrument should not be deterministic and each firm has a positive probability to get each level of Z, regardless of its covariates.

- $(Y^Z, D^Z) \perp\!\!\!\perp Z|X$ - After controlling for X, Z is as good as random and is unrelated to the potential outcomes and treatment.

- $Y^{zd} = Y^d$ - This assumption implies that Z should only affect the outcome Y through the treatment and not directly.

- Z is monotone with respect to the treatment - For a binary treatment and a binary IV, Z, this means that $D^1 \geq D^0$ with probability 1 for all units. Units that are characterized by $D^1 = D^0 = 0$ are called never takers. In contrast, units for which $D^1 = D^0 = 1$ are called always takers. Lastly, the compliers are units for which $D^0 = 0$ and $D^1 = 1$. For a continuous IV, this can be extended to the assumption that if $Z\prime > Z$, then $D^{Z\prime} \geq D^Z$ with probability 1 for all units. This means that increasing the instrumental variable, either encourages treatment or it does not affect it at all for all firms. This also divides the population into never takers, always takers and compliers, where the compliers at z are now defined as units for which $D^z = 1$ and $D^{z-\epsilon} = 0$ for any $\epsilon > 0$. This can also be written in terms of an unobserved latent threshold $T : D^Z = 1\{Z \geq T\}$ (see Vytlacil (2002)). Units with higher thresholds T, are less likely to receive the treatment compared to units with lower values of T.

- There are at least some units that would take the treatment for a level of $Z = 1$ (for the $LATE$) and $Z \geq T$ (for MTE).

With the availability of such IVs we can identify $LATE$ in case of a discrete IV, i.e., the average treatment effect of those firms that would have $D = 1$ if Z is 1 and $D = 0$ if Z is 0 , or average MTE in case of a continuous IV, i.e., the average effects of treatment on firms for which $D = 1$ when the instrument Z passes some threshold, otherwise $D = 0$.
For a binary instrumental variable let

$$Y = \mu(X, Z) + U, \quad E[U|X, Z] = 0,$$

$$D = m(X, Z) + V, \quad E[V|X, Z] = 0,$$

$$Z = p(X) + \zeta, \quad E[\zeta|X] = 0.$$

The parameter of interest can be written as:

$$\theta^{LATE} = \frac{E[\mu(X,1)] - E[\mu(X,0)]}{E[m(X,1)] - E[m(X,0)]}.$$

Note that this definition of θ^{LATE} departs from what is typically used in the literature thus, the typical causal interpretation of the estimand (i.e., weighted average of heterogeneous treatment effect) would not immediately apply here. The double robust score for such estimator is given by:

$$\begin{aligned}\psi^{LATE}(W;\theta,\eta) = &\mu(X,1) - \mu(X,0) + \frac{Z(Y-\mu(X,1))}{p(X)} - \frac{(1-Z)(Y-\mu(X,0))}{1-p(X)} \\ &- \left(m(X,1) - m(X,0) + \frac{Z(D-m(X,1))}{p(X)} - \frac{(1-Z)(D-m(X,0))}{1-p(X)}\right) \times \theta. \end{aligned} \tag{3}$$

Using this score, $\hat{\theta}^{LATE}$ can be estimated by solving $\mathbb{P}_n[\psi^{LATE}(W;\theta,\hat{\eta})] = 0$, where the nuisance functions $\hat{\eta}$ is estimated non-parametrically in the first step (see Chernozhukov et al. (2018)).
For a continuous instrumental variable, it is possible to identify the local instrumental variable effect curve (LIV) or the effects of a treatment on the a firm that is at the margin of entering treatment at a particular level of the IV, also known as marginal treatment effects MTE. This is the average treatment effects of the firms for which $D = 1$ when the instrument Z passes some threshold T, otherwise when the threshold is below T, $D = 0$. Formally we can define the LIV curve as:

$$\theta^{LIV}(t,v) = E[Y_1 - Y_0|T = t, V = v].$$

In other words, this is the effect among those units with some arbitrary baseline covariate subset $V = v$ that are taking the treatment when the instrument passes the threshold $T = t$, but do not take the treatment for lower values of T. In addition to the assumptions above, we need to further assume that T is continuously distributed and that $\theta^{LIV}(t,v)$ is continuous in t. Then, the LIV curve is identified for any $t \in \mathscr{T}$ by:

$$\theta^{LIV}(t,v) = \frac{\frac{\partial}{\partial z}E[E(Y|X,Z=z)|V=v]}{\frac{\partial}{\partial z}E[E(D|X,Z=Z]|V=v)}|_{z=t}. \tag{4}$$

If the LIV curve can be modeled parametrically as $\theta^{LIV}(t,v;\gamma)$ for parameters $\gamma \in \mathbb{R}^q$, then the weighted least squares projection can be written as:

$$\gamma_0 = \arg\min_{\gamma\in\mathbb{R}^q} E\left[\omega(T,V)(\theta^{LIV}(T,V) - \theta^{LIV}(T,V;\gamma)\right)^2], \tag{5}$$

where the weight $\omega(t,v)$ is a user specified function, usually set based on the instrument density (for example $\omega(t,v) = p(Z = T, v)$ or $\omega(t,v) = \omega(t) = p(Z = t)$). While one can estimate the $\theta^{LIV}(t,v)$ non-parametrically using equation (5) based on equation (4), robust estimation, which also models the instrument mechanism, is preferred for the following reasons: (1) consistency is achieved whenever the instrument equation or the treatment/outcome equation is correctly specified but not necessarily all three, (2) convergence rates are fast enough, even though the nuisance parameters on which they are based and estimated non-parametrically (e.g., with machine learning methods), convergence slowly. The double robust estimation is based on the efficient influence function for the LIV curve, derived in Kennedy et al. (2019):

