D6.1: Interim report with literature review on estimation of parameters and elasticities with respect to public/private R&D

Deliverable:	D6.1: Interim report with literature review on estimation of parameters and elasticities with respect to public/private R&D
Author(s):	Torben Schubert, U Lund, Maikel Pellens, ZEW
Version:	Revised Draft
Quality review:	All partners
Date:	31/07/2018
Grant Agreement N°:	727073
Starting Date:	01/04/2017
Duration:	24 months
Coordinator:	Dr. Georg Licht
E-mail:	licht@zew.de

Table of Contents

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement No 727073

Project Information Summary

Project Acronym	FRAME
Project Full Title	Framework for the Analysis of Research and Adoption Activities and their Macroeconomic Effects
Grant Agreement	727073
Call Identifier	H2020 - SC6 - CO-CREATION - 2016 -1
Topic	CO-CREATION-08-2016/2017: Better integration of evidence on the impact of research and innovation in policy making
Funding Scheme	Medium-scaled focused research project
Project Duration	1st April 2017 - 31st March 2019 (24 months)
Project Officer(s)	Hinano SPREAFICO (Research Executive Agency) Roberto MARTINO (DG Research and Innovation)
Co-ordinator	Dr. Georg Licht, Zentrum für Europäische Wirtschaftsforschung GmbH (ZEW), Mannheim
Consortium Partners	Centre for Economic Policy Research, London Lunds Universitet, Lund Università Luigi Bocconi, Milan Universitat Pompeu Fabra, Barcelona London Business School
Website	http://www.h2020frame.eu/frame/home.html

Table 1: Project Information Summary

Deliverable Documentation Sheet

Number	D6.1
Title	Interim report with literature review on estimation of parameters and elasticities with respect to public/private R&D
Related WP	WP6
Lead Beneficiary	ULUND
Author(s)	Torben Schubert (ULUND), Maikel Pellens (ZEW)
Contributor(s)	
Reviewer(s)	All partners
Nature	R (Report)
Dissemination level	PU (Public)
Due Date	30.09.2017
Submission Date	
Status	

Table 2: Deliverable Documentation Sheet

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement No 727073

Quality Control Assessment Sheet

Issue	Date	Comment	Author
V0.1	26.06.2017	First draft	Torben Schubert
V0.2	09.08.2017	Second draft	Maikel Pellens
V0.3	05.09.2017	Peer review	Torben Schubert
V1.1	17.07.2018	Revised draft	Maikel Pellens
V1.2	18.07.2018	Peer review	Torben Schubert

Table 3: Quality Control Assessment Sheet

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement No 727073

Disclaimer

The opinion stated in this report reflects the opinion of the authors and not the opinion of the European Commission.

All intellectual property rights are owned by the FRAME consortium members and are protected by the applicable laws in accordance with the FRAME Collaboration Agreement.

All FRAME consortium members are also committed to publish accurate and up to date information and take the greatest care to do so. However, the FRAME consortium members cannot accept liability for any inaccuracies or omissions nor do they accept liability for any direct, indirect, special, consequential or other losses or damages of any kind arising out of the use of this information.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement No 727073

Acknowledgment

This document is a deliverable of the FRAME project, which has received funding from the European Union's Horizon 2020 Programme for Research, Technological Development and Demonstration under Grant Agreement number 727073.

Executive Summary

The models developed in WP1-WP4 describe the general theoretical relationships between the key variables and determine their short and long-run dynamics. The models, for example include key equations describing how accumulated knowledge stock evolves as a function of private or public resources devoted to activities aimed at expanding the technology stock. Other equations describe how, once generated, knowledge stocks need to be adopted before they can be brought into economic use and which kinds/amounts of resources are necessary for that. While the models describe general relationships, the precise outcome of simulation studies based on them, namely the evolution of the endogenous variables, will greatly depend on the specification of the exogenous parameters. For any practical purpose, the models are therefore in need of a sensible parameterization.

Principally, two ways exist to achieve that objective. First, the models can be parameterized so that their predictions are as similar as possible to existing historic data. This approach has the advantage that the parameterized models at least resemble the past well. The disadvantage is that it is unclear whether the high resemblance is due to the inherent qualities of the theoretical model that capture the main features of the real world or just due to specific parameter choices hiding away theoretical deficiencies. If the structural equations indeed provided a bad description of the underlying real world relationships and the parameters mask these deficiencies, there is usually a high risk that models deliver bad predictions of future developments even though they described well the historic evolution of the key endogenous variables.

A second approach takes the models as given and estimates the key parameters of particular equations by econometric techniques. The disadvantage of this approach is its partial nature, as typically only subparts of the model can be estimated simultaneously (in the extreme only one equation can be estimated at a time). The advantage is that the focus is not on making the theoretical overall model, often artificially, fit to the real world but rather on deriving causal estimates of the key parameters. This results in a high degree of statistical validity, as well-developed statistical tests (e.g. considering endogeneity issues) can be used to check the quality of the estimates. Using these estimates has a further advantage: if simulations of the theoretical models based on them fail to deliver good approximations of past data, there is strong evidence that either parameters or the theoretical models are not well-specified. Instead of displaying a tendency to hide away modeling deficiency, as does the historic approach, the econometric approach delivers an apparatus to unveil such problems. Because of that the parameterization efforts in WP6 will rely mostly on econometric estimations aiming at the identification of causal relationships between the key endogenous variables. In specific the following parameters will need to be identified:

- P1: The elasticity of the firms' knowledge stock with respect to private R&D (ρ)
- P2: The elasticity of the public knowledge stock with respect to public R&D (ρ_P)
- P3: The elasticity of the productivity of private R&D spending with respect to public R&D spending (γ)
- P4: The average adoption lag in each country ($\bar{\lambda}$).
- P5: The elasticity of adoption with respect to adoption investments (ρ_λ)
- P6: The elasticity of technology adoption with respect to spending in application-oriented public research organizations (ρ_λ^P)

Estimating these parameters requires at least two important decisions. First, because the parameters often refer to abstract economic quantities it needs to be considered which real world quantities best reflect them and whether data measuring them is available. Second, once decisions on the data have been made, a decision needs to be made on the econometric methodology delivering the good estimates of parameters.

The remainder of the interim report is structured as follows. We will continue by describing estimation of parameters P1, P2, P3, and P5 in Section 1. The estimation of P4 will be described in Section 0. The estimation of P6 will be described in Section 3.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement No 727073

1. Estimating the relationship between public and private patents and R&D (P1, P2, P3)

1.1. The one-sector model

The estimation of the key parameters P1, P2, P3, and P5 will be based on a well-established model initially developed by Bottazzi and Peri (2007). The general model suggests the following relationship:

$$\ln(PAT_{i,t}) = \theta \ln(R\&D_{i,t}) + \phi \ln A_{i,t-1} + \zeta \ln A_{ROW,t-1} + u_{it} \quad (1)$$

where $PAT_{i,t}$ is the change in the knowledge stock of country i in period t as measured by patents. $A_{i,t-1}$ is the one period lag of the knowledge stock of country i, $A_{ROW,t-1}$ is the one period lag of the knowledge stock available in the world and v_i is a country-specific time-constant error-term. If the data are stationary, simple panel-data models such as fixed effects could be applied to estimate Eq. (1a) and (1b). However, under non-stationarity, the estimation becomes more complex. Also note that the model proposed in the context of this project is more complex, because interest is not only in estimating one parameter θ, measuring the effect of total R&D on total patenting, but rather to differentiate the models according to public patents and public R&D as well as private patenting and private R&D. Using the parameter notation from above, we are interested in the following equation describing the private patents, public and private R&D and past knowledge stocks (giving P1 and P3)

$$\ln(A_{i,t}) = \rho \ln(R\&D_{i,t}) + \lambda(\ln R\&D_{it} \cdot \ln R\&D^{P}_{i,t}) + \phi \ln A_{i,t-1} + \zeta \ln A_{ROW,t-1} + u_{it} \quad (2)$$

where Eq. (2) in accordance with WP1 assumes that public R&D can make private R&D more productive as represented by the interaction.

