I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15708

# The Identification of Time-Invariant Variables in Panel Data Model: Exploring the Role of Science in Firms' Productivity

Sara Amoroso
Randolph Luca Bruno
Laura Magazzini

NOVEMBER 2022

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15708

# The Identification of Time-Invariant Variables in Panel Data Model: Exploring the Role of Science in Firms' Productivity

**Sara Amoroso**
*Joint Research Centre and European Commission*

**Laura Magazzini**
*Institute of Economics and EMbeDS and Sant'Anna School of Advanced Studies*

**Randolph Luca Bruno**
*University College London - SSEES, IZA Bonn and Rodolfo DeBenedetti Foundation*

NOVEMBER 2022



ISSN: 2365-9793

**IZA – Institute of Labor Economics**

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

# ABSTRACT

## The Identification of Time-Invariant Variables in Panel Data Model: Exploring the Role of Science in Firms' Productivity

Recent literature has raised the attention on the estimation of time-invariant variables both in a static and a dynmamic framework. In this context, Hausman-Taylor type estimators have been applied, relying crucially on the distinction between exogenous and endogenous variables (in terms of correlation with the time-invariant error component). We show that this provision can be relaxed, and identification can be achieved by relying on the milder assumption that the correlation between the individual effect and the time-varying regressors is homogenous over time. The methodology is applied to identify the role of inputs from "Science" (firm-level publications' stock) on firms' labour productivity, showing that the effect is larger for those firms with higher level of R&D investments. The results further support the dual – direct and indirect – role of R&D.

**JEL Classification:** C23, O32, L20

**Keywords:** panel data, time-invariant variables, science, productivity, R&D

**Corresponding author:**
Laura Magazzini
Institute of Economics and EMbeDS
Sant'Anna School of Advanced Studies
Piazza Martiri della Liberta 33
56127 Pisa
Italy

Email: laura.magazzini@santannapisa.it

# 1 Introduction

One of the main virtue of the availability of panel data is the possibility to control for unobserved heterogeneity at the unit level. The fixed effects (FE) model allows to identify the slope coefficient of time-variant variables even in case of omitted variables that are correlated with the regressors of interest, provided that these unobserved effects are constant over time. A notable drawback of this approach is that the specific effect of time-invariant variable —such as individuals' gender and race or firms' nationality— cannot be identified. However, the identification and estimation of the slope coefficients of such time-invariant variables are certainly of interest in applied research.

Different approaches have been proposed in the literature, allowing for estimation of the effect of time-invariant variables at the cost of introducing additional assumptions.

The pioneer work of Hausman and Taylor (1981, henceforth, HT81) allows to identify the slope coefficient of time-invariant variables imposing the assumption that some regressors (both time-variant and time-invariant) are uncorrelated with the individual component. Between variability of exogenous variables can be used to build legitimate instruments of the endogenous variables. Amemiya and MaCurdy (1986) and Breusch et al. (1989) extend the framework proposed by HT81 by suggesting additional (legitimate) instrumental variables. See Ahn and Schmidt (1995) and So Im et al. (1999) for an excellent review of these approaches. Baltagi and Bresson (2012) have proposed a robust version of the HT81 estimator in the presence of outliers.

Plümper and Troeger (2007) suggest an approach that is closely related to

HT81. They propose a multi-stage approach, labelled 'fixed effects vector decomposition', which relies on an exogeneity assumption of the time-invariant variables for the identification of their own effect (Greene, 2011). More recently, Pesaran and Zhou (2018) have proposed a two-stage approach, labelled 'fixed effect filtering', that can be applied when the time-invariant variables are exogenous. The author also consider the case of endogenous time-invariant variables; hence identification is achieved by relying on the existence of *external* instruments (Pesaran and Zhou, 2018; Chen et al., 2020).

The two-stage approach has also been applied in a dynamic framework, still the distinction between exogenous and endogenous (i.e. correlated with the individual effect) variable is crucial for identification of the effect of time-invariant variables, as well as the availability of external instruments (Kripfganz and Schwarz, 2019).

In this paper, we show that, relying on the assumption that the correlation between (a subset of) time-varying variables and the individual effect is homogenous over time (originally suggested by Breusch et al. 1989), it is possible to achieve identification of the effect of time-invariant variables without the need to choose exogenous variables or relying on external instruments. This result has been largely overlooked by the econometric literature and constitutes the main contribution of this paper.

We apply the methodology to estimate the effect of of study the relationship between a firm's publications' stock and its productivity. We aim to gauge whether and how much firm-level publications' stock (proxy for"Science") affects labour productivity. The results show that input from Science can be beneficial to those firms with higher level of R&D investments already in place, further supporting the dual role (direct and indirect) of R&D (Cohen and Levinthal, 1989; Leten

et al., 2021).

The paper proceeds as follows. The next section introduces the model and briefly reviews the existing literature studying the identification of the slope coefficient of time-invariant variables in a fixed effect framework. Section 3 presents the proposed methodology and the assumptions needed for the identification of the effect of time-invariant variables. Section 4 reports the results of Monte Carlo experiments. The empirical application is presented in Section 5 and Section 6 concludes.

# 2 Estimating the effect of time-invariant variables

Consider the linear panel data model with fixed effects ($i = 1, ..., N; t = 1, ..., T$):

$$y_{it} = \mathbf{x}'_{it}\beta + \mathbf{z}'_i\gamma + \varepsilon_{it} = \mathbf{x}'_{it}\beta + \mathbf{z}'_i\gamma + \alpha_i + \tau_t + e_{it} \tag{1}$$

in which $\mathbf{x}_{it}$ is a vector of $k$ time-varying independent variables, $\mathbf{z}_i$ contains observations on the $g$ time-invariant variables that vary across individuals but are constant over time; the error term $\varepsilon_{it}$ is decomposed into three sources of variation (Baltagi, 2021): an individual effect $\alpha_i$ that only varies across individuals and is constant over time for each $i$, a time component $\tau_t$ that only varies over time and is constant across individuals, capturing economy-wide effects, and $e_{it}$, an idiosyncratic error term. We consider the case of economic applications in which $N$ is "large" and $T$ is "small" (fixed), so that the time effect can be modelled by the inclusion of a set of time dummy variables $\tau_t$.

In a fixed effects framework, correlation is unrestricted between $\alpha_i$ and the independent variables, whereas the strict exogeneity assumption requires lack of correlation between $x_{is}$ and $e_{it}$ at any time period $(s, t = 1, ..., T)$. Despite the ability to control for unobservable individual characteristics, the fixed effects approach only allows identification of the parameter $\beta$, whereas the effect of specific time-invariant variables (the parameter $\gamma$) cannot be identified. The within-group transformation is used in the estimation process, in which group-demeaned variables are taken into account, that is for each variable in $\mathbf{x}$: $x_{j,it} - \bar{x}_{j,i}$ with $\bar{x}_{j,i} = \sum_t x_{j,it}/T$ $(j = 1, ..., k)$. In the case of the time-invariant variables $\mathbf{z}_i$, the within-group transformation would be identically equal to zero. Identification of $\gamma$ is not achieved, unless additional assumptions are imposed on model (1).