$$\varphi(W;\gamma,\eta) = \int_{\mathcal{T}} \left\{ g_1(t,V;\gamma)E(D|X,Z=t) - g_2(t,V;\gamma)E(Y|X,Z=t) \right\} dt$$
$$+ g_1(Z,V;\gamma)\left\{\frac{D - E(D|X,Z)}{E(Z|X)}\right\} - g_2(Z,V;\gamma)\left\{\frac{Y - E(Y|X,Z)}{E(Z|X)}\right\}, \tag{6}$$

where

$$g_1(z,v;\gamma) = \frac{\partial}{\partial t}\left\{\frac{\partial}{\partial \gamma^*}\theta^{LIV}(t,v;\gamma^*)|_{\gamma^*=\gamma}\omega(t,v)\theta^{LIV}(t,v;\gamma)\right\}|_{t=z} \tag{7}$$
$$g_2(z,v;\gamma) = \frac{\partial}{\partial t}\left\{\frac{\partial}{\partial \gamma^*}\theta^{LIV}(t,v;\gamma^*)|_{\gamma^*=\gamma}\omega(t,v)\right\}|_{t=z}. \tag{8}$$

Closed form solutions are available when assuming that the LIV curve is projected into a constant, i.e., $\theta^{LIV} = \gamma$, or when it is projected into a linear space, i.e., $\theta^{LIV} = h(t,v)\gamma$ for a known mapping $h : \mathcal{T} \times supp(v) = \mathbb{R}^q$ (see Kennedy et al. (2019)). Nevertheless, when estimation of the LIV curve is carried out non-parametrically, closed form solutions are not available. We must rely on optimization methods and the procedure described in section 5.3 to obtain the estimates of the LIV curve. This is advantageous because the shape of the LIV curve is not restricted as it is for usual parametric models[5] and it also allows for nonmonotone shapes.

We utilize a continuous IV satisfying the previous assumptions, the logarithm of total general government revenue. This variable is directly correlated to the public subsidy since, in general, the latter constitutes a fixed portion of total government revenues. When revenues are high it is more likely that this portion will expand, and when revenues are low it is more likely that this portion will experience a budget cut. Thus, the variable is expected to be strongly correlated to the subsidies. In addition, it is not expected to be related to the unobserved characteristics that affect the R&D output, such as the management characteristics mentioned earlier and thus is expected to be a valid IV. Data on total government revenue is taken from Eurostat.

[5]In fully parametric models it is assumed that the errors follow a normal distribution and the LIV curve shape is determined by simply the inverse of the standard normal distribution multiplied by a constant.

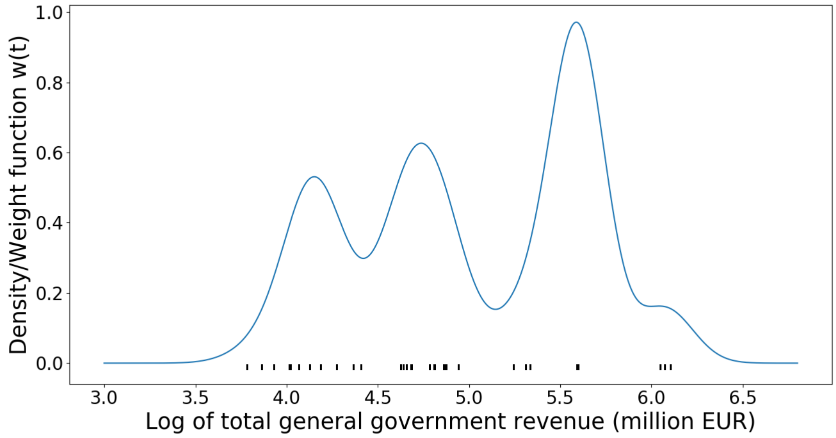

Figure 4: Gaussian kernel estimate of the marginal density $p(z = t)$ of the instrument Z . This is also the weight function $\omega(t)$ used in equation 6.

For the marginal density of Z, which is also used as the weight function $w(t)$, we utilize a kernel density estimator based on a Gaussian kernel. While it is possible to use a uniform weight that assigns mass equally across the support, in practice this could lead to poor efficiency. Kennedy et al. (2019) find that weights based on the instrument density work well, thus we use the estimated marginal density of the instrument as weight. Bandwidth selection is quite important in finding a suitable density. A very narrow bandwidth will result in over-fitting (high variance) and a very wide one will result in under-fitting (high bias). Thus, the optimal bandwidth is found using 3-fold cross validation[6]. The plot of the density is in Figure 4. The validity of the instrument is tested with an F-statistics which should exceed 10 as a rule of thumb.

5.3 Double Machine Learning

The estimators discussed in Section 5.1 are consistent as long as the outcome or the treatment mechanism is correctly specified. The estimator in Section 5.2 is consistent whenever either the treatment/outcome mechanism or the IV mechanism is correctly specified. The famous quote from George E. P. Box concerning misspecified but useful econometric models notwithstanding (see Kang et al. (2007)) a modeling approach that relies on flexible non-parametric methods such as data-driven machine learning methods to estimate these functions rather than making unrealistic functional form assumptions has much appeal. The procedure we use is referred to as “Double

[6]The 3-fold cross validation is used as an alternative to the leave one out cross-validation since the latter is more computationally expensive.

Machine Learning" (see Chernozhukov et al. (2018)) and the estimators based on such procedure can attain faster rates of convergence than the nuisance estimators they depend on, different from the plug-in estimators that have slow convergence rates. Such procedures also offer uniformly valid confidence intervals. Moreover, if the nuisance functions are learned with any machine learning method, under mild conditions and as long as these methods converge each at a rate of at least $n^{-1/4}$, the double robust scores in equation (1), (2), (3), or the estimator from (4) give $\sqrt{n}$ consistent estimators.

Sample splitting is another crucial factor in obtaining desirable asymptotic properties of the treatment parameter of interest. Chernozhukov et al. (2018) suggest that the estimation of the parameter of interest should be obtained using a separate partition of the data that is not used in the learning of the nuisance parameters. The procedure follows these 5 steps:

1. Construct the equal length K - fold random partition ($I_1, \ldots, I_K$) of the sample $W = (X, Z, Y)$. Define $I_k^c = W \setminus I_k$, where $I_j \cap I_k = \varnothing$ for $j \neq k$ and $\cup_k I_k = W$.