Public R&D however affects private patenting in more ways than making private R&D more productive. It also affects the public knowledge stock, which again affects patenting. The further elasticity of interest is therefore the elasticity of public knowledge stock with respect to public R&D investments. A structurally similar equation can thus be defined as:

$$ln(A^{P}_{it}) = \rho_P(ln\, R\&D^{P}_{i,t}) + \phi_P\, ln\, A_{i,t-1} + \zeta_P\, ln\, A_{ROW,t-1} + u^{P}_{i,t} \quad (3)$$

which incorporates P3. So principally, Eq. (2) and Eq (3) describe central causal relationships necessary is a basis of estimation P1-P3. If the data are stationary, simple panel-data models such as fixed effects could be applied to estimate Eq. (2) and (3). However, under non-stationarity, estimation becomes more complex. Also, note that cointegrated data provides a way of defining also P5. This approach will be described below.

1.1.1. Methods

As written above, estimating Eq. (2) and Eq. (3) is straightforward when data are stationary. However, past research has shown that the underlying time series are non-stationary (Bottazzi and Peri 2007, Bottasso et al. 2017). Such non-stationarity is of limited impact when the time-series are short and the sample size is large, which would at least approximately warrant the adequacy of large-n-fixed-t-asymptotics underlying conventional panel data methods. Since, however, the aim of this project is to determine the dynamics over time, such an approach does not conform with the stated objectives of the project. A time-series approach instead however requires an explicit treatment of non-stationarity. A conventional approach is to transform the underlying time series (e.g. by integration) so that they become stationary. A further approach rests on the concept of cointegration. Non-stationary (I1) time series are said to be cointegrated, if there exists a linear combination of them which is stationary (I0). If time series are cointegrated consistent estimates can be obtained by regular time series estimators such as Dynamic OLS (DOLS) (Mark and Sul 2003). Therefore, if cointegration can be asserted, Eq. (2) and Eq. (3) can be estimated by standard time-series techniques. The resulting coefficients can be interpreted as reflecting the long-run relationship tying the time series together. However, if time series are cointegrated, there exist short-run dynamics, which reflect the speed by which any deviances from the cointegrating long-run relationship are eliminated over time. Engle and Granger (1987) have proven that such cointegrated relationships also have an error correction (ECM) representation (Eq. 4) where the β-coefficients represent the long-run dynamics, the term in brackets in the second row is

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement No 727073

the long-run relationship and ξ is an estimate of the time reversion to the long-run relationship takes.

$$\Delta \ln\left(PAT_{i,t}\right) = \beta_0 + \beta_1 \Delta \ln\left(PAT_{i,t-1}\right) + \beta_2 \Delta\ln\left(R\&D_{i,t-1}\right) + \beta_3 \Delta\ln R\&D^P_{i,t} + \beta_4 \Delta\ln A_{i,t-1} + \beta_5 \Delta\ln A_{ROW,t-1} + \xi\left(\ln\left(PAT_{i,t}\right) - \rho \ln\left(R\&D_{i,t}\right) - \rho_P \ln R\&D^P_{i,t} - \phi \ln A_{i,t-1} - \zeta \ln A_{ROW,t-1}\right) + u_{it} \quad (4)$$

For single-time series models there exist a variety of techniques to estimate Eq. (4) by regular ECM-procedures. Our data however does not easily lend itself so such an estimation because the time series are based only on annual observations and are thus too short. One way is estimate Eq. (4) explicitly based on a dynamic panel data approach (e.g. Arellano and Bond 1991, Arellano and Bover 1995). To implement that the long-run relationship can be replaced by the residual of Eq. (2) which is identical to the long-run relationship. The overall methodology for estimating P1-P3 thus consists of four basic steps:

1. Check all time series for non-stationarity using panel-stationarity tests.
2. If non-stationary I1-relationships are asserted, check for panel cointegration (see e.g. Pedroni (2004) or Westerlund (2007) test)
3. If the time series are cointegrated estimate Eq. (2) by DOLS.
4. Extract the residual from step 3 and estimate Eq. (4) by a system-GMM dynamic panel-data estimator.

The same procedure works analogously for the estimation of Eq. (3) and its short-term dynamics. To obtain results for differentiated by country we will subdivide the overall sample of countries into homogenous groups of countries.

1.1.2. Data

Estimating Eq. (2), Eq. (3), and Eq. (4) requires country-level data on public and private R&D, as well as data on public and private knowledge stocks. The R&D data are publicly available from the OECD at the country level for most OECD countries since 1985 until 2013. Due to missing data points for some countries the data can be missing. The data on the public and private knowledge stocks can be based on either patent data or - at least as concerns the public knowledge stocks on stocks of scientific publications. In principle, both patent and publication data can be extracted either from the OECD/Eurostat (patents) or from the Worldbank (publications), which is again freely available. None of the data would however be available on the sectoral level as is required for the purpose of providing parameters also for WP4, we propose to calculate the country and the sector-level patent data directly on the basis of the ZEW's in-house access to the PATSTAT database. The advantage lies on the possibility to guarantee a common data standard for the all parameter estimation exercises in all work packages.

The patent statistics will be calculated in accordance with international standards as described in the OECD Patent Statistics Manual (OECD, 2009). Specifically, national patent stocks will be calculated as follows:

- The primary measure is the stock of patent applications to the European Patent Office (EPO). The EPO enables single patent filing and granting in European Patent Convention (EPC) member states. As an international patent office, statistics based on EPO patent applications should be less susceptible to home country bias than single national offices (De Rassenfosse et al., 2013). However, as some bias is likely to remain, transnational - or world market - patent applications are also considered. These combine applications at the EPO and PCT applications, and are more reflective of developments in the world technology market than analyses based on a single office or other approaches such as triadic patent families (Frietsch and Schmoch, 2010).

- The reference date for each patent application has been determined based on the patent's priority date, which represents the patent's first date of filing. Compared to other options (date of application, date of publication, date of granting), the priority date is closest to the invention date, and does not suffer from biases due to administrative differences between patent offices

(OECD, 2009). Timelines have been constructed from 1978 to 2013.

- The reference country for each patent application has been determined based on the inventor addresses listed on the application, as these best reflect where the inventive activities have taken place (OECD, 2009). In cases where inventors from multiple countries are listed on one application, fractional counting is applied.