To the best of our knowledge, Hausman and Taylor (1981) was the first work to propose a methodology for the identification of the effect of time-invariant variables in model (1). Their approach requires to partition the vectors $\mathbf{x}_{it}$ and $\mathbf{z}_i$ into two sets of variables on the basis of the assumptions on the correlation between the variables and the individual effect $\alpha_i$:

$$\mathbf{x}_{it} = (\mathbf{x}'_{1it}, \mathbf{x}'_{2it})' \quad \text{and} \quad \mathbf{z}_i = (\mathbf{z}'_{1i}, \mathbf{z}'_{2i})'$$

with $\mathbf{x}_{1it}$ of dimension $k_1 \times 1$, $\mathbf{x}_{2it}$ of dimension $k_2 \times 1$, $\mathbf{z}_{1i}$ of dimension $g_1 \times 1$, and $\mathbf{z}_{2i}$ of dimension $g_2 \times 1$ ($k_1 + k_2 = k$ and $g_1 + g_2 = g$). The individual effect $\alpha_i$ is assumed to be uncorrelated with variables in $\mathbf{x}_{1it}$ and $\mathbf{z}_{1i}$; whereas any pattern of correlation is allowed with the variables in $\mathbf{x}_{2it}$ and $\mathbf{z}_{2i}$.[1] The model (1) can

[1] The assumption of strict exogeneity is maintained for all variables in $\mathbf{x}_{it}$ and $\mathbf{z}_i$.

therefore be written as:

$$y_{it} = \mathbf{x}'_{1it}\beta_1 + \mathbf{x}'_{2it}\beta_2 + \mathbf{z}'_{1i}\gamma_1 + \mathbf{z}'_{2i}\gamma_2 + \alpha_i + \tau_t + e_{it} \tag{2}$$

Within this framework, an instrumental variable approach can be adopted for estimating the parameters in (1) or, equivalently, (2). The assumption of lack of correlation between variables in $\mathbf{x}_{1it}$ and $\alpha_i$ allows using transformation of these variables as internal instruments for $\mathbf{z}_{2i}$. In particular, the set of instruments proposed by HT81 includes $\mathbf{x}_{1it} - \bar{\mathbf{x}}_{1i}$, $\mathbf{x}_{2it} - \bar{\mathbf{x}}_{2i}$, $\mathbf{x}_{1it}$ and $\mathbf{z}_{1i}$. A necessary condition for the identification of $\gamma = (\gamma'_1, \gamma'_2)'$ is that the number of variables in $\mathbf{x}_{1it}$ is at least as large as the number of variables in $\mathbf{z}_{2i}$ (Hausman and Taylor, 1981).

This approach has been extended by augmenting the set of legitimate instruments by Amemiya and MaCurdy (1986) and Breusch et al. (1989) (Ahn and Schmidt, 1995; So Im et al., 1999).[2] In order to increase efficiency, Amemiya and MaCurdy (1986) propose to replace the instrument $\mathbf{x}_{1it} - \bar{\mathbf{x}}_{1i}$ with the time-invariant variables

$$\mathbf{x}_{1i1} - \bar{\mathbf{x}}_{1i},\ \mathbf{x}_{1i2} - \bar{\mathbf{x}}_{1i},\ ...,\ \mathbf{x}_{1iT} - \bar{\mathbf{x}}_{1i}$$

Breusch et al. (1989, henceforth BMS89) propose to further expand the set of instruments by also using as instruments the (endogenous) time-invariant variables

$$\mathbf{x}_{2i1} - \bar{\mathbf{x}}_{2i},\ \mathbf{x}_{2i2} - \bar{\mathbf{x}}_{2i},\ ...,\ \mathbf{x}_{2iT} - \bar{\mathbf{x}}_{2i}$$

The validity of this approach would require the following additional assumption

[2] Baltagi and Bresson (2012) propose a robust version of the HT81 estimator.

(see also Ahn and Schmidt, 1995):

$$\text{For all } i,\ E(\mathbf{x}_{2it}\alpha_i) \text{ is the same for all } t \tag{3}$$

Note that this assumption is also required for the application of the system GMM estimator in dynamic panel data models (Blundell and Bond, 1998; Blundell et al., 2001).

Recently, some estimators have been proposed that relax the assumption on the exogeneity of $\mathbf{x}_{1it}$ but, on the other hand, they require either all the variables in $\mathbf{z}_i$ to be exogenous or, when endogeneity of $\mathbf{z}_i$ is allowed, they require the availability of *external* instrumental variables (Plümper and Troeger, 2007; Greene, 2011; Pesaran and Zhou, 2018; Chen et al., 2020), which may be problematic to retrive.

Plümper and Troeger (2007) propose a three-stage procedure that allows identification of the effect of time-invariant variables in the case they are orthogonal to the individual effects (Plümper and Troeger, 2007; Greene, 2011). More recently, Pesaran and Zhou (2018) have proposed a two-stage procedure that:

(i) in the first step, computes the fixed effects estimator of $\beta$ in model (1) and the associated residuals $\hat{\varepsilon}_{it}$;

(ii) in the second step, considers a regression of the time average of the residuals ($\bar{\hat{\varepsilon}}_i = \sum_t \hat{\varepsilon}_{it}/T$) on the time-invariant variables (including an intercept).

The regression in (ii) can be based on a OLS estimator when the exogeneity of all time-invariant variables is considered – in this case, the Pesaran and Zhou (2018) procedure is equivalent to the three-stage approach by Plümper and Troeger

(2007). The case of endogenous time-invariant variables can be taken into account, and in that case the regression in (ii) is based on an instrumental variable approach but the availability of valid *external* instrumental variables is required for the identification of $\gamma$.

The econometric literature has so far overlooked the implication of the BMS89 assumption (3) in a static framework. Indeed, homogenous-correlation variables (i.e. variables that satisfies condition in (3)) can be exploited to identify the parameter $\gamma$, without the need to rely on an exogeneity assumption, i.e. without the need to partition the variables in the model into an endogenous and exogenous group. We further elaborate on this insight in the next section.

# 3 Taking full advantage of BMS89 homogeneity assumption

Consider the model in (1) where, for simplicity, we focus on the case $k = 1$, $g = 1$ and $\tau_t = 0$, as the inclusion of a full set of time dummy variables in the model would not change the validity of our approach. The attention will be focused on taking into account the issue of correlation between the regressors ($x_{it}$, $z_i$) and the individual effect $\alpha_i$. Within a fixed effects framework, correlation is unrestricted between the variables on the right hand side and the individual fixed effect. The model is estimated by considering the within-group transformation, that is individual-demeaned variables are considered in the estimating equation:

$$y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)\beta + (z_i - z_i)\gamma + (\alpha_i - \alpha_i) + e_{it} - \bar{e}_i = (x_{it} - \bar{x}_i)\beta + e_{it} - \bar{e}_i$$

This approach allows consistent estimation even when arbitrary correlation with $\alpha_i$ is present, as the individual effect is removed from the estimating equation by the within-group transformation. However, the within-group transformation also remove the time-invariant variables $z_i$, so that the parameter $\gamma$ is not identified/identifiable.

As noticed by Verbeek (2008, pag. 354), the within-group estimator can also be obtained by an instrumental variable regression in which the within-group transformed $x_{it}$ (i.e., $x_{it} - \bar{x}_i$) is used as instrument in the level equation (1) (see also So Im et al. 1999). Still, only the parameter $\beta$ can be identified as there is no information (internal instrumental variable) for the estimation of $\gamma$.