2. For each partition $I_k, k \in \{1, \ldots, K\}$, use the complement I_k^c to learn the nuisance parameter η. For L candidate models, train $\hat{\eta}_{k,l}$ for $l \in L$ with I_k^c and pick the one with the smallest mean squared error loss $MSE(\cdot)$ on the I_k: $\hat{l} = \arg\min_{l \in L} MSE(I_k)$. Then use $\hat{\eta}_k = \hat{\eta}_{k,\hat{l}}$ for the second step.

3. For each partition $I_k, k \in \{1, \ldots, K\}$, construct the estimator $\hat{\theta}_k$ by solving
$$|I_k|^{-1} \sum_{w \in I_k} \psi(w; \theta_k, \hat{\eta}_k) = 0,$$
where $|I_k|$ is the cardinality of set I_k and ψ is any score in (1), (2), (3).

4. Aggregate the estimators to get the estimated ATE, ATT and $LATE$ for each score in Equation (1), (2), (3) respectively:
$$\hat{\theta} = \frac{1}{K} \sum_{k=1}^{K} \hat{\theta}_k.$$

5. For the estimation of the LIV curve, to select from the candidate models $\theta_{1,k}^{LIV}, \ldots, \theta_{R,k}^{LIV}$ trained with I_k^c, compute the following loss in the I_k:
$$L(I_k; \theta_{r,k}^{LIV}, \hat{\eta}_k) = \int_{\mathcal{T}} \left\{ f_1(t, V; \theta_{r,k}^{LIV}) E(Y|X, Z=t) - f_2(t, V; \theta_{r,k}^{LIV}) E(D|X, Z=t) \right\} dt \\ + f_1(Z, V; \theta_{r,k}^{LIV}) \left\{ \frac{Y - E(Y|X,Z)}{E(Z|X)} \right\} - f_2(Z, V; \theta_{r,k}^{LIV}) \left\{ \frac{D - E(D|X,Z)}{E(Z|X)} \right\},$$

where $\mathscr{T}$ is the support of t and f_1 and f_2 are defined as follows:

$$f_1(z, v, \theta_r^{LIV}) = 2\frac{\partial}{\partial t}\left\{\omega(t, v)\theta_r^{LIV}(t, v)\right\}|_{t=z}$$
$$f_2(z, v, \theta_r^{LIV}) = \frac{\partial}{\partial t}\left\{\omega(t, v)\theta_r^{LIV}(t, v)^2\right\}|_{t=z}.$$

Select the model $\hat{\theta}_{r,k}^{LIV}$ with the smallest loss for each k. Next, pick the model that has the lowest average loss across K:

$$\hat{r} = \arg\min_{r\in\mathscr{R}} E_K \int L\big(w; \hat{\theta}_{r,K}^{LIV}, \hat{\eta}_K\big)\, dP_K(w).$$

This gives the LIV function $\hat{\theta}^{LIV}(\cdot)$. Then, learn the weights γ using optimization methods to solve $\sum_k \sum_{w\in I_k} \varphi(w; \gamma, \hat{\eta}_k) = 0$ for $\varphi(\cdot)$ given in Equation (6).

The ATE,ATT and $LATE$ estimators from such procedure obey $\sigma^{-1}\sqrt{n}(\hat{\theta} - \theta) \rightarrow N(0, 1)$, with the variance given by

$$\sigma^2 = E[\psi^2(W; \theta, \eta)]. \tag{9}$$

The estimated variance is evaluated by replacing θ with each $\hat{\theta}$ found from the scores and taking the empirical equivalent of (9). Moreover, the confidence intervals can be calculated as usual by $[\hat{\theta} \pm \Phi^{-1}(1-\alpha/2)\hat{\sigma}/\sqrt{n}]$.

To make the results more robust to sample splitting, the above procedure is repeated S times, preferably with $S \geq 100$. After obtaining $\{\hat{\theta}_s\}_{s=1}^S$, we report the $median\{\hat{\theta}_s\}_{s=1}^S$, since it is more robust to outliers compared to the mean and a conservative adjusted standard error to incorporate the variation introduced by sample splitting:

$$\hat{\sigma}^{median} = median\left\{\sqrt{\hat{\sigma}_s^2 + (\hat{\theta}_s - \hat{\theta}^{median})^2}\right\}_{s=1}^S.$$

The standard errors for the LIV curve are computed with bootstrapping. The advantages of machine learning methods and in particular NNs and DNNs, compared to some of the usual non-parametric methods broadly used in econometrics are discussed in depth in Section D of the supplementary material along with statistical details of how they are specified and implemented.

6 Results

In this section we report results obtained from the double machine learning method. We first discuss results for the treatment effects of public subsidies on R&D intensity. We then report results for the treatment effects on R&D output based on the exogeneity assumption and on endogenous treatments. Our reported findings the LIV curve are for the average firm in the sample.

6.1 Effects of public subsidies on R&D intensity

There are two important effects of interest to analyze: the existence of total crowding out and the existence of partial crowding out effects. Total crowding out means that the firm would have had the same or lower total R&D expenditure had it financed the R&D privately. This means that the efforts financed with private funding would exceed the efforts induced by public funding, thus making the subsidy undesirable. With partial crowding out total expenditures after subsidies increase but the private portion of the subsidies is lower than the private expenditure had it not received the subsidies. What we observe in the dataset is only the total expenditure (private plus the subsidy) and thus while we can test for total crowding out we can not test for partial crowding out as we do not observed the private expenditures of firms.