- To differentiate public and private knowledge stocks, each patent application has been assigned to the public or private sector based on the sector in which the assignee is active. This information, used by Eurostat and included in the Patstat database, is generated using the methodology described in Van Looy et al. (2006). For the purposes of this exercise, the sectors identified in the database are aggregrated to the public and private sector as shown in Table 4. Patents which are assigned to the public as well as private sector are fractionally counted.

  We follow the literature (Bottazzi and Peri, 2007) and construct knowledge stocks as the depreciated cumulative sum of patent applications, as:

$$Stock_{i,t+1} = Patent\ applications_{i,t} + (1-\delta) * Stock_{i,t} \quad (5)$$

  The depreciation rate, δ, represents the rate at which new ideas become obsolete and is set at 10%, in line with Bottazzi and Peri. The initial value of the knowledge stock is calculated through the perpetual inventory method, where:

$$Stock_{i,t_0} = \frac{Patent\ applications_{i,t_0}}{g_i+\delta} \quad (6)$$

  The country-specific growth rate g_i is calculated as the average annual patent application growth rate between t_0 and $t_{0+\tau}$. We set τ to 10.[1]

The output of this exercise will therefore be country-specific public and private knowledge stocks, as well as stocks for specific economic sectors (as described in the next section).

Assignee Type	Public	Private
Government non-profit	•	
Company hospital	•	•
University	•	
Company government non-profit	•	•
Company university	•	•
Company government non-profit university	•	•
Government non-profit university	•	
Government non-profit hospital	•	
Company		•
University Hospital	•	
Hospital		•
Individual		•

Table 4: Assignee sector classification. Note: Assignees with sector classification “Unknown” are not included.

[1] Bottazzi and Peri set τ to 5. However, a wider window for τ can reduce problems cause by large time variation in patenting in the early years of the analysis. The analysis will include robustness tests to the parameters chosen.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement No 727073

1.2. Extensions: the multi-sector model

The multi-sector model can principally be estimated by the same methodologies outlined in Section 1.1.1. Eq. (2), Eq. (3), and Eq. (4) are then referring only to a specific sector rather than a specific country. While the estimation procedure does not require any specific adjustment, data availability is more limited and requires additional collection efforts. As concerns R&D data, the OECD is providing data at the ISIC-2-digit level, which can principally serve as basis of for the estimation of the estimations also by sectors. One issue concerns the availability of data, because data is more often missing on the sectoral level. Whether this can be ignored or whether imputation efforts will be warranted will be decided upon implementation. A more serious issue is the fact that in 2008 the ISIC rev. 3.1 classification scheme was replaced by the ISIC rev. 4 classification scheme. The OECD does not reclassify the R&D data but provides still older series classified by rev. 3.1 and newer series classified by rev. 4. Harmonization of the time series is difficult and not unambiguous. This holds despite the existence of concordance tables, because the concordance tables, which are themselves partial, refer to the four digit-level which is unavailable in OECD R&D-data. In particular, in the field of manufacturing (rev. 4 codes 10-35) considerable restructuring has taken place, which makes a manual harmonization even more problematic.

In summary, an inspection of the available data has demonstrated that for many sectors consistent long time series cannot be constructed. Fortunately, for several two digit sectors, the inconsistencies were either not large or even completely absent when a one-to-one correspondence between ISIC 3.1 and ISIC 4 existed. The sectors in Table 1 contain ISIC sectors which are particularly relevant in terms of technology generation and for which a roughly consistent concordance exists. For these sectors it is principally possible to create time series from 1985 until 2013 (with the caveat of course of potentially remaining smaller inconsistencies or missing data). Patent statistics will be generated for these sectors based on the mapping between patent application's International Patent application codes and NACE rev. 2 sectors developed by Van Looy et al (2015). Fractional counting will be applied in cases where patents are assigned to more than one of these ISIC sectors.

	ISIC rev. 3	NACE rev. 2/ISIC rev. 4
Chemicals & Pharmaceuticals	**C42**	**20-21**
Office accounting and medical precision instruments	**C73, C76**	**26**
Electrical equipment n.e.c.	**C74**	**27**
Machinery n.e.c.	**C72**	**28**

Table 5: ISIC rev 3 – NACE rev.2 concordance

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement No 727073

2. The diffusion and adoption parameters (P4, P5)

Estimating the diffusion and adoption parameters requires the existence of data giving information on the long-term adoption of technologies. Aggregate macro-economic time-series do not generally exist. Authors have therefore resorted to specialized and/or survey based datasets. Estimates of P4, i.e. the average adoption lag per country have been determined by Comin and Mestieri (2016) based on the Cross-country Historical Adoption of Technology (CHAT) dataset covering the diffusion of 104 technologies in 161 countries over the last 200 years (Comin and Hobijn 2009). We therefore propose to reuse these sets of parameters.

P5 is only estimable indirectly due to a lack of data. Anzoategui et al. (2016) determine P5, i.e. the elasticity of private adoption with respect to private adoption investment to be 0.925 and thus exhibit slight decreasing returns to scale. It should be noted that the parameter has not been determined by a structural economic model due to the lack of data. Rather, the parameter was chosen so that the simulation results of structural model described in Anzoategui et al. (2016) are consistent with the observed R&D intensities (R&D as share of GDP in the US after 1970). In the absence of sector-level country or sector level data on private diffusion investments, we still propose using the already derived estimator. An alternative would be to exploit the Fraunhofer dataset (see next Section). Although this strategy would lead to an econometrically validated estimate, it should be noted that the Fraunhofer dataset can only deliver information on the elasticity of public adoption investments but not private ones. In that respect, there is a risk that we, potentially unduly, equate the public and the private adoption elasticity. While this option may be discussed, possibly also as a robustness check, the preferred solution is to use the already existing estimate of 0.925.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement No 727073

3. Exploiting the Fraunhofer case (P6)

3.1. Introduction

To derive P6, the elasticity of technology adoption with respect to spending in application-oriented public research organisations, we rely on microeconometric estimations on the level of the firm to obtain a causal estimate of the adoption parameter. For that purpose, the project database of the Fraunhofer Society will be combined with information on German firms derived from the Mannheim Innovation Panel. Identification is based on an instrumental variables approach which exploits first-stage heteroskedasticity in order to identify parameters in settings involving issues of omitted variables (Lewbel, 2012). To the best of our knowledge, there are no other reliable sources that can be exploited to obtain the parameter. The few other studies which have examined the impact of extra universitary research organisations (Robin & Schubert 2013; Kaiser & Kuhn 2012) were forced to rely on much coarser information than the one at our disposal.

Assuming that microeconomic issues of selection and omitted variables can be overcome through the instrumental variables approach, the critical issue arises whether this approach can be used to derive macroeconomic conclusions, especially when considering policy implications. To draw a conclusion regarding the net benefit of applied research organisations, it is not sufficient to show positive effects on the microeconomic level; one also need to show that these effects translate into macroeconomic gains. Even when individual firms robustly gain from interacting with applied research organisations, the net macroeconomic effect might be negative in case gains in one firm come at the detriment of competitors. Therefore, we need to assume additionallity when extrapolating the results to the macroeconomic level. This is a strong assumption, which will likely not hold fully. Hence, the results should be treated with care. Similarly, we need to assume away any other competitive effects caused by interacting with applied research. Note, however, that the microeconomic analysis might at the same time underestimate the macroeconomic impact when spillover effects exist. Again, we need to assume these away as they are difficult to quantify. Another critical issue for extrapolation is the representativity of the analysis for the German economy as a whole. Even though the MIP sample is designed to yield a respresentative sample, allowing us to reasonably assume representativity, we treat this issue with care and provide robustness checks using population weights.