Rooted in the HT81 approach, BMS89 noticed that, by assuming homogeneity over time in the correlation structure between time-varying (endogenous, i.e. correlated with $\alpha_i$) regressors and the individual component, further valid instruments for the level equation could be defined, and it is possible to write

$$E[(x_{is} - \bar{x}_i)(\alpha_i + e_{it})] = 0 \quad \text{for all } t \text{ and } s \tag{4}$$

The set of $T$ instruments only exploits the within variability of $x_{it}$, but by increasing the number of instruments also offers the possibility to identify $\gamma$, without the need to distinguish endogeous and exogenous regressors as in the original HT81 framework. What the literature has, so far, failed to recognize is that the additional moment conditions can be effectively employed for the identification of $\gamma$.

The estimation approach we propose in this paper builds on the BMS89 insight and suggest exploiting the $T-1$ additional instruments spanning from the assump-

tion of homogenous correlation in order to identify the effect of time-invariant variable(s):[3]

$$x_{i2} - \bar{x}_i, ..., x_{iT} - \bar{x}_i \tag{5}$$

Overall, the following set of moment conditions/instruments can be exploited in estimation:

$$\begin{aligned} E[(x_{it} - \bar{x}_i)(\alpha_i + e_{it})] &= E[(x_{it} - \bar{x}_i)(y_{it} - \beta x_{it} - \gamma z_i)] = 0 && (6) \\ E[(x_{i2} - \bar{x}_i)(\alpha_i + e_{it})] &= E[(x_{i2} - \bar{x}_i)(y_{it} - \beta x_{it} - \gamma z_i)] = 0 && (7) \\ &\vdots \\ E[(x_{iT} - \bar{x}_i)(\alpha_i + e_{it})] &= E[(x_{iT} - \bar{x}_i)(y_{it} - \beta x_{it} - \gamma z_i)] = 0 && (8) \end{aligned}$$

with condition (6) spanning from within-group/fixed-effect estimation, that allows estimation of $\beta$; and the additional $T-1$ moment conditions in (7)-(8), spanning from the homogeneity assumption (3) that can be exploited for the identification of $\gamma$.

Of course, besides exogeneity, the proposed set of instruments should also be "strong", such that weak instrument issues do not emerge in estimation. For that we require $z_i$ to be related to within-deviations of $x_{it}$, that is the within-evolution of $x_{it}$ should be related to $z_i$. To better understand this requirement, consider a linear combination of the full set of instruments:

$$x_{it} - x_{i1},\ x_{it} - x_{i2},\ ...,\ x_{it} - x_{iT} \tag{9}$$

[3] $x_{i1} - \bar{x}_i$ is omitted from the set of instrument to avoid multicollinearity in the full instrument set that also includes $x_{it} - \bar{x}_i$ as instrument for $x_{it}$.

for $t = 1, ..., T$.[4] All in all, correlation between $z_i$ and $x_{it}$ should be driven by correlation with its growth rate, and different evolution of $x_{it}$ is associated with different values of $z_i$.

Summing up, the following conditions are needed for a consistent estimation:

(A1) Homogeneity: For all $i$, $E(x_{it}\alpha_i)$ is the same for all $t$, as in (3)

(A2) When $k = g = 1$, $T \geq 2$

(A3) $z_i$ is related to the within evolution of $x_{it}$: $E(z_i(x_{it} - \bar{x}_i)) \neq 0$ (or, equivalently, $E(z_i(x_{it} - x_{is})) \neq 0$, $s, t = 1, ..., T$)

Condition (A1) is needed for consistent estimation of $\beta$ and $\gamma$: if the condition does not hold, the set of instrumental variables in (5) does not satisfy the exogeneity requirement. Note that this assumption is customarily exploited when estimating dynamic panel data model with endogenous or predetermined variables using the system-GMM estimator (Blundell and Bond, 1998; Blundell et al., 2001).

Condition (A2) is the order condition for identification. As we are estimating two parameters ($\beta$, $\gamma$) associated to two endogenous variables, we need at least two instrumental variables. In the more general case of $k > 1$ and $g > 1$, condition (A2) would be generalized to $(T - 1)k \geq g$. Condition (A1) needs not to hold for *all* the variables included in the model, it should nonetheless holds for a subset of variables, $k_{(A1)}$, satisfying condition (A1). In this case, condition (A2) would be written as $(T - 1)k_{(A1)} \geq g$.

Condition (A3) requires that instruments are relevant for $z$: the parameter $\gamma$ could only be identified when $z$ is correlated with the instruments in (5), that is,

[4] Note that the linear combination of the these instruments reproduces the within-group transformation as: $(x_{it} - x_{i1} + x_{it} - x_{i2} + ... + x_{it} - x_{iT})/T = x_{it} - \bar{x}_i$.

to the within evolution of $x$ (Stock and Yogo, 2005).

Standard theory of instrumental variables and GMM applies for estimation (Hansen, 1982). As an advantage, the proposed framework easily allows residual serial correlation and/or heteroskedasticity. Extension to unbalanced panel datasets is straightforward too.

The 'fixed effect filtering' (FEF) method proposed by Pesaran and Zhou (2018) can also be applied for estimation (Newey, 1984). In the first step, the fixed-effect estimator of $\beta$ is computed, and, relatedly, the fixed effect residuals. Then, the group mean of the fixed effect residuals is taken as the dependent variable and regressed on a constant term and the time-invariant variable $z_i$. This step is performed using a two-stage least squares estimator with variables in (5) used as instruments for $z_i$, therefore without reliance on external instruments.

# 4 Monte Carlo experiments

Our Monte Carlo set up is based on the experiments of Pesaran and Zhou (2018), henceforth referred to as PZ18. The data generating process is nonetheless simplified as we only generate one endogenous time-variying variable $x_{it}$ and one endogenous time-invariant variable $z_i$ $(i = 1, ..., N; t = 1, ..., T)$:

$$y_{it} = 1 + \alpha_i + \beta x_{it} + \gamma z_i + \varepsilon_{it} \tag{10}$$

with $\beta = \gamma = 1$. Focus will be on the estimation of $\gamma$. The fixed effects $\alpha_i$ are generated as $0.5(\chi^2_2 - 2)$ for $i = 1, ..., N$. As for $\varepsilon_{it}$, we follow PZ18, and consider three different designs:

- Homoskedastic $\varepsilon_{it}$:

$$\varepsilon_{it} \sim IIDN(0,1), i = 1, ..., N; t = 1, ..., T$$

- Heteroskedastic $\varepsilon_{it}$:

$$\varepsilon_{it} \sim IIDN(0, \sigma_i^2), i = 1, ..., N; t = 1, ..., T$$

  with $\sigma_i^2 \sim 0.5(1 + 0.5\chi_2^2)$ for all $i$.

- Serially correlated and heteroskedastic $\varepsilon_{it}$:

$$\varepsilon_{it} = \rho_{\varepsilon,i}\varepsilon_{it-1} + \sqrt{1 - \rho_{\varepsilon,i}^2}u_{it}$$

  for $t = -49, -48, ..., T$ (the first value of $\varepsilon_{it}$ is set to 0 for all $i$), $u_{it} \sim IIDN(0, \sigma_{ui}^2)$ for all $i$ and $t$, $\sigma_{ui}^2 \sim 0.5(1+0.5IID\chi_2^2)$, and $\rho_{\varepsilon,i} \sim IIDU(0, 0.98)$ for all $i$. We discard the first 50 observations, using the remaining $T$ observations in the experiments.