Network	Architecture	Learning Rate	Reg. Const.	ATE (in %)	ATT (in %)
Panel A: All firms					
Outcome	{64}	[0.01, 0.001, 1e-04]	0	6.99*** (0.4613)	8.51*** (0.4955)
Prop. score	{64}	[0.001, 1e-04, 1e-05]	0		
Outcome	{64,64}	[0.001, 1e-04, 1e-05]	0.0001	6.72*** (0.3768)	8.43*** (0.3899)
Prop. score	{64,64}	[0.001, 1e-04, 1e-05]	0.001		
Panel B: High-tech and knowledge-intensive firms					
Outcome	{128}	[0.05, 0.005, 5e-04]	0	10.47*** (0.7968)	11.74*** (0.9241)
Prop. score	{64}	[0.001, 1e-04, 1e-05]	0		
Outcome	{256,256}	[0.01, 0.001, 1e-04]	0	10.08*** (0.9809)	10.81*** (1.2560)
Prop. score	{64,64}	[0.001, 1e-04, 1e-05]	0		

Table 4: ATE and ATT of public subsidies on R&D intensity for all firms (Panel A) and high-tech and knowledge-intensive firms (Panel B) in CIS 20010-CIS2014. The results are obtained using a 2-fold random partition ($K = 2$) and 100 different sample splits ($S = 100$) with point estimates calculated with the median. The first block in each panel shows the results from a shallow neural network with a Relu activation function and the second block shows the results from a deep neural network with two layers and Relu activation function. The adjusted standard errors are reported in parenthesis.

To test the total crowding out hypothesis the expected outcome of firms receiving the funding has to be compared with the counterfactual outcome had they not received the funding, i.e., $E[Y_1 - Y_0]$ or $E[Y_1 - Y_0|D = 1]$. Table 4, Panel A, shows that on average, funding induces a 6.99% (significant at the 1% level) increase in firm expenditures in R&D relative to turnover, and higher R&D intensities. While ATE is of interest if policy-makers are considering a full economy shift from a baseline into an alternative state (an intervention that could be universally expanded), or if the subsidy assignment is random, another treatment effect could be of more interest, the ATT. The ATT is more relevant if the policy-makers are questioning or considering the elimination of a program already in place and it shows the average gain from the program for a treated firm randomly drawn from the treated population. The ATT coefficients are shown in the last column of Table 4 and indicate that, among firms that receive funding, the difference in R&D intensity resulting from the subsidy assignment is around 8.5% (significant at 1% level). Thus, the hypothesis of full crowding out effects between public and private innovation funds can be ruled out. Similar results but smaller in magnitude are also found for the German service sector, with a difference of 5.7% (Czarnitzki and Fier (2002)) or for Eastern Germany with a difference of 3.9% Almus and Czarnitzki (2003).

Table 4, Panel B, shows that for firms in the high-tech or knowledge-intensive industries[7] that receive public support the expenditure as a percentage of turnover is about 10% higher than the firms that do not receive the support. The ATT is also positive with a coefficient of about 11%. Both, the ATE and ATT are significant at 1% significance level. Full crowding out also can be ruled out for high-tech and knowledge-intensive firms.

Next, as a robustness check, we report results based on random forests to learn the nuisance parameters. We use random forests as opposed to gradient boosting methods since with the latter there is a higher risk of overfitting and poor performance out of sample (see Hastie et al. (2009)).The results are similar to those obtained with a neural network and a deep neural network and are positive and significant, for both the sample of all firms and the subsample of high-tech and knowledge-intensive firms (see Table 5). Decision trees and Lasso give higher coefficients, as shown in Table 6 and Table 7. Lasso also gives higher adjusted standard errors compared to other methods, possibly due to the higher out of sample prediction error from a linear model. The mean squared error also is higher with a Lasso model, as shown in Table 12, while the neural network provides the lowest means squared error for both the outcome and treatment model.

[7] See Section C of the supplementary material for a classification of industries based on technology and knowledge intensity.

Random Forests	ATE (in %)	ATT (in %)
Sample of all firms	7.56***	9.32***
	(0.804)	(1.3948)
Subsample of high-tech and knowledge-intensive firms	11.10***	11.32***
	(1.4484)	(1.6448)

Table 5: ATE and ATT of public subsidies on R&D intensity for all firms (first row) and high-tech and knowledge-intensive firms (second row) in CIS 20010-CIS2014. The nuisance functions are learned with a random forest composed of 1000 trees. The results are obtained using a 2-fold random partition ($K = 2$) and 100 different sample splits ($S = 100$) with point estimates calculated with the median. The adjusted standard errors are reported in parenthesis.

Decision Trees	ATE (in %)	ATT (in %)
Sample of all firms	7.815***	9.961***
	(0.9965)	(1.743)
Subsample of high-tech and knowledge-intensive firms	11.088***	13.12***
	(1.556)	(1.8981)

Table 6: ATE and ATT of public subsidies on R&D intensity for all firms (first row) and high-tech and knowledge-intensive firms (second row) in CIS 20010-CIS2014. The nuisance functions are learned with decision trees. The results are obtained using a 2-fold random partition ($K = 2$) and 100 different sample splits ($S = 100$) with point estimates calculated with the median. The adjusted standard errors are reported in parenthesis.

Lasso	ATE (in %)	ATT (in %)
Sample of all firms	11.86***	12.515***
	(1.9161)	(2.044)
Subsample of high-tech and knowledge-intensive firms	13.09***	14.94***
	(2.128)	(2.1852)

Table 7: ATE and ATT of public subsidies on R&D intensity for all firms (first row) and high-tech and knowledge-intensive firms (second row) in CIS 20010-CIS2014. The results are obtained using a 2-fold random partition ($K = 2$) and 100 different sample splits ($S = 100$) with point estimates calculated with the median. The adjusted standard errors are reported in parenthesis.

6.2 Effects of public subsidies on R&D output under exogeneity

In this section the treatment effects of public subsidies on innovation output are reported under the exogeneity assumption. Panel A of Table 8 shows that the likelihood of innovation increases by 0.49 if the firm is subsidized. For the subgroup of high-tech and knowledge-intensive firms this effect is slightly smaller, with a coefficient of approximately 0.47 (see Panel B of Table 8). In all cases, the estimator of ATE is significant at the 1% significance level. The ATT on the other hand is much smaller but still significant, with a coefficient of 0.28 for all firms and approximately 0.24 for the high-tech and knowledge-intensive firms. These results suggest that public subsidies lead to a higher propensity to generate innovation output. These results are similar, but less in magnitude compared to what is usually found in the literature, for example Bronzini and Piselli (2016)[8] who find a coefficient of around 0.7. They find an even higher coefficient (1.114) for small firms and an insignificant effect for large firms.