3.2. Data

The empirical analysis is based on two main data sources. The first is the project database provided by the Fraunhofer Society, which covers all projects started between 1997 and 2014.[2] The database contains information on the Fraunhofer institute and department involved, the client's name and address, the title, short description and time span of the project, and any payments related to the project. The database covers 131,158 projects. Section 2.2 presents an in-depth description of the information in the database.

Care was taken to guarantee the confidentiality of the agreements delivered by FhG, particularly with regard to the identities of the client firms. The individuals responsible for executing and checking the match between the FhG data and MIP data did not have access to the agreement data, but only to the name and address of the entities to match to the MIP. Anonymous identifiers were constructed based on the matched data for use in the remainder of the analysis. Furthermore, individuals involved in the database matching were not involved in the remainder of the analysis.

These projects were merged to the Mannheim Innovation Panel (MIP), which is an annual survey collected since 1993 by the Centre for European Economic Research for the Federal Ministry for Education and Research (BMBF). The MIP provides a representative annual sample of German firms with five or more employees (See Aschhoff et al., 2013 for further details). The MIP follows the methodology outlined in the Oslo Manual (OECD and Eurostat, 2005) and is also the German

[2] Approximately 10% of the projects in the database listed start dates before 1997. As these do not seem to represent a full picture of the projects, we omit these from the further analysis. Any payments made to Fraunhofer in the context of these projects in 1997 onwards, however, are taken into account.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement No 727073

contribution to the European Community Innovation Survey. This data has further been amended by data from Germany's largest credit rating agency, Creditreform, for information on firm's age. The present analysis makes use of the 2014 edition of the MIP, including information up to calendar year 2013. Excluding firms which were observed less than three times, the MIP covers 198,385 observations of 30,125 firms between 1996 and 2014.[3]

Both datasets were merged by comparing firm names and address information.[4] Of the 131,158 projects in the Fraunhofer project database, 46.651 projects could be related to 7.781 distinct firms which were surveyed at least once in the MIP survey. Due to nonresponse and the condition of observing a firm at least three times, 32,568 projects, representing 4,495 firms in the MIP panel, were used in the final analysis. This represents 24.8% of the projects in the database.

There are several reasons for the large number of unmatched projects. First, 17% of projects relate to clients outside of Germany. Second, any public clients (such as universities, research centers, and government institutes) are not covered by the MIP and hence remain unmatched. Third, the MIP only presents a representative sample of German firms of roughly 10% of the population (Aschhoff et al, 2013),[5] which does not capture all firms that potentially would contract with Fraunhofer. Fourth, projects were assigned to MIP firms conservatively, requiring a match in both name and address. While this avoids errors based on namesakes, it might also lead to potential underestimation of the degree to which firms make use of Fraunhofer's services.[6]

In the next section, we present a statistical description of the Fraunhofer project database. We base this analysis on the full database of projects starting from 1997 onwards, and not only the part of the data matched to the MIP. After that, we present the variables used for the multivariate analysis, and describe differences between MIP firms which interact with Fraunhofer and firms which do not.

3.2.1. Project descriptions

To gain some insight into the goals and organization of FhG projects, a key word analysis was performed based on the short project descriptions available in the database. Table 6 lists the 20 most common harmonized keywords in the project descriptions.[7] The keywords show that FhG projects cover the full spectrum from studies and analysis to development, application, and implementation. It is likely not the case that the impact of FhG projects is constant across different types of projects. However, the broad nature of the project descriptions limits the inference to be made. In the multivariate analysis, we differentiate between projects which show through their description a clear intent to implement whatever is in the focus of the project, which could be a technology, product, process, or still something different. This allows us to assess whether more downstream projects have an impact that differs from more upstream, abstract projects.[8]

[3] We retain information from 1996 to allow control variables to be lagged with one year.

[4] The matching algorithm takes spelling deviations into account and assigns a score to each potential match. Potential matches with some uncertainty were manually screened for accuracy.

[5] Sample size and coverage varies throughout time.

[6] This is not a crucial issue in the analysis, as we define interactions with Fraunhofer on firms paying at least a certain amount of money. Therefore, the analysis presented here should be robust to some underestimation.

[7] Descriptions were short: the average description is 7 words long, and 90% of descriptions consist of 14 words or less. Keywords in the descriptions were translated from German and harmonized. Common words as well as brands and any identifying information has been removed from the data.

[8] To achieve this we developed and applied the following key: Projects were deemed ‚implementative' when they included words indicating a change or development, such as 'adapt', 'build', 'create', 'construct', 'develop', 'improve', 'innovate', 'integrate', 'intervene', 'install', 'manufacture', 'modify', 'realize', 'restructure'.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement No 727073

Rank	Term	Number projects	Share projects	Rank	Term	Number projects	Share projects
1	Development	6906	5.27%	11	Creation	1363	1.04%
2	Analysis	5348	4.08%	12	Feasibility	1354	1.03%
3	Study	4366	3.33%	13	Process	1336	1.02%
4	System	2481	1.89%	14	Application	1308	1.00%
5	Manufacturing	1776	1.35%	15	Technology	1248	0.95%
6	Supply	1740	1.33%	16	Structure	1112	0.85%
7	Project	1713	1.31%	17	Concept	1077	0.82%
8	Optimization	1687	1.29%	18	Simulation	1064	0.81%
9	Evaluation	1665	1.27%	19	Implementation	1059	0.81%
10	Test	1621	1.24%	20	Phase	1038	0.79%

Table 6: Project keywords

3.3. Variables

In this section, we describe the variables used in the analysis (described in Table 7). This includes the variables that measure the interaction with Fraunhofer, the various outcomes and controls.

3.3.1. Interaction with Fraunhofer

The key explanatory variable of the study captures whether the firm interacted with FhG. As many projects involve little or no payment to FhG, indicating that they are small in size, a threshold needs to be defined to to indicate when projects are of significant size.[9] At the same time, the project data needs to be transposed into the firm-year framework of the MIP. Therefore, the project data was aggregated to the money paid to the Fraunhofer-Gesellschaft for each firm in each year. As most firms are involved in one interaction at a time, this is not a strong assumption to make.[10] A significant interaction was then defined as making a total payment of 13.000 EUR or more to Fraunhofer over the course of a given year.[11] We name this variable FHG_INT. We also define a broader interaction indicator, FHG, that takes value 1 if the firm ever interacted with FHG over the timeframe of the data. Lastly, we define FHG_AMOUNT to capture the size of the annual payment made to FhG.