We consider $N = 100, 500$ and $T = 4, 8$. In all experiments, $x_{it}$ is correlated with $\alpha_i$, but uncorrelated with $\varepsilon_{is}$ at all time periods ($s = 1, ..., T$) so that the strict exogeneity assumption is satisfied and the fixed-effect estimation (within-group transformation) provides consistent estimation of $\beta$. In particular, $x_{it}$ is generated as:

$$x_{it} = \alpha_i g_t + w_{it}$$

with the time effects $g_t$ generated as $U(1,2)$ and kept fixed across all replications. For assumption (A1) to hold, we set $g_t = g_1$ for all $t$.

We follow PZ18 in generating the stochastic component of $x_{it}$, $w_{it}$, as an heterogenous stationary AR(1) process:

$$w_{it} = \mu_i(1-\rho_{w,i}) + \rho_{w,i} w_{it-1} + \sqrt{1-\rho_{w,i}^2}\epsilon_{it}$$

with $\epsilon_{it} \sim IIDN(0, \sigma_{\epsilon,i}^2)$ for all $i$, $\sigma_{\epsilon,i}^2 \sim 0.5(1 + 0.5IID\chi_2^2)$, $w_{i0} \sim IIDN(\mu_i, \sigma_{\epsilon,i}^2)$, $\rho_{w,i} \sim IIDU(0, 0.98)$, $\mu_i \sim IIDN(0,2)$.

As for the time-invariant variable $z_i$, PZ18 consider the following data generating process:

$$z_i = 1 + \bar{w}_i + \alpha_i + \xi_i \tag{11}$$

that we change to

$$z_i = 1 + (w_{iT} - w_{i1}) + \alpha_i + \xi_i \tag{12}$$

with $\xi_i \sim IIDN(0,1)$. When $z_i$ is generated according to (12), condition (A3) is satisfied, i.e. within-variability of $x_{it}$ is relevant for $z_i$, and the proposed methodology allows identification of $\gamma$. We will also consider equation (11) to evaluate the effect of irrelevant instruments in the estimation process. In all experiments, $z_i$ is endogenous, i.e. correlated with $\alpha_i$.

No external instrument (denoted as $r_i$ in PZ18) is generated, and identification of $\gamma$ can be achieved by exploiting the homogeneity assumption (A1) and the correlation of $z_i$ and the growth rate ("within evolution") of $x_{it}$.

Results of Monte Carlo experiments, based on 1000 replications, are in Tables 1-3, that report Monte Carlo mean and standard deviation of the FEF-IV estimator, and joint GMM estimation (first and second step) of $\beta$ and $\gamma$ (intercept is estimated but not reported). In order to compute FEF-IV estimates, the fixed-effect estimator is first obtained. Residuals of the within-group estimation are then computed and their group-average is considered as the dependent variable in the second stage. This entails an IV estimation with dependent variable $z_i$ (and the intercept) and the BMS variables (5) used as instruments.[5] The tables also report, in the GMM framework, the 5% rejection rate of the Hansen $J$ test of overidentifying restriction,[6] and the Kleibergen-Paap (KP) $F$ test for assessing instrument strength.

Table 1 is our baseline estimation framework, in which all the assumptions needed for the application of the proposed methodology are satisfied, i.e. correlation between $\alpha_i$ and $x_{it}$ is homogenous over time (i.e. A1 satisfied: when generating $x_{it}$, $g_t = g_1$ for all $t$), and $z_i$ is related to within deviations of $x_{it}$ (A3 satisfied, eq. 12 used to generate $z_i$). As $T > 2$ in all experiments, and we consider the case of one endogenous time-varying variable $x_{it}$ and one endogenous time-invariant variable $z_i$, condition (A2) is always satisfied.

Results in Table 1 shows that identification of $\gamma$ is achieved, and similar results emerge with IV and GMM joint estimation as compared to the FEF-IV estimation (Pesaran and Zhou, 2018). The $J$-test of overidentified restrictions has generally the correct size (being oversized in the smaller sample of $N = 100$). The KP $F$

[5]Estimates are obtained using Stata. FEF-IV estimates are obtained using the $xtfef$ command available at https://qiankunzhou.weebly.com/research.html. IV and GMM estimators are obtained with the $ivreg2$ command (Baum et al., 2010).

[6]Being the number of replication 1,000, the 95% confidence interval for the test size at that 5% level is 3.65%-6.35% (Morris et al., 2019).

test is large, pointing to instruments' strenght.

*** TABLE 1 ABOUT HERE ***

In Table 2, $x_{it}$ is generated such that the homogeneity condition (A1) is not satisfied, i.e. the parameter $g_t$ varies over time. In this case, FEF-IV has the advantage of providing a consistent estimator of $\beta$, whereas the estimate of $\gamma$ is biased. However, the $J$-test of overidentifying restriction has power in detecting departures from the homogeneity assumption.

*** TABLE 2 ABOUT HERE ***

Finally, Table 3 reports the result in the case when (A3) is not satisfied, i.e. $z_i$ is generated according to (11): within-evolution of $x_{it}$ is not dependent upon $z_i$, so that the BMS instruments are irrelevant. Again, a consistent estimate of $\beta$ is obtained with FEF-IV, and the estimate of $\gamma$ is biased. In this context, the KP statistics is small, pointing to a problem related to instruments' strength in this setting.

*** TABLE 3 ABOUT HERE ***

# 5 Science and productivity at the firm level: an application

The methodology presented in the previous sections is particularly suitable for empirical applications in which one or more independent variables are invariant over time, and the traditional FE model does not allow for the identification of

such variables. Precious information on the impact of such fixed components could be lost in a catch-all black box. There are many empirical models in which such occurrences appear: for example in longitudinal household survey containing individual as unit where the gender/race are hidden in the FE; firm level panels where the location/sector effect cannot be identified; macro-panels where fixed effects capture geographical or time invariant features (such as membership to alliances or blocks, language, legal origin, etc.). Examples could be extended in diverse empirical contexts.

We exploit the proposed methodology to gauge whether and how much publications' stock at the firm level (proxy for "Science") might affect its labour productivity. The data for the analysis have been compiled by matching firm level data from three different data sources: R&D Scoreboard, Orbis (Bureau van Dijk), and the JRC-OCED COR&DIP database (v.2, 2019).

Our baseline regression builds on the model by Hall and Mairesse (1995):

$$\begin{aligned} \ln\left(\frac{VA_{it}}{L_{it}}\right) &= \beta_1 \ln\left(\frac{K_{it}}{L_{it}}\right) + \beta_2 \ln\left(\frac{RD_{it}}{L_{it}}\right) + \beta_3 \ln(L_{it}) + \\ &+ \beta_4 S_i + \beta_5 \ln\left(\frac{RD_{it}}{L_{it}}\right) \times S_i + \alpha_i + \tau_t + e_{it} \end{aligned} \tag{13}$$

where the dependent variable is firm's $i$ labour productivity at time $t$, defined as the (natural log) ratio of value added, $VA_{it}$, over employees, $L_{it}$. The regressors include physical capital stock, $K_{it}$, the stock of R&D expenditure $R\&D_{it}$ (both over employees). The perpetual inventory method has been applied for computation of the stock.[7] and the (natural log) number of employees, $L_{it}$. These variables

[7]For physical capital we follow Gal (2013), whereas in the case of R&D we consider a 5%

are extracted from Orbis ($VA$, $K$) and the R&D Scoreboard ($RD$, $L$) and are available over the period 2007-2016.