Network	Architecture	Learning Rate	Reg. Const.	ATE	ATT
Panel A: All firms					
Outcome	{256}	[0.01, 0.001, 1e-04]	0	0.491*** (0.0029)	0.283*** (0.0031)
Prop. score	{64}	[0.001, 1e-04, 1e-05]	0		
Outcome	{64,64}	[0.01,0.001,1e-04]	0	0.495*** (0.0050)	0.288*** (0.0059)
Prop. score	{64,64}	[0.01,0.001,1e-04]	0		
Panel B: High-tech and knowledge-intensive firms					
Outcome	{64}	[0.01, 0.001, 1e-04]	0	0.462*** (0.0039)	0.251*** (0.0043)
Prop. score	{64}	[0.001, 1e-04, 1e-05]	0		
Outcome	{256,256}	[0.01,0.001,1e-04]	0	0.471*** (0.0024)	0.237*** (0.0036)
Prop. score	{256,256}	[0.05,0.005,5e-04]	0		

Table 8: ATE and ATT of public subsidies on R&D output for all firms (Panel A) and high-tech and knowledge-intensive firms (Panel B) in CIS 2008-CIS2014. The results are obtained using a 2-fold random partition ($K = 2$) and 100 different sample splits ($S = 100$) with point estimates calculated with the median. The first block of each panel shows the results from a shallow neural network with a Relu activation function and the second block shows the results from a deep neural network with two layers and Relu activation function. The adjusted standard errors are reported in parenthesis.

As a robustness check, random forests are used as an alternative to neural networks and deep neural networks to learn the nuisance parameters. The results are shown in Table 9. The coefficients are all significant at the 1% significance level and they are quite similar to those obtained with the

[8] They analyze the effects of subsidies on innovation for the Italian firms in the Emilia-Romagna region via a regression discontinuity design and use as a proxy for innovation whether or not the firm has applied for at least one patent.

neural and deep neural network. Also, the results obtained with decision trees and with Lasso do not differ significantly, though Lasso appears to overestimate the treatment effects and its standard errors are larger. Looking at Table 12, neural networks give the lowest mean squared error out of sample and thus estimates from a neural network are more reliable based on this metric

Random Forests	**ATE**	**ATT**
Sample of all firms	0.496***	0.271***
	(0.0048)	(0.0056)
Subsample of high-tech and knowledge-intensive firms	0.445***	0.251***
	(0.0039)	(0.0053)

Table 9: ATE and ATT of public subsidies on R&D output for all firms (first row) and high-tech and knowledge-intensive firms (second row) in CIS 20010-CIS2014. The nuisance functions are learned with a random forest composed of 1000 trees. The results are obtained using a 2-fold random partition ($K = 2$) and 100 different sample splits ($S = 100$) with point estimates calculated with the median. The adjusted standard errors are reported in parenthesis.

Decision Trees	**ATE**	**ATT**
Sample of all firms	0.502***	0.293***
	(0.0041)	(0.0048)
Subsample of high-tech and knowledge-intensive firms	0.483***	0.254***
	(0.0032)	(0.0046)

Table 10: ATE and ATT of public subsidies on R&D output for all firms (first row) and high-tech and knowledge-intensive firms (second row) in CIS 20010-CIS2014. The nuisance functions are learned with decision trees. The results are obtained using a 2-fold random partition ($K = 2$) and 100 different sample splits ($S = 100$) with point estimates calculated with the median. The adjusted standard errors are reported in parenthesis.

Lasso	**ATE**	**ATT**
Sample of all firms	0.510***	0.301***
	(0.0049)	(0.0058)
Subsample of high-tech and knowledge-intensive firms	0.514***	0.271***
	(0.0067)	(0.0069)

Table 11: ATE and ATT of public subsidies on R&D output for all firms (first row) and high-tech and knowledge-intensive firms (second row) in CIS 20010-CIS2014. The nuisance functions are learned with Lasso. The results are obtained using a 2-fold random partition ($K = 2$) and 100 different sample splits ($S = 100$) with point estimates calculated with the median. The adjusted standard errors are reported in parenthesis.

Model	Neural Network	Random Forest	Regression Tree	Lasso
R&D Intensity	0.0009	0.0011	0.0013	0.0022
R&D Output	0.0650	0.0831	0.0894	0.1322
Propensity Score	0.0698	0.0932	0.1160	0.1247

Table 12: Mean squared error of R&D intensity, R&D output and propensity score between different machine learning methods in the test set (out of sample).

6.3 Does the funding source matter?

We next report the average treatment effects for different types of public support: (0) no public support received, (1) received public support only from local and regional authorities, (2) received public support only from the central government and (3) received public support only from the EU. Results are in Figure 13.

ATE	R&D Intensity	R&D Output
$Y_{LOC} - Y_0$	5.79*** (1.131)	0.41*** (0.020)
$Y_{GMT} - Y_0$	6.97*** (1.277)	0.38*** (0.015)
$Y_{EU} - Y_0$	7.26*** (1.352)	0.47*** (0.022)
$Y_{EU} - Y_{LOC}$	1.62* (0.921)	0.03 (0.002)
$Y_{EU} - Y_{GMT}$	0.94 (0.893)	0.07 (0.005)
$Y_{GMT} - Y_{LOC}$	0.60 (0.638)	-0.05 (0.004)

Table 13: Average treatment effects from a multiple treatment framework. Y_{LOC} denotes the potential outcome when receiving support from local regional authorities, Y_{GMT} denotes the potential outcome when receiving support from central government, Y_{EU} denotes the potential outcome when receiving support from the EU, and Y_0 denotes the potential outcome when no support is received. The standard errors are shown in parenthesis.