3.3.2. Outcomes

We approach the characterization of the effect of interaction with FhG on firms from different perspectives. First, firms might be able to grow larger as a result of their interactions. The size of the firm is measured by (TURNOVER; million EUR) and by employee counts (EMPLOYEES). Second, implementing technology with support from FhG might be an efficient way to increase productivity. To capture that, we calculate added value per employee (ADDVAL). Third, Fraunhofer might support firms in the development and commercialization of their own innovative products and processes. We capture these in a direct way through the share of sales stemming from new or improved products introduced by the firm (INNOSALES). Additionally, measures of average employee cost (CPE) and the share of employees with tertiary education (EMP_HIGHED) capture any changes in firm strategy with regard to innovation and R&D by tracing changes in the composition of the workforce.

[9] A small minority of projects involved negative payment, i.e. money going from Fraunhofer to the firm.

[10] In 64% of cases in which a MIP firm interacts with FhG, there is only one interaction in that year. In 18% of cases there are 2, and 3 or more only in 10% of cases.

[11] The Fraunhofer database lists payments made by year. While these could in principle occur at any point after starting a project, payments are typically made in the year after the project is started. Given the fact that the median project lasts one year, this means that payments can be used as a proxy for Fraunhofer activity in that time. The amount of 13.000 EUR was chosen to eliminate projects which are too small in scale to show a significant impact on firm-level performance measures, and approximately resembles the median payment made by firms in the MIP to Fraunhofer in a given year, considering the total payment across all projects in which the firm is engaged. In the robustness checks, we show that this definition holds to stricter definitions of an interaction.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement No 727073

Name	Source		Description
Interaction with Fraunhofer			
FHG	FhG data	Binary	1 if firm ever paid at least 13.000 EUR to FhG
FHG_AMOUNT	FhG data	Numeric	Payment made by firm to FhG in year (tho. EUR), considering all projects in which the firm is involved.
FHG_INT	FhG data	Binary	1 if firm paid at least 13.000 EUR to FhG in year.
Outcomes			
TURNOVER	Mip	Numeric	Turnover of firm in year (million Eur)
EMPLOYEES	Mip	Numeric	Number of employees in year
ADDVAL	Mip	Numeric	Added value per employee (million EUR)
INNOSALES	Mip	Numeric	Share of sales stemming from new or improved products
CPE	Mip	Numeric	Average employee cost (tho. EUR)
EMP_HIGHED	Mip	Numeric	Share of employees with tertiary education
Controls			
RD_INT	Mip	Numeric	R&D expenditures scaled by turnover (ratio)
AGE	Creditreform	Numeric	Years since firm's founding
GROUP	Mip	Binary	1 if firm is member of a group of firms
EXPORT	Mip	Binary	1 if firm indicates to export in year
EAST	Mip	Binary	1 if firm is located in former Eastern Germany
SIZE_(SMALL,_MEDIUM, _LARGE)	Mip	Binary	Categoric indicator of firm size. Small: up to 49 employees. Medium: 50-249 employees. Large: 250+ employees.
Sector	Mip	Categoric	Categoric indicator: 21 sectors (see Table 3)
Year	Mip	Categoric	Categoric indicator: calendar year

Table 7: Variable Definitions

3.3.3. Controls

The analysis aims to estimate the effect of interacting with FhG on firm performance. However, there are some factors that need to be held constant. The degree up to which a firm can profit from interacting with FhG is likely a function of internal R&D capacities (Cohen and Levinthal, 1990). To control for this, we include a measure of own R&D intensity (RD_INT, R&D expenditures scaled by turnover). Likewise, R&D intensity is expected to play an important role for self-selection into Fraunhofer interaction, as firms with more innovation-focused strategies are more likely to have projects with FhG.

We also include a number of more general indicators that capture the competitive situation of the firm. These include the firm's age (AGE) and a dummy indicating whether or not the firm exports (EXPORT). Additionally, we control for broad economic differences within Germany by including a dummy that takes value one if the firm is located in former Eastern Germany (EAST), and control for broad differences along firm size through the inclusion of three firm size categories[12] (SIZE_SMALL, SIZE_MEDIUM, and SIZE_LARGE). We further control for the economic activities of the firm through the inclusion of 21 broad sector indicators and include year fixed effects to account for shared macroeconomic trends.

3.3.4. Firm-level descriptives

Table 8 compares the outcome and control variables for firms that interacted or did not interact with Fraunhofer in the project database. Table 4 shows the same for sector distribution. As shown in the upper panel of table 3, 6% of firm-year observations in the MIP are found to contain interactions with Fraunhofer. On average, a year in which a firm paid money to FhG involves a payment of approximately 37.000 EUR.

Firms that interact with FhG through projects are significantly (p<0.01, two-sided t-test) larger in terms

[12] Small: up to 49 employees. Medium: 50-249 employees. Large: 250+ employees. In estimations not related to size, we control for firm size by including the number of employees as a control variable.

of turnover and employees. The difference is strong, approximating a ten-fold size differential. This is reflected in the firm size categories: whereas 14% of firms that did not interact with FhG are classified as large, 54% of the firms that interact with FhG are large firms. At the same time, firms that engage with FhG draw more sales from new or improved products (18% versus 6%). They also seem to be more productive, as added value per employee is approximately 20% higher among firms that interact with FhG than among firms who do not. Lastly, FhG firms report higher average labour costs per employee (47.49 versus 35.22 Tho. EUR) and a higher share of employees with higher education (30% versus 20%). A similar pattern emerges in terms of the controls: FhG firms are more R&D-intense (10% compared to 3%), tend to be older (37 years versus 28), are more likely to export their products (45% versus 25%), and are more likely to be part of a group (68% versus 52%). FhG firms are more likely to be situated in former Western Germany than in former Easter Germany.

Taken together, these descriptive differences underline the importance of accounting for positive selection bias in the empirical analysis. If left uncontrolled for, the impact of interacting with Fraunhofer will be biased upwards.

	Total			By Fraunhofer Firm Interaction				
				Interacted		Did not Interact		Difference
	Mean	St. Dev	Obs.	Mean	Obs.	Mean	Obs.	
Interaction with FhG								
FHG	0.06	0.24	198385	1.00	17103			
FHG_INT	0.02	0.14	198385	0.24	17103			
FHG_AMOUNT	3.23	53.80	198385	37.23	17103			
Outcomes								
TURNOVER	199.00	3593.82	131822	906.95	11239	133.02	120583	-773.93***
EMPLOYEES	531.56	7253.71	191065	2735.27	16571	322.28	174494	-2412.99***
INNOSALES	0.07	0.17	112029	0.18	7734	0.06	104295	-0.12***
ADDVAL	0.10	0.38	61955	0.12	5641	0.10	56314	-0.02***
CPE	36.23	17.16	77831	47.49	6376	35.22	71455	-12.27***
EMP_HIGHED	0.21	0.25	99873	0.30	8163	0.20	91710	-0.10***
Controls								
RD_INT	0.04	0.58	77974	0.10	6989	0.03	70985	-0.07***
AGE	29.08	32.27	190804	37.44	16707	28.28	174097	-9.16***
EXPORT	0.27	0.44	198385	0.45	17103	0.25	181282	-0.20***
GROUP	0.54	0.50	198385	0.68	17103	0.52	181282	-0.16***
EAST	0.33	0.47	198385	0.27	17103	0.34	181282	0.07***
Firm Size								
SIZE_SMALL	0.56	0.50	191065	0.20	16571	0.59	174494	0.39***
SIZE_MEDIUM	0.27	0.44	191065	0.26	16571	0.27	174494	0.01**
SIZE_LARGE	0.17	0.38	191065	0.54	16571	0.14	174494	-0.40***

Notes: Firm-years. Difference: outcome of two-sided t-test. Stars indicate significance level of t-statistic. ***(,**,*): p < 0.01(,0.05, 0,10). Interaction with Fraunhofer: split made along having ever had Fraunhofer project between 1997 and 2014

Table 8: Firm Summary Statistics

Table 9 shows the sectoral distribution of firms which interacted with FhG or not. Firms interacting with FhG are more likely to be situated in medium and high-tech manufacturing industries (specifically, petroleum and chemical industry, machinery and domestic appliances, electrical machinery, communication equipment, instruments, and automotive), and comparatively less likely to be active in low-tech manufacturing or service industries. The multivariate analysis needs to correct for these differences in sample composition of the FhG and non-FhG firm samples.