The baseline regression is augmented by including a proxy of the firm's engagement in basic research via links with the scientific community $S_i$, as well as its interaction with R&D stock. In order to measure firms' linkages to the scientific community, we rely on the count of a firm' publications (Cockburn and Henderson, 2000; Leten et al., 2021). The number of publications is drawn from the COR&DIP dataset and it is only available over the time period 2014-2016. In order to compute $S_i$, we pool the total number of publication at the firm level over the available time span (stock). There is a clear sectoral pattern in the average number of publications, from 'Aerospace & Defence' having on average 774.8 publications per firm over the observed time span to 'Industrials' with 188.1. To mitigate this erratic pattern, $S_i$ is defined as a firm-specific dummy variable equal to 1 to identify firms with "above-sector-average" number of publications. This variable is time-invariant over the observed time span, and we extend its value over the whole time period analysed.[8] We argue that the propensity of the firm to publish to be an "open science" structural characteristic of the firm unlikely to change substantially over time. In other words, a firm could be changing the overall number of publications, but it is unlikely to change its relative position above or below the sector-average. We motivate the use of such proxy based on the extensive literature analysing the role of firms' interaction with "Science" and science effect on innovation (see e.g., Mansfield, 1991, 1998, Cockburn and Henderson,

pre-sample growth rate and a rate of obsolescence of knowledge equal to 15%.

[8]The sectors considered are from the R&D Scoreboard data: Aerospace & Defence, Automobiles & other transport, Chemicals, Health industries, ICT producers, ICT services, Industrials, and a residual sector Others.

2000, Leten et al., 2021).

Descriptive statistics of the variables included in the equation are reported in Table 4. Results of the fixed-effect (FE) estimation of (13) are reported in Table 5. Column (1) reports the baseline estimation as in Hall and Mairesse (1995) where there is no assumption of constant return to scale (the coefficient of $\ln(L)$ is not statistically different from zero, though).

*** TABLE 4 ABOUT HERE ***

*** TABLE 5 ABOUT HERE ***

Column (2) includes $S$ and its interaction term with R&D stock (being time-invariant, the coefficient of $S$, $\beta_4$, cannot be identified). Column (3)-(5) add the interaction between sector dummies and time FE, as well as the interaction between regional and time FE. Finally, column (6) imposes constant return to scale by omitting $\ln(L)$. Tangible capital and R&D stocks have significant coefficients of an order of magnitude similar to other studies.

As the failure of the strict exogeneity assumption may be a concern in this setting, the bottom of Table 5 includes the $p$-value of a simple test of strict exogeneity based on the augmented regression that also include the forward values of $\ln(K/L)$ , $\ln(RD/L)$, and $\ln(L)$ (when included among the regressors; see Wooldridge, 2002). The reported $p$-value is related to the null hypothesis that the coefficients of the variables in $t+1$ are equal to zero.

The regression coefficient $\beta_4$ is not estimated with the traditional FE approach as it would be treated just like the unobserved time-invariant individual effect $\alpha_i$, whereas the interaction between $S_i$ and R&D stock can be identified. The interaction term is positive, so that firms that have "above-sector-average" publications

(i.e., a pro-publication behaviour) have higher returns from R&D expenditures in terms of larger productivity.

It would be extremely important to identify $\beta_4$ in this context. It can be argued that closer interaction with the Science domain can be beneficial to firms' productivity, by gathering quicker access to and better understanding of new scientific knowledge (Griliches, 1986; Rosenberg, 1990; Leten et al., 2021). However, different values of $\beta_4$ would have different implications. On the one side, were $\beta_4$ positive, pro-publication behaviour would be always beneficial to firms' productivity. On the contrary, a negative $\beta_4$ would provide support to the idea that R&D expenditures play a dual role within the firm: at low level of R&D the generation of scientific knowledge does not accrue benefits to firms, which fail to *absorb* and exploit external knowledge (Cohen and Levinthal, 1989). If we are unable to estimate $\beta_4$, we are not addressing such important dimension on the role of science in firms' productivity. That is the reason we move to our proposed methodology.

In Table 6, the model is estimated using the methodology proposed in this paper, allowing the estimation of the effect of the time-invariant variable $S$. Both the two-step version of the proposed procedure (FEF-IV) and joint estimation using GMM are considered.[9] The number of observations is reduced because we omitted some sector-region with few observations that were causing problems for computation.

We explore the results obtained by changing the variable treated as homogeneous, i.e. the variable(s) assumed to satisfy (A1). Evidence of weak instruments, violating assumption (A3), emerges when $\ln(K/L)$, $\ln(L)$, and $\ln(RD/L)$ .[10] In

[9]Time FE, and its interaction with sector and region FE are treated as exogenous variables when the model is estimated using GMM, as customary.

[10]Full table of results is available from the authors upon request.

these model, the value of the KP $F$ statistic is extremely low. As our preferred specification, reported in Table 6, both $\ln(RD/L)$ and the interaction term between $S$ and $\ln(RD/L)$ are treated as homogeneous. In this case, the weak instrument problem is mitigated: the KP-$F$ turns into 17.70, and 34.87 when constant return to scale are imposed. As for the validity of the homogeneity assumption, the Hansen-$J$ statistics does not allow rejecting the null of instrument validity. The coefficients of the time-invariant variables are coherent with FE estimation, and the interaction term $\ln(RD/L) \times S$ is positive, pointing to the fact that the effect of $S$ is increasing with increasing R&D expenditure. The direct effect of $S$ is negative and statistically significant, though.

*** TABLE 6 ABOUT HERE ***

In order to better interpret the results, Figure 1 shows the predicted productivity, $\ln(VA/L)$, as a function of $\ln(RD/L)$ in the two cases $S = 1$ and $S = 0$. The first, second and third quartile of the distribution of $\ln(RD/L)$ are also reported in the graph. Coherently with the interpretation of the coefficients of the variables involved, there seems to be an advantage in being linked to science for those firms that also exhibits larger R&D expenditures, otherwise the effect is negative. Results further provide support to the view that previous knowledge (absorptive capacity, Cohen and Levinthal, 1989) is required to better exploit inputs from basic science, so that publication stock at the firm level should be complemented by investments in 'absorptive capacity' (Rosenberg, 1990; Cohen and Levinthal, 1989).

*** FIGURE 1 ABOUT HERE ***

Finally, Table 7 reports some robustness checks, that use alternative measures for the variable $S$. Concerns may arise as the total number of publications may be linked to the size of the firm. As we are including top R&D spenders worldwide, we claim that this is a minor concern; however, we also compute the variable $S$ by taking into account the "above-average" number of publications weighted by the number of employees (Model 11, 12), and by R&D expenditures (Model 13, 14). The variable $\ln(L)$ is omitted from the equation because never statistically different from zero, and better statistics related to instruments' strength are obtained (i.e., larger KP $F$). When the number of publications is weighted by the number of employees, the sign of the effects is broadly coherent but no longer statistically significant. Weighting the number of publications by R&D expenditures, on the contrary, does confirm the results of our main specification.

*** TABLE 7 ABOUT HERE ***

# 6 Discussion

This paper exploits an homogeneity assumption in a fixed effects framework to achieve identification of the slope of time-invariant variables. The assumption is not without content, and it would not be necessary for identification of the effect of time-variant variables using the fixed effect approach. If satisfied, the additional moment conditions spanning from this assumption allow to identify the effects of time-invariant variables, that is unfeasible in the fixed effects framework. To the best of our knowledge, our proposal is the first to suggest to exploit this assumption in a static framework, without requiring the distinction between variables that are related/unrelated to the individual effect. However, the assumption is analogous to

the mean stationarity assumption customarily exploited in a dynamic framework.