These results suggests that there are no significant differences in the potential outcome (R&D intensity and output) from different sources of the funds. The difference between EU funds and local funds is only significant at 10% significance level for R&D intensity. This is probably due to the fact that EU subsidies are much higher than the local subsidies. The ATE on R&D intensity from receiving any subsidy versus not receiving any subsidy at all is around 6 – 7%, while the ATE on R&D output is around 0.4.

6.4 Effects of public subsidies on R&D output under endogeneity

We next discuss the MTE estimation results[9] when we allow for endogeneity in the output model[10]. We obtain treatment effects for the marginal firm, i.e., the firm that is at the margin of receiving a subsidy for a particular level of the instrument. The MTE or LIV curve for an average firm is shown in Figure 5 and indicates that subsidies are more effective for firms with low levels of resistance, i.e., those firms that switch to the treatment for very low values of the instrument. The maximum treatment effect is achieved at the lowest resistance level and it amounts to around 0.2. This is lower than any of the treatment effects found under exogeneity, but it is still significant. For more resistant firms, the effects decrease and slowly converge to zero. On average, across levels of resistance, these effects are approximately 0.04 . While lower than the coefficients found under exogeneity, these effects are still non negligible and suggest that public support causes a 15% increase in the probability of generating innovative output. Thus, even when accounting for endogeneity, we observe positive and significant effects of public support on R&D output. This suggests that in addition to higher innovation intensity, public subsidies also translate into higher innovation output.

[9]The MTE estimation is used as an alternative and more informative way of exploiting the continuous instrument. But it is possible to use LATE instead by dichotomizing the continuous instrument.

[10]We focus on the endogeneity in the output model motivated by Bloom et al. (2009); Nallari and Bayraktar (2010); Ali et al. (2017) and others. This literature shows evidence that management characteristics are important determinants of the probability to generate R&D output. Given that this information is not observed in our dataset it is possible that the conditional treatments are no longer randomly assigned and causing endogeneity in our output model. Endogeneity in the input model is not considered in this paper.

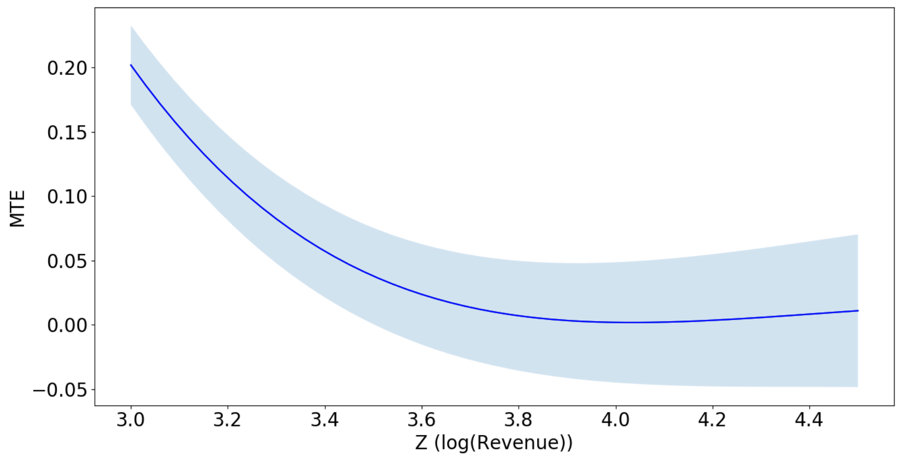

Figure 5: Local IV/MTE curve for an average firm.

7 Conclusions

The literature on the effects of public subsidies on innovation is broad, yet there is very limited work on exploiting fully data-driven models. While many research studies on this topic acknowledge selection into treatment, many assume that after controlling for the observed firm's characteristics, the assignment of the subsidy is random. Our paper introduces another empirical tool for examining the effects of public subsidies on innovation–a fully data-driven model–and accounts for the endogeneity of subsidy assignments. In addition, we use a more representative sample of more than 272,000 observations from European firms, with information about firms' activities over the period 2008 – 2014.

Using the double machine learning semiparametric approach and assuming that unconfoundedness holds, we find that there exists a significant and positive effect of public subsides on both innovation input and innovation output as suggested by the ATE and ATT. These results rule out total crowding out effects, even though data limitations prevent us from making inferences about partial crowding out effects. Treatment effects of subsidies on high-tech firms' investments are higher compared to the effects on the whole sample. Our results suggest that a high-tech firm with, for example, EUR 100,000 of turnover would spend EUR 11,000 less, had it not received a subsidy[11]. This increase in R&D input and output is important because it induces learning by doing, and the production of new goods and services creates more effective ways to use existing resources and elements that have been shown to translate into an increase in productivity (see for e.g. Griliches (1979); Harhoff (1998);

[11] These numbers are based on the estimation results using (deep) neural networks since the out of sample (test set) mean squared errors were lower compared to the other estimation methods. Nevertheless, all methods give comparable results.

Ortega-Argilés et al. (2011); Grifell-Tatjé et al. (2018)), thus also increasing per capita income and overall welfare based on this measure.

Even though we observe higher differences between high-tech and knowledge-intensive firms in R&D intensity, the treatment effects on high-tech firms's R&D output do not seem to be very different from the treatment effects on the sample of all firms. The higher increase in innovation intensity for high-tech firms does not seem to be followed by a higher increase in R&D output. This suggests that the subsidy increases the R&D output through different channels, not only through the effect it has on R&D intensity. Nevertheless, public funding increases the propensity to generate innovative output by around 0.5 on average, and the propensity for the treated firms by 0.28. This means that a subsidized firm with 50% chance of innovating would have had less than a 25% chance of innovating without the subsidy.

It is also found that there are marginal differences in potential outcomes from receiving one type of support versus the other. There exist a 1.6% difference between the potential R&D intensity for support from the EU and the potential R&D intensity from receiving local support. This is potentially due to the fact that the amount of funding and grants from the EU are higher compared to the local ones. However, the source of the support had no effect on R&D output.