Sector	Total	Yes	No	Difference
	Share	Share	Share	
Mining	0.02	0.01	0.02	0.01***
Food and Tobacco	0.04	0.01	0.05	0.04***
Textiles and Leather	0.03	0.01	0.03	0.02***
Wood, paper, and printing	0.06	0.02	0.06	0.04***
Petroleum, coke, and chemical industry	0.04	0.08	0.03	-0.05***
Rubber and plastics	0.04	0.03	0.04	0.00
Glass, ceramics, other non-metallic minerals	0.03	0.03	0.02	-0.01***
Basic metals and metal products	0.07	0.09	0.07	-0.01***
Machinery and domestic appliances	0.07	0.16	0.06	-0.10***
Office appliances and computers, electrical machinery, communication equipment	0.05	0.11	0.04	-0.06***
Medical, precision, and optical instruments	0.05	0.11	0.04	-0.07***
Transportation equipment	0.03	0.07	0.03	-0.05***
Furnitury, jewellery, musical instruments, sports equipment, games and toys	0.02	0.01	0.02	0.01***
Intermediation of trade and wholesale (excl. Trade in motor vehicles)	0.04	0.01	0.05	0.03***
Trade of motor vehicles, maintenance and repair of motor vehicles, petrol stations, repairs	0.02	0.01	0.03	0.02***
Transportation, traffic, and courier services	0.08	0.03	0.08	0.05***
Credit and insurance	0.05	0.03	0.05	0.02***
Data processing and databases; telecommunications	0.05	0.05	0.05	-0.00
R&D and engineering	0.08	0.08	0.08	-0.01**
Legal and tax advice; consultancy; marketing	0.05	0.02	0.05	0.03***
HR, information, and security services, other services for firms	0.07	0.02	0.08	0.06***
Estate and housing, renting of movable items	0.02	0.01	0.02	0.01***

Notes: difference: outcome of two-sided t-test. Stars indicate significance level of t-statistic. ***(,**,*): p < 0.01(,0.05, 0,10). Interaction with Fraunhofer: split made along having ever had Fraunhofer project between 1997 and 2014

Table 9: Sector Distribution of Firms that interacted with Fraunhofer versus firms that did not

3.4. Methods

Estimating the causal effects of project interactions with Fraunhofer institutes on firm and innovation performance can be accomplished by regression techniques. Estimating the causal effect is however not straightforward because of selection on unobservables and endogeneity. This section describes the methods employed in this study to deal with the endogeneity in the relationship between firm performance and Fraunhofer interactions. We will specifically present an in-depth description of how we can exploit scale heteroscedasticity to identify the causal effects.

In order to illustrate how identification works assume the following simple model of the relationship between the firm performance $\mathrm{y_{it}}$ and the cooperation variable FHG_{it}::

$$\mathrm{y_{it}} = \mathrm{x_{it}}\beta + \mathrm{FHG_{it}}\delta + \mathrm{u_{it}} \ (7)$$

where $\mathrm{x_{it}}$ is a vector of control variables and $\mathrm{u_{it}}$ is a structural error term. δ is the central parameter of interest and measures how the interaction variable affects firm performance. Interactions with Fraunhofer institutes cannot be expected to be randomized firms but will depend on a process of mutual selection. However, if the time-varying factors governing the selection process can be sufficiently controlled for in $\mathrm{x_{it}}$ and if $\mathrm{u_{it}}$ contains at best time-constant unobserved heterogeneity, δ can be structurally identified by fixed effects. If, however, selection is based on (time-varying) unobservables, the estimates δ will generally be biased, because the central identification condition that $\mathrm{u_{it}}$ is uncorrelated with any of the vector of observed variables in $1, \dots \mathrm{T}$ (strict exogeneity) will not generally hold.

A prime reason why $\mathrm{u_{it}}$ is correlated with the included regressors, in particular $\mathrm{FHG_{it}}$, is that Fraunhofer interactions positively depend on the firm's innovation capabilities, because more capable firms will be more likely to self-select into collaborative projects and will be more likely to be selected by the Fraunhofer institutes. Moreover, the innovation capabilities are unlikely to be constant over time because of skill accumulation.

Using fixed effects in (7) then will not lead to consistent estimation of δ. To prevent that, we need to identify δ from exogenous variation in the interaction with Fraunhofer induced by instrumental variables. Recently, Lewbel (2012) has demonstrated how scale heteroscedasticity can help to generate instrumental variables. The advantage of using scale heteroscedasticity in our setting is appealing - as compared to regular exclusion restrictions, because it turns out that heteroscedasticity

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement No 727073

is tremendous in our dataset and thus provides a strong basis for identification avoiding issues of weak instruments typically pertaining to the exploitation of natural experiments. We will now describe the statistical mechanism by which scale heteroscedasticity can lead to causal identification. In the following section we will then proceed by investigating the precise nature of heteroscedasticity in our dataset to show that the theoretical assumptions in Lewbel (2012) are met in our case. First applications of the general idea can be traced to Wright (1928). Less general but related applications relying on time-dependent heteroscedasticity in longitudinal data can be found in King et al. (1994) and Sentana and Fiorentini (2000). We based our presentation on simplified cross-sectional models. We note, however, that under regular assumptions necessary for the consistent estimation of panel-data IV models, also the Lewbel approach is consistent. Assume a simplified model without control variables:[13]

$$\mathrm{y_i} = \mathrm{FHG_i}\delta + \mathrm{a_1 capabil_i} + \mathrm{e_{1i}},$$

$$\mathrm{FHG_i} = \mathrm{a_2 capabil_i} + \mathrm{e_{2i}}. \text{ (8a,b)}$$

where we allow that $\mathrm{e_{2i}}$ is heteroscedastic, i.e. it may depend on some vector h_i. Estimating Eq. (8a) by OLS without taking the capability-term into account will result in a biased estimate $\hat{\delta}$. In particular, setting $X = (\mathrm{FHG_1}, \dots, \mathrm{FHG_n})'$, $z = (\mathrm{capabil_1}, \dots, \mathrm{capabil_n})'$ and $y = (\mathrm{y_1}, \dots, \mathrm{y_n})'$, $\hat{\delta}$ can be written as:

$$\hat{\delta} = (X'X)^{-1}X'y$$

$$= (1/n\sum_{i=1}^{n} x_i'x_i)^{-1}\, 1/n\sum_{i=1}^{n} x_i'y_i = \delta + (1/n\sum_{i=1}^{n} x_i'x_i)^{-1}\, 1/n\sum_{i=1}^{n} x_i'(a_1 z_i + e_{1i}) \text{ (9)}$$