Monte Carlo experiments show that the proposed approach can allow to estimate the coefficients of time-invariant variables. Estimation can rely on the FEF-IV estimator proposed by Pesaran and Zhou (2018) or by jointly estimating all parameters via IV or GMM.

The proposed methodology is applied to the identification of the role of scientific knowledge at the level of the firm on its productivity. Due to data constraints, the main variable of interest (firm's publications stock) is invariant over time; however, we argue that this way of treating the variable can be supported by the fact that this firm characteristic is likely to be substantially stable over time. The results show that scientific knowledge at the firm level can be beneficial in terms of productivity if associated to larger R&D expenditure, supporting the dual role (direct and indirect) of R&D at the firm level (Cohen and Levinthal, 1989).

## Acknowledgements

We gratefully acknowledge comments and suggestions from conference participants to the $9^{th}$ Italian Congress of Econometrics and Empirical Economics (virtual, January 2021), the $26^{th}$ International Panel Data Conference (virtual, July 2021), and the European Winter Meeting of the Econometric Society 2021 (virtual, December 2021).

The views expressed are purely those of the authors and may not in any circumstances be regarded as stating an official position of the European Commission.

# References


Ahn, S. C. and P. Schmidt (1995). Efficient estimation of models for dynamic panel data. *Journal of econometrics*, 68(1):5–27.

Amemiya, T. and T. E. MaCurdy (1986). Instrumental-variable estimation of an error-components model. *Econometrica: Journal of the Econometric Society*, pages 869–880.

Baltagi, B. H. (2021). *Econometric analysis of panel data*. Springer Nature.

Baltagi, B. H. and G. Bresson (2012). A robust hausman–taylor estimator. In *Essays in Honor of Jerry Hausman*. Emerald Group Publishing Limited.

Baum, C., M. Schaffer, and S. Stillman (2010). ivreg2: Stata module for extended instrumental variables/2sls, gmm and ac/hac, liml and k-class regression.

Blundell, R. and S. Bond (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of econometrics*, 87(1):115–143.

Blundell, R., S. Bond, and F. Windmeijer (2001). Estimation in dynamic panel data models: improving on the performance of the standard gmm estimator. In *Non-stationary Panels, Panel Cointegration, and Dynamic Panels*, Advances in Econometrics. Emerald Group Publishing Limited.

Breusch, T. S., G. E. Mizon, and P. Schmidt (1989). Efficient estimation using panel data. *Econometrica*, 57(3):695–700. ISSN 00129682, 14680262.

Chen, J., R. Yue, and J. Wu (2020). Testing for individual and time effects in the two-way error component model with time-invariant regressors. *Economic Modelling*, 92(C):216–229.

Cockburn, I. M. and R. M. Henderson (2000). Publicly funded science and the productivity of the pharmaceutical industry. *Innovation policy and the economy*, 1:1–34.

Cohen, W. M. and D. A. Levinthal (1989). Innovation and learning: the two faces of r & d. *The economic journal*, 99(397):569–596.

Gal, P. N. (2013). Measuring total factor productivity at the firm level using oecd-orbis. *OECD Economics Department Working Papers*, No. 1049.

Greene, W. (2011). Fixed effects vector decomposition: A magical solution to the problem of time-invariant variables in fixed effects models? *Political Analysis*, 19(2):135–146. ISSN 10471987, 14764989.

Griliches, Z. (1986). Productivity, r and d, and basic research at the firm level in the 1970's. *The American Economic Review*, 76(1):141–154. ISSN 00028282.

Hall, B. H. and J. Mairesse (1995). Exploring the relationship between r&d and productivity in french manufacturing firms. *Journal of econometrics*, 65(1):263–293.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054. ISSN 00129682, 14680262.

Hausman, J. A. and W. E. Taylor (1981). Panel data and unobservable individual effects. *Econometrica*, 49(6):1377–1398. ISSN 00129682, 14680262.

Kripfganz, S. and C. Schwarz (2019). Estimation of linear dynamic panel data models with time-invariant regressors. *Journal of Applied Econometrics*, 34(4):526–546.

Leten, B., S. Kelchtermans, and R. Belderbos (2021). How does basic research improve innovation performance in the world's major pharmaceutical firms? *Industry and Innovation*, online:1–29.

Mansfield, E. (1991). Academic research and industrial innovation. *Research policy*, 20(1):1–12.

Mansfield, E. (1998). Academic research and industrial innovation: An update of empirical findings. *Research policy*, 26(7-8):773–776.

Morris, T. P., I. R. White, and M. J. Crowther (2019). Using simulation studies to evaluate statistical methods. *Statistics in medicine*, 38(11):2074–2102.

Newey, W. K. (1984). A method of moments interpretation of sequential estimators. *Economics Letters*, 14(2-3):201–206.

Pesaran, M. H. and Q. Zhou (2018). Estimation of time-invariant effects in static panel data models. *Econometric Reviews*, 37(10):1137–1171.

Plümper, T. and V. E. Troeger (2007). Efficient estimation of time-invariant and rarely changing variables in finite sample panel analyses with unit fixed effects. *Political Analysis*, 15(2):124–139.

Rosenberg, N. (1990). Why do firms do basic research (with their own money)? *Research Policy*, 19(2):165–174.

So Im, K., S. Ahn, P. Schmidt, and J. Wooldridge (1999). Efficient estimation of panel data models with strictly exogenous explanatory variables. *Journal of Econometrics*, 93(1):177–201.

Stock, J. and M. Yogo (2005). *Testing for Weak Instruments in Linear IV Regression*, pages 80–108. Cambridge University Press, New York.

Verbeek, M. (2008). *A guide to modern econometrics*. John Wiley & Sons.

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

Table 1: Results of Monte Carlo experiments: mean of $\hat{\beta}$ and $\hat{\gamma}$, standard deviation in parenthesis. DGP with (A1) homogenous $x_{it}$; and (A3) $z_i$ related to within deviation of $x_{it}$ – order condition (A2) satisfied in all experiments.