The results for the effects of public support on R&D output obtained under exogeneity are still robust, though smaller, when we account for treatment endogeneity. The treatment effects for the marginal firm are positive and significant for low resistance firms[12], and they slowly converge to zero for higher resistance firms. On average, for these marginal firms, the magnitude is small, but non-trivial. This has important policy implications, as it suggests that public support in addition to increasing innovation intensity significantly impacts innovation output. Thus, the vast amount of funding injected in the economy delivers the intended outcomes: it increases R&D efforts and R&D output, which in turn drive long-run growth, improvements in standards of livings, prosperity and welfare.

[12]Low resistance refers to firms that are very likely to be assigned the treatment even when the instrument values are very low (government revenue in this case). These are firms that are considered to be more "deserving"/"in need" of the treatment compared to firms that would get the treatment only in cases when the instrument value is very large (a larger government revenue). The latter firms would be considered "high resistance".

References

Acs, Z. J. and Audretsch, D. B. (2003). Innovation and technological change. In *Handbook of entrepreneurship research*, pages 55–79. Springer.

Aghion, P. and Howitt, P. (1990). A model of growth through creative destruction. Technical report, National Bureau of Economic Research.

Ahn, S. C., Lee, Y. H., and Schmidt, P. (2013). Panel data models with multiple time-varying individual effects. *Journal of econometrics*, 174(1):1–14.

Aigner, D., Lovell, C. K., and Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. *Journal of econometrics*, 6(1):21–37.

Ali, Z., Sun, H., and Ali, M. (2017). The impact of managerial and adaptive capabilities to stimulate organizational innovation in SMEs: A complementary PLS–SEM approach. *Sustainability*, 9(12):2157.

Almus, M. and Czarnitzki, D. (2003). The effects of public R&D subsidies on firms' innovation activities: The case of Eastern Germany. *Journal of Business & Economic Statistics*, 21(2):226–236.

Angrist, J. D. and Imbens, G. W. (1995). Identification and estimation of local average treatment effects. Technical report, National Bureau of Economic Research.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University press.

Arrow, K. J. (1962). *Economic Welfare and the Allocation of Resources for Invention*, pages 609–626. Princeton University Press.

Arrow, K. J. (1971). The economic implications of learning by doing. In *Readings in the Theory of Growth*, pages 131–149. Springer.

Audretsch, D. B. (1995). *Innovation and Industry Evolution*. MIT Press.

Audretsch, D. B., Link, A. N., and Scott, J. T. (2002). Public/private technology partnerships: Evaluating SBIR-supported research. *Research Policy*, 31(1):145–158.

Austin, P. C. (2007). The performance of different propensity score methods for estimating marginal odds ratios. *Statistics in Medicine*, 26(16):3078–3094.

Austin, P. C., Grootendorst, P., Normand, S.-L. T., and Anderson, G. M. (2007). Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: A Monte Carlo study. *Statistics in Medicine*, 26(4):754–768.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.

Bloom, N., Kretschmer, T., and Van Reenan, J. (2009). Work-life balance, management practices and productivity. In *International Differences in the Business Practices and Productivity of Firms*, pages 15–54. University of Chicago Press.

Breiman, L. et al. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231.

Bronzini, R. and Piselli, P. (2016). The impact of R&D subsidies on firm innovation. *Research Policy*, 45(2):442–457.

Busom, I. (2000). An empirical evaluation of the effects of R&D subsidies. *Economics of Innovation and New Technology*, 9(2):111–148.

Carmichael, J. (1981). The effects of mission-oriented public R&D spending on private industry. *The Journal of Finance*, 36(3):617–627.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314.

Czarnitzki, D. and Fier, A. (2001). Do R&D subsidies matter? Evidence for the German service sector. Technical report, ZEW Discussion Papers.

Czarnitzki, D. and Fier, A. (2002). Do innovation subsidies crowd out private investment? Evidence from the German service sector. Technical report, ZEW Discussion Papers.

David, P. A., Hall, B. H., and Toole, A. A. (2000). Is public R&D a complement or substitute for private R&D? A review of the econometric evidence. *Research Policy*, 29(4-5):497–529.

Farrell, M. H., Liang, T., and Misra, S. (2018). Deep neural networks for estimation and inference. *arXiv preprint arXiv:1809.09953*.

Fier, A. (2002). *Staatliche Förderung industrieller Forschung in Deutschland: eine empirische Wirkungsanalyse der direkten Projektförderung des Bundes*. Nomos-Verlag-Ges.

Frölich, M. (2007). Nonparametric iv estimation of local average treatment effects with covariates. *Journal of Econometrics*, 139(1):35–75.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. (2017). Neural message passing for quantum chemistry. *arXiv preprint arXiv:1704.01212*.

Görg, H. and Strobl, E. (2007). The effect of R&D subsidies on private R&D. *Economica*, 74(294):215–234.

Grifell-Tatjé, E., Lovell, C. K., and Sickles, R. C. (2018). *The Oxford Handbook of Productivity Analysis*. Oxford University Press.

Griliches, Z. (1979). Issues in assessing the contribution of research and development to productivity growth. *The Bell Journal of Economics*, pages 92–116.

Griliches, Z. (1998). Patent statistics as economic indicators: A survey. In *R&D and Productivity: The Econometric Evidence*, pages 287–343. University of Chicago Press.

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410.

Hamberg, D. (1966). *R&D: Essays on the Economics of Research and Development*. Random house, New York, NY.

Harhoff, D. (1998). R&D and productivity in German manufacturing firms. *Economics of Innovation and New Technology*, 6(1):29–50.

Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. (2017). Deep iv: A flexible approach for counterfactual prediction. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1414–1423. JMLR. org.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Tibshirani, R., and Friedman, J. (2009). Random forests. *The elements of statistical learning: Data mining, inference, and prediction*, pages 587–604.