The probability limes of Eq. (9) is:

$$\hat{\delta} \overset{p}{\rightarrow}= \delta + a_1 \frac{E(FHG_{it}capabil_i)}{E(FHG_i^2)} = \delta + a_1 \frac{a_2 E(capabil_i^2)}{a_2^2 E\left(capabil_i^2\right)+E(e_{2i}^2)} \text{ (10)}$$

where the second equality follows from replacing $\mathrm{FHG_{it}}$ with Eq. (2b). Although the OLS estimate is generally biased, interestingly, if $E\left(e_{2it}^2\right)$ is large, then the bias will be small. One approach of using this result is to define subsamples with very high residual variance and running regressions only on the subsample. Fisher (1976) calls the dependence of the bias on the first stage error variance near identifiability. We present a graphical representation in Figure 1, where we simulated the Eqs. (8a, b) using $\delta = \alpha_1 = \alpha_2 = 1, \mathrm{e_{1i}} \sim \mathrm{capabil_i} \sim N(0,1)$. The left panel is generated with $\mathrm{e_{2i}} \sim N(0,1^2)$ and the right panel is generated with $\mathrm{e_{2t}} \sim N(0,5^2)$. Obviously, the true parameter is δ is unit. But when running the regression $\mathrm{y_i}$ on $\mathrm{FHG_i}$ we obtain a biased estimate of about 1.5 in the left panel. If we increase the second stage error to variance to 25 (right panel), the estimated slope parameter drops to about 1.04 and is already quite close to the true parameter. The reason for the drop is that the increase in the variance of $\mathrm{e_{2i}}$ weakens the strength of the direct relationship between $\mathrm{FHG_i}$ and the omitted variable $\mathrm{capabil_i}$, which is defined by Eq. (2b), leading to a drop in the bias.

[13] Suppressing the control variables leads to a closed form expression of the bias without matrix algebra, but otherwise does not inhibit the generality of the illustration.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement No 727073

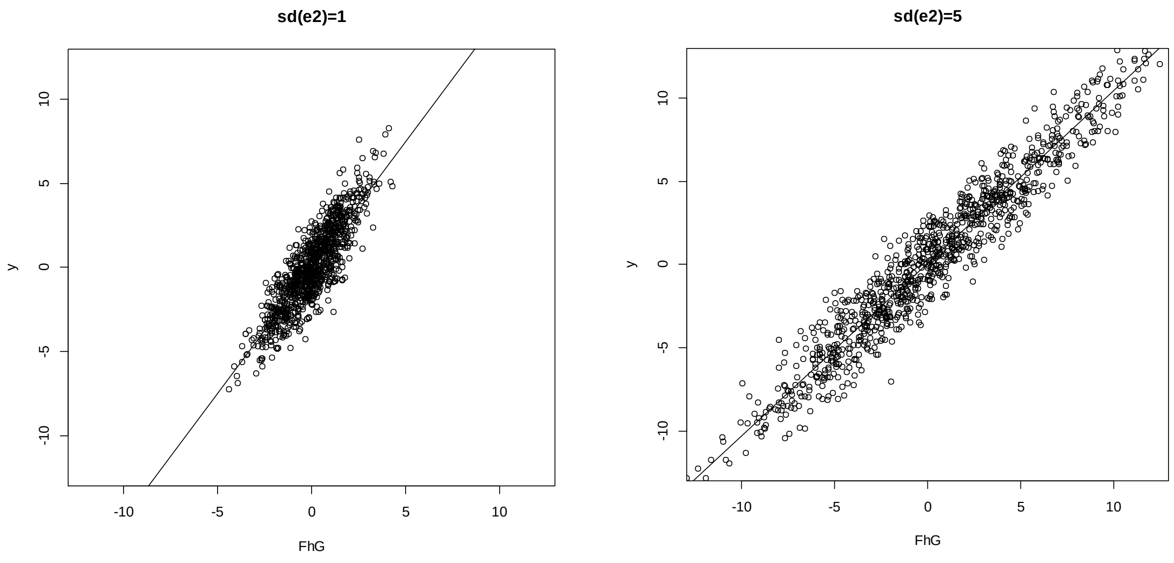

Figure 1: Illustration of near identifiability under low and high error term variance

Two principal ways to exploit the dependence of the bias on the error variance have emerged in the literature. The first approach is the event-study design, which assumes that in specific events the error variance so large that OLS leads approximate identification. However, unless the variance becomes infinite, identification will never be exact. Under certain conditions it is however possible to use heteroscedasticity as a basis for defining instrumental variables, which can solve the identification problem even if the second stage error variance is finite. Eq. (10) gives an intuition: since the omitted variable bias is a function of the first stage error variance, heteroscedasticity implies that not only $E(e_{2i}^2)$ but also the bias in Eq. (10) is a function of the vector h_i. If for example we assume positive scale heteroscedasticity, the bias is the smaller the larger the individual elements of h_i are. Moreover, since h_i appears nowhere else in the model, h_i induces exogenous variation in the model: it affects $\mathrm{FHG_i}$, more precisely its volatility, but it has no effect on $\mathrm{capabil_i}$. Indeed instruments can be defined, which makes use this exogenous information to identify the causal effect.

To illustrate that, we turn to more general version of Eqs. (8a, b) allowing for a vector of control variables $x_i \in \mathbb{R}^k$:

$$\mathrm{y_i} = \mathrm{x_i}\beta + FHG_i\delta + \mathrm{u_i}$$

$$FHG_i = \mathrm{x_i}\zeta + \mathrm{v_i}\text{(11a,b)}$$

with $\mathrm{u_{it}} = \mathrm{a_1 capabil_i} + \mathrm{e_{1i}}$, and $\mathrm{v_i} = \mathrm{a_2 capabil_i} + \mathrm{e_{2i}}$ and $E(e_{2i}^2)$ is allowed to depend on x_{it}. Again, we are not able to consistently estimate the model because of omitted variable bias induced by unobserved $\mathrm{capabil_i}$.