| | | $T = 4$ | | | $T = 8$ | | |
|---|---|---|---|---|---|---|---|
| | | FEF-IV | GMM1 | GMM2 | FEF-IV | GMM1 | GMM2 |
| | | Homoskedastic $\varepsilon_{it}$ | | | | | |
| $N = 100$ | $\hat{\beta}$ | 1.004 | 1.043 | 1.019 | 1.002 | 1.082 | 1.027 |
| | | (0.075) | (0.086) | (0.077) | (0.044) | (0.067) | (0.048) |
| | $\hat{\gamma}$ | 1.015 | 1.015 | 1.014 | 1.031 | 1.028 | 1.026 |
| | | (0.091) | (0.088) | (0.089) | (0.081) | (0.074) | (0.076) |
| | 5% rej $J$ | | | 5.9% | | | 9.1% |
| | avg. KP | | | 1022.5 | | | 452.8 |
| $N = 500$ | $\hat{\beta}$ | 0.999 | 1.008 | 1.001 | 1.001 | 1.023 | 1.005 |
| | | (0.033) | (0.036) | (0.034) | (0.020) | (0.026) | (0.020) |
| | $\hat{\gamma}$ | 1.001 | 1.001 | 1.001 | 1.007 | 1.006 | 1.007 |
| | | (0.040) | (0.040) | (0.040) | (0.036) | (0.035) | (0.036) |
| | 5% rej $J$ | | | 5.0% | | | 6.7% |
| | avg. KP | | | 437.3 | | | 247.2 |
| | | Heteroskedastic (uncorrelated) $\varepsilon_{it}$ | | | | | |
| $N = 100$ | $\hat{\beta}$ | 1.003 | 1.044 | 1.018 | 1.003 | 1.080 | 1.027 |
| | | (0.076) | (0.089) | (0.079) | (0.045) | (0.070) | (0.050) |
| | $\hat{\gamma}$ | 1.005 | 1.005 | 1.007 | 1.027 | 1.024 | 1.022 |
| | | (0.091) | (0.088) | (0.089) | (0.078) | (0.072) | (0.073) |
| | 5% rej $J$ | | | 6.4% | | | 9.3% |
| | avg. KP | | | 1109.4 | | | 209.0 |
| $N = 500$ | $\hat{\beta}$ | 1.002 | 1.011 | 1.004 | 1.000 | 1.022 | 1.004 |
| | | (0.034) | (0.036) | (0.034) | (0.020) | (0.025) | (0.021) |
| | $\hat{\gamma}$ | 1.001 | 1.001 | 1.002 | 1.004 | 1.004 | 1.004 |
| | | (0.039) | (0.039) | (0.039) | (0.035) | (0.034) | (0.034) |
| | 5% rej $J$ | | | 4.2% | | | 6.0% |
| | avg. KP | | | 432.4 | | | 244.0 |
| | | Heteroskedasticity and serial correlation in $\varepsilon_{it}$ | | | | | |
| $N = 100$ | $\hat{\beta}$ | 0.998 | 1.039 | 1.008 | 1.002 | 1.083 | 1.023 |
| | | (0.061) | (0.081) | (0.063) | (0.044) | (0.073) | (0.048) |
| | $\hat{\gamma}$ | 1.009 | 1.009 | 1.009 | 1.028 | 1.025 | 1.023 |
| | | (0.104) | (0.100) | (0.103) | (0.083) | (0.077) | (0.082) |
| | 5% rej $J$ | | | 6.4% | | | 7.9% |
| | avg. KP | | | 902.1 | | | 273.1 |
| $N = 500$ | $\hat{\beta}$ | 1.000 | 1.010 | 1.001 | 1.001 | 1.023 | 1.003 |
| | | (0.027) | (0.032) | (0.027) | (0.020) | (0.026) | (0.020) |
| | $\hat{\gamma}$ | 1.004 | 1.004 | 1.004 | 1.004 | 1.004 | 1.004 |
| | | (0.045) | (0.044) | (0.045) | (0.039) | (0.038) | (0.039) |
| | 5% rej $J$ | | | 6.0% | | | 5.4% |
| | avg. KP | | | 442.0 | | | 244.4 |

Table 2: Results of Monte Carlo experiments: mean of $\hat{\beta}$ and $\hat{\gamma}$, standard deviation in parenthesis. DGP with non-homogenous $x_{it}$: (A1) not satisfied.

| | | $T=4$ | | | $T=8$ | | |
|---|---|---|---|---|---|---|---|
| | | FEF-IV | GMM1 | GMM2 | FEF-IV | GMM1 | GMM2 |
| | | Homoskedastic $\varepsilon_{it}$ | | | | | |
| $N=100$ | $\hat{\beta}$ | 1.003 | 1.307 | 1.199 | 1.002 | 1.267 | 1.146 |
| | | (0.072) | (0.101) | (0.095) | (0.044) | (0.086) | (0.071) |
| | $\hat{\gamma}$ | 1.061 | 1.072 | 1.130 | 1.148 | 1.089 | 1.094 |
| | | (0.074) | (0.083) | (0.110) | (0.092) | (0.062) | (0.072) |
| | 5% rej $J$ | | | 94.3% | | | 92.9% |
| | avg. KP | | | 28630 | | | 7564850 |
| $N=500$ | $\hat{\beta}$ | 0.998 | 1.331 | 1.195 | 1.001 | 1.281 | 1.124 |
| | | (0.032) | (0.044) | (0.046) | (0.020) | (0.041) | (0.034) |
| | $\hat{\gamma}$ | 1.117 | 1.053 | 1.063 | 1.130 | 1.079 | 1.082 |
| | | (0.049) | (0.034) | (0.037) | (0.041) | (0.028) | (0.034) |
| | 5% rej $J$ | | | 100% | | | 100% |
| | avg. KP | | | 773.8 | | | 660.3 |
| | | Heteroskedastic (uncorrelated) $\varepsilon_{it}$ | | | | | |
| $N=100$ | $\hat{\beta}$ | 1.003 | 1.310 | 1.196 | 1.003 | 1.262 | 1.141 |
| | | (0.072) | (0.099) | (0.097) | (0.044) | (0.088) | (0.073) |
| | $\hat{\gamma}$ | 1.120 | 1.053 | 1.067 | 1.139 | 1.084 | 1.089 |
| | | (0.108) | (0.075) | (0.082) | (0.086) | (0.060) | (0.067) |
| | 5% rej $J$ | | | 94.0% | | | 92.9% |
| | avg. KP | | | 49813 | | | 6298.8 |
| $N=500$ | $\hat{\beta}$ | 1.002 | 1.333 | 1.198 | 1.000 | 1.281 | 1.123 |
| | | (0.032) | (0.044) | (0.044) | (0.020) | (0.042) | (0.034) |
| | $\hat{\gamma}$ | 1.117 | 1.053 | 1.063 | 1.128 | 1.077 | 1.079 |
| | | (0.048) | (0.033) | (0.038) | (0.041) | (0.027) | (0.033) |
| | 5% rej $J$ | | | 100% | | | 100% |
| | avg. KP | | | 8837.0 | | | 670.6 |
| | | Heteroskedasticity and serial correlation in $\varepsilon_{it}$ | | | | | |
| $N=100$ | $\hat{\beta}$ | 0.997 | 1.306 | 1.146 | 1.001 | 1.266 | 1.123 |
| | | (0.057) | (0.102) | (0.089) | (0.043) | (0.090) | (0.071) |
| | $\hat{\gamma}$ | 1.122 | 1.058 | 1.080 | 1.144 | 1.087 | 1.100 |
| | | (0.117) | (0.087) | (0.098) | (0.091) | (0.068) | (0.076) |
| | 5% rej $J$ | | | 92.1% | | | 85.7% |
| | avg. KP | | | 41355 | | | 283630 |
| $N=500$ | $\hat{\beta}$ | 1.000 | 1.331 | 1.144 | 1.001 | 1.281 | 1.101 |
| | | (0.025) | (0.047) | (0.043) | (0.019) | (0.045) | (0.033) |
| | $\hat{\gamma}$ | 1.117 | 1.054 | 1.076 | 1.128 | 1.077 | 1.088 |
| | | (0.052) | (0.038) | (0.045) | (0.042) | (0.031) | (0.038) |
| | 5% rej $J$ | | | 100% | | | 100% |
| | avg. KP | | | 791.4 | | | 676.2 |

Table 3: Results of Monte Carlo experiments: mean of $\hat{\beta}$ and $\hat{\gamma}$, standard deviation in parenthesis. DGP with lack of association between $z_i$ and within deviation in $x_{it}$: condition (A3) not satisfied.