Heaton, J., Polson, N., and Witte, J. H. (2017). Deep learning for finance: deep portfolios. *Applied Stochastic Models in Business and Industry*, 33(1):3–12.

Heckman, J., Ichimura, H., Smith, J., and Todd, P. (1998). Characterizing selection bias using experimental data. Technical report, National Bureau of Economic Research.

Heckman, J. J., Ichimura, H., Smith, J., and Todd, P. (1996). Sources of selection bias in evaluating social programs: An interpretation of conventional measures and evidence on the effectiveness of matching as a program evaluation method. *Proceedings of the National Academy of Sciences*, 93(23):13416–13420.

Heckman, J. J. and Vytlacil, E. (2005). Structural equations, treatment effects, and econometric policy evaluation 1. *Econometrica*, 73(3):669–738.

Helpman, E. (2009). *The Mystery of Economic Growth*. Harvard University Press.

Heshmati, A., Loof, H., et al. (2005). The impact of public funds on private R&D investment: New evidence from a firm level innovation study. Technical report.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97.

Hussinger, K. (2008). R&D and subsidies at the firm level: An application of parametric and semiparametric two-step selection models. *Journal of Applied Econometrics*, 23(6):729–747.

Johansson, F., Shalit, U., and Sontag, D. (2016). Learning representations for counterfactual inference. In *International Conference on Machine Learning*, pages 3020–3029.

Kang, J. D., Schafer, J. L., et al. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):523–539.

Kennedy, E. H., Lorch, S., and Small, D. S. (2019). Robust causal inference with continuous instruments using the local instrumental variable curve. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(1):121–143.

King, G. and Nielsen, R. (2019). Forthcoming. "Why propensity scores should not be used for matching.". *Political Analysis*, 27(4).

Klette, T. J., Møen, J., and Griliches, Z. (2000). Do subsidies to commercial R&D reduce market failures? Microeconometric evaluation studies. *Research Policy*, 29(4-5):471–495.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105.

Lach, S. (2002). Do R&D subsidies stimulate or displace private R&D? Evidence from Israel. *The Journal of Industrial Economics*, 50(4):369–390.

Levin, R. C., Cohen, W. M., and Mowery, D. C. (1985). R&D appropriability, opportunity, and market structure: New evidence on some Schumpeterian hypotheses. *The American Economic Review*, 75(2):20–24.

Lichtenberg, F. R. (1987). The effect of government funding on private industrial research and development: A re-assessment. *The Journal of Industrial Economics*, 30(1):97–104.

Lichtenberg, F. R. (1988). The privateR&D investment response to Federal design and technical competitions. *The American Economic Review*, 78(3):550–559.

Lipsitz, S. R., Ibrahim, J. G., and Zhao, L. P. (1999). A weighted estimating equation for missing covariate data with properties similar to maximum likelihood. *Journal of the American Statistical Association*, 94(448):1147–1160.

Lucas, R. E. (1998). On the mechanics of economic development. *Econometric Society Monographs*, 29:61–70.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, 23(19):2937–2960.

Marino, M., Lhuillery, S., Parrotta, P., and Sala, D. (2016). Additionality or crowding-out? An overall evaluation of public R&D subsidy on private R&D expenditure. *Research Policy*, 45(9):1715–1730.

Mullainathan, S. and Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

Nallari, R. and Bayraktar, N. (2010). *Micro efficiency and macro growth*. The World Bank.

Nelson, R. R. (1959). The simple economics of basic scientific research. *Journal of Political Economy*, 67(3):297–306.

Neugebauer, R. and van der Laan, M. (2005). Why prefer double robust estimators in causal inference? *Journal of Statistical Planning and Inference*, 129(1-2):405–426.

Oecd, E. (2005). Oslo manual: Guidelines for collecting and interpreting innovation data. *Paris 2005, Sp*, 46.

Ortega-Argilés, R., Potters, L., and Vivarelli, M. (2011). R&D and productivity: Testing sectoral peculiarities using micro data. *Empirical Economics*, 41(3):817–839.

Robins, J. and Rotnitzky, A. (2001). Comment on the Bickel and Kwon article,"On double robustness.". *Statistica Sinica*, 11(4):920–936.

Robins, J. M. (2000). Marginal structural models versus structural nested models as tools for causal inference. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pages 95–133. Springer.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121.

Romer, P. M. (1987). Growth based on increasing returns due to specialization. *The American Economic Review*, 77(2):56–62.

Romer, P. M. (1990). Endogenous technological change. *Journal of Political Economy*, 98(5, Part 2):S71–S102.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120.

Shalit, U., Johansson, F. D., and Sontag, D. (2017). Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3076–3085. JMLR. org.

Sickles, R. C. (2022). The ET Interview: Peter Schmidt interviewed by Robin C. Sickles, Rice University. *Econometric Theory*, pages 1–19.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Solow, R. M. (1956). A contribution to the theory of economic growth. *The Quarterly Journal of Economics*, 70(1):65–94.

Stiglitz, J. E. (1988). *Economics of the Public Sector*. WW Norton.

Toivanen, O. and Niininen, P. (2000). Investment, R&D, subsidies, and credit constraints. Working Paper No. 264, Helsinki School of Economics.

Van der Laan, M. J., Laan, M., and Robins, J. M. (2003). *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media.

Varaku, K. (2020). Stock price forecasting and hypothesis testing using neural networks. *Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3597684*.

Vytlacil, E. (2002). Independence, monotonicity, and latent index models: An equivalence result. *Econometrica*, 70(1):331–341.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Wallsten, S. J. (2000). The effects of government-industry R&D programs on private R&D: the case of the small business innovation research program. *The RAND Journal of Economics*, 31(1):82–100.

Wang, N. and Yeung, D.-Y. (2013). Learning a deep compact image representation for visual tracking. In *Advances in Neural Information Processing Systems*, pages 809–817.

Westreich, D., Lessler, J., and Funk, M. J. (2010). Propensity score estimation: neural networks, support vector machines, decision trees (cart), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology*, 63(8):826–833.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT press.