To achieve identification by exploiting heteroscedasticity we make the usual minimal identification assumption that $\mathrm{x_i}$ is exogenous: $E(x_i u_i) = 0$ and $E(x_i v_i) = 0$. Then it can be shown that $\mathrm{z_i} = (\mathrm{x_i} - \mathrm{E(x_i)})\mathrm{v_i}$ is a vector of valid instrument for $\mathrm{FHG_{it}}$ provided that:

$$\mathrm{cov}(\mathrm{x_i} - \mathrm{E(x_{it})}, \mathrm{u_i v_i}) = 0$$

$$\mathrm{cov}\big(\mathrm{x_i} - \mathrm{E(x_i)}, v_i^2\big) \neq 0 \text{ (12a, b)}$$

Because the proof is lengthy and somewhat tedious, we omit here. Yet, it is easy to create some intuitions why these assumptions identify the parameters of interest. Intuitively, Eq. (12b), i.e.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement No 727073

heteroscedastic first stage errors, implies that the instrument $\mathrm{z_i}$ and the endogenous variable are correlated. Using Eq. (11a,b) we can write:

$$cov\left(x_i - \mathrm{E}(\mathrm{x_i}), \mathrm{v_i^2}\right) = E((x_i - \mathrm{E}(\mathrm{x_i}))v_i(FHG_i - \mathrm{x_i}\zeta))$$

$$= E(x_i v_i FHG_i - x_i v_i \mathrm{x_i}\zeta - \mathrm{E}(\mathrm{x_i}) v_i FHG_i + \mathrm{E}(\mathrm{x_i}) v_i \mathrm{x_i}\zeta) = \mathrm{E}(z_i FHG_i) \neq^{!} 0 \quad (13)$$

On the other hand, Eq. (12a) guarantees that $\mathrm{x_i}$ does not simultaneously affect the variance of the unobserved variable. Assuming without loss of generality that the expectation of the unobserved variable is zero, note that

$$\mathrm{cov}(\mathrm{x_i} - \mathrm{E}(\mathrm{x_i}), \mathrm{u_i v_i}) = E(z_i u_i)$$

$$= \mathrm{E}\left(\mathrm{x_i}(\mathrm{a_1 a_2 capabil_i^2} + \mathrm{a_1 capabil_i e_{2i}} + \mathrm{a_2 capabil_i e_{1i}} + \mathrm{e_{1i} e_{2i}})\right) =^{!} 0 \quad (14)$$

Thus, Eq. (12b) is the equivalent of the regular rank condition in IV ensuring that the instruments display some sort of correlation with the endogenous variable. Eq. (12a) is the exogeneity condition which is seen also from the fact that Eq. (14) shows that it is equivalent to requiring that the instruments and the structural error term are uncorrelated. Furthermore, Eq. (14) illustrates the identification assumption: the variation in FHG_i induced by heteroscedastic first stage errors is exogenous only if it does not also affect the variance of the unobserved variable $\mathrm{capabil_i}$.

Implementing the Lewbel estimator is easy by using the sample equivalent of $\mathrm{z_i}$:

$$\widehat{\mathrm{z_i}} = (\mathrm{x_i} - \bar{\mathrm{x}})\widehat{\mathrm{v_i}} \quad (15)$$

where $\widehat{\mathrm{v_i}}$ is the residual from reduced form regression of $\mathrm{FHG_i}$ on the exogenous regressors $\mathrm{x_i}$. $\widehat{\mathrm{v_i}}$ is structurally identified because the parameters in the reduced form regression can always be consistently estimated (Wooldridge, 2002).[14] Since our identification relies on the existence of scale heteroscedasticity, we will explore the precise nature of scale heteroscedasticity in our dataset. In particular, we will identify the size of the firm as the single factor driving heteroscedasticity, while none of the other control variables seems to provide any identification power.

[14] It should be noted that Lewbel-methodology works in broader settings than the omitted variable bias considered here. In specific, even full simultaneity in Eq. (2a) and Eq. (2b) is admissible.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement No 727073

4. References

Anzoategui, D., Comin, D., Gertler, M., & Martinez, J. (2016). Endogenous Technology Adoption and R&D as Sources of Business Cycle Persistence (No. w22005). National Bureau of Economic Research.

Arellano, M., & Bond, S.R. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. Review of Economic Studies, 58.

Arellano, M., & Bover, O. (1995). Another look at the instrumental variable estimation of error-components models. Journal of Econometrics, 68(1), 29-51.

Aschhoff, B., Baier, E., Crass, D., Hud, M., Hünermund, P., Köhler, C., Peters, B., Rammer, C., Schricke, E., Schubert, T., & Schwiebacher, F. (2013). Innovation in Germany – Results of the German CIS 2006 to 2010. ZEW Documentation No. 13-01, Mannheim.

Bottasso, A., Castagnetti, C., & Conti, M. (2015). R&D, Innovation and Knowledge Spillovers: A Reappraisal of Bottazzi and Peri (2007) in the Presence of Cross-Sectional Dependence. Journal of Applied Econometrics, 30(2), 350-352.

Bottazzi, L., & Peri, G. (2007). The International Dynamics of R&D and Innovation in the Long Run and in The Short Run. The Economic Journal 117(518): 486-511.

Cohen, W. M., & Levinthal, D. A. (1990). Absorptive capacity: A new perspective on learning and innovation. Administrative Science Quarterly, 128-152.

Comin, D., & Mestieri, M. (2016). If technology has arrived everywhere, why has income diverged? Mimeo.

Comin, D., &Hobijn, B. (2009). The CHAT Dataset. Working Paper 15319

De Rassenfosse, G., Dernis, H., Guellec, D., Picci, L., & de la Potterie, B. V. P. (2013). The worldwide count of priority patents: A new indicator of inventive activity. Research Policy, 42(3), 720-737.

Engle, R. F., & Granger, C. W. (1987). Co-integration and error correction: representation, estimation, and testing. Econometrica: journal of the Econometric Society, 251-276.

Fisher, F. M. (1976). The Identification Problem in Econometrics. Robert E. Krieger Publishing Co., New York, second edition.

Frietsch, R., & Schmoch, U. (2010). Transnational patents and international markets. Scientometrics 82:185-200.

Kaiser, U., & Kuhn, J. M. (2012). Long-run effects of public–private research joint ventures: The case of the Danish Innovation Consortia support scheme. Research Policy 41(5): 913-927.

King, Mervyn, Enrique Sentana, and Sushil Wadhwani (1994), Volatility and Links between National Stock Markets, Econometrica 62, 901-933.

Lewbel, A. (2012). Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models. Journal of Business & Economic Statistics, 30(1), 67-80.

Mark, N. C., & Sul, D. (2003). Cointegration vector estimation by panel DOLS and long-run money demand. Oxford Bulletin of Economics and statistics 65.5, 655-680.

OECD (2009). OECD Patent Statistics Manual.

OECD, & Eurostat, 2005. Oslo Manual: Proposed guidelines for collecting and interpreting innovation data, 3rd edition. OECD, Paris

Pedroni, P. (2004). Panel cointegration: asymptotic and finite sample properties of pooled time series tests with an application to the PPP hypothesis. Econometric theory, 20(3): 597-625.

Robin, S., & Schubert, T. (2013). Cooperation with public research institutions and success in innovation: Evidence from France and Germany. Research Policy, 42, 149–166. doi:10.1016/j.respol.2012.06.002

Sentana, E., & Fiorentini, G. (2001). Identification, estimation and testing of conditionally heteroskedastic factor models. Journal of econometrics, 102(2), 143-164.

Van Looy, B., Du Plessis, M., & Magerman, T. (2006). Data production methods for harmonized patent statistics: Patentee sector allocation. KU Leuven FEB working paper MSI 0606.

Van Looy, B., Vereyen, C., & Schmoch, U. (2015). Patent Statistics: concordance IPC V8 – NACE rev.2 (version 2.0). Eurostat.

Westerlund, J. (2007). Testing for error correction in panel data. Oxford Bulletin of Economics and statistics, 69(6): 709-748.

Wooldridge, J. M. (2002). Econometric analysis of cross section and panel data. Cambridge, MA: MIT Press.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement No 727073

Wright, P. G. (1928). Tariff on animal and vegetable oils.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement No 727073