| | | $T=4$ | | | $T=8$ | | |
|---|---|---|---|---|---|---|---|
| | | FEF-IV | GMM1 | GMM2 | FEF-IV | GMM1 | GMM2 |
| | | Homoskedastic $\varepsilon_{it}$ | | | | | |
| $N=100$ | $\hat{\beta}$ | 1.004 | 1.019 | 1.008 | 1.002 | 1.040 | 1.011 |
| | | (0.075) | (0.078) | (0.076) | (0.044) | (0.055) | (0.046) |
| | $\hat{\gamma}$ | 1.207 | 1.210 | 1.221 | 1.233 | 1.201 | 1.211 |
| | | (0.437) | (0.452) | (0.441) | (0.212) | (0.209) | (0.218) |
| | 5% rej $J$ | | | 3.1% | | | 2.7% |
| | avg. KP | | | 0.949 | | | 1.251 |
| $N=500$ | $\hat{\beta}$ | 0.999 | 1.002 | 0.999 | 1.001 | 1.009 | 1.002 |
| | | (0.033) | (0.034) | (0.033) | (0.020) | (0.021) | (0.020) |
| | $\hat{\gamma}$ | 1.135 | 1.132 | 1.134 | 1.207 | 1.200 | 1.200 |
| | | (0.378) | (0.378) | (0.384) | (0.195) | (0.194) | (0.197) |
| | 5% rej $J$ | | | 4.0% | | | 3.9% |
| | avg. KP | | | 1.074 | | | 1.083 |
| | | Heteroskedastic (uncorrelated) $\varepsilon_{it}$ | | | | | |
| $N=100$ | $\hat{\beta}$ | 1.003 | 1.021 | 1.008 | 1.003 | 1.039 | 1.012 |
| | | (0.076) | (0.080) | (0.078) | (0.045) | (0.056) | (0.047) |
| | $\hat{\gamma}$ | 1.198 | 1.183 | 1.193 | 1.221 | 1.190 | 1.199 |
| | | (0.448) | (0.443) | (0.461) | (0.213) | (0.209) | (0.218) |
| | 5% rej $J$ | | | 2.5% | | | 3.7% |
| | avg. KP | | | 0.951 | | | 1.234 |
| $N=500$ | $\hat{\beta}$ | 1.002 | 1.005 | 1.002 | 1.000 | 1.009 | 1.002 |
| | | (0.034) | (0.034) | (0.034) | (0.020) | (0.022) | (0.021) |
| | $\hat{\gamma}$ | 1.160 | 1.157 | 1.157 | 1.192 | 1.185 | 1.188 |
| | | (0.470) | (0.468) | (0.478) | (0.206) | (0.205) | (0.204) |
| | 5% rej $J$ | | | 3.4% | | | 3.7% |
| | avg. KP | | | 1.041 | | | 1.072 |
| | | Heteroskedasticity and serial correlation in $\varepsilon_{it}$ | | | | | |
| $N=100$ | $\hat{\beta}$ | 0.998 | 1.015 | 1.001 | 1.002 | 1.040 | 1.010 |
| | | (0.061) | (0.068) | (0.062) | (0.044) | (0.056) | (0.046) |
| | $\hat{\gamma}$ | 1.208 | 1.193 | 1.202 | 1.217 | 1.184 | 1.198 |
| | | (0.492) | (0.489) | (0.520) | (0.233) | (0.231) | (0.245) |
| | 5% rej $J$ | | | 2.3% | | | 3.3% |
| | avg. KP | | | 0.962 | | | 1.260 |
| $N=500$ | $\hat{\beta}$ | 1.000 | 1.003 | 1.000 | 1.001 | 1.009 | 1.002 |
| | | (0.027) | (0.028) | (0.027) | (0.020) | (0.021) | (0.020) |
| | $\hat{\gamma}$ | 1.153 | 1.150 | 1.150 | 1.197 | 1.190 | 1.192 |
| | | (0.453) | (0.452) | (0.459) | (0.221) | (0.220) | (0.220) |
| | 5% rej $J$ | | | 2.8% | | | 2.7% |
| | avg. KP | | | 1.005 | | | 1.086 |

Table 4: Descriptive statistics of the variables included in the regression ($N = 6,681$)

| Variable | mean | std. dev. | min. | max. |
|---|---|---|---|---|
| $\ln(VA/L)$ | 10.66 | 1.913 | -0.397 | 19.51 |
| $\ln(K/L)$ | 10.34 | 2.324 | -9.711 | 21.244 |
| $\ln(RD/L)$ | 10.75 | 1.465 | 4.785 | 17.15 |
| $\ln(L)$ | 9.219 | 1.625 | 1.386 | 13.35 |
| $S$ | 0.286 | 0.452 | 0 | 1 |

Table 5: Fixed effects estimator of the productivity-Science relationship.

| Variable | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| $\ln(K/L)$ | 0.578*** | 0.579*** | 0.579*** | 0.570*** | 0.570*** | 0.570*** |
| | (0.030) | (0.030) | (0.029) | (0.030) | (0.030) | (0.030) |
| $\ln(RD/L)$ | 0.334*** | 0.265*** | 0.285*** | 0.231** | 0.247** | 0.288*** |
| | (0.101) | (0.101) | (0.102) | (0.108) | (0.110) | (0.085) |
| $\ln(L)$ | -0.027 | -0.022 | -0.039 | -0.059 | -0.076 | – |
| | (0.092) | (0.091) | (0.09) | (0.101) | (0.102) | |
| $S$ | | (omitted) | (omitted) | (omitted) | (omitted) | (omitted) |
| $\ln(RD/L) \times S$ | | 0.350* | 0.361* | 0.363* | 0.369** | 0.373** |
| | | (0.198) | (0.195) | (0.191) | (0.188) | (0.186) |
| Firm FE | yes | yes | yes | yes | yes | yes |
| Time FE | yes | yes | yes | yes | yes | yes |
| Sector×Year FE | no | no | yes | no | yes | yes |
| Region×Year FE | no | no | no | yes | yes | yes |
| $N$ | 6,681 | 6,681 | 6,681 | 6,681 | 6,681 | 6,681 |
| $p$ strict exo. | 0.874 | 0.859 | 0.923 | 0.832 | 0.863 | 0.663 |

Table 6: IV estimate of the effect of the time-invariant variable $S_i$; homogenous variables: $\ln(RD/L)$ and $\ln(RD/L) \times S$.

| Variable | (7) FEF-IV | (8) GMM | (9) FEF-IV | (10) GMM |
|---|---|---|---|---|
| $\ln(K/L)$ | 0.572*** | 0.564*** | 0.572*** | 0.564*** |
| | (0.029) | (0.028) | (0.029) | (0.028) |
| $\ln(RD/L)$ | 0.229** | 0.276*** | 0.270*** | 0.306*** |
| | | (0.085) | (0.087) | (0.069) |
| $\ln(L)$ | -0.076 | -0.060 | | |
| | (0.102) | (0.076) | | |
| $S$ | -3.354* | -2.312** | -3.378* | -2.440** |
| | (1.902) | (1.137) | (1.874) | (1.121) |
| $\ln(RD/L) \times S$ | 0.355* | 0.229** | 0.359* | 0.236** |
| | (0.186) | (0.109) | (0.185) | (0.109) |
| Firm FE | yes | yes | yes | yes |
| Time FE | yes | yes | yes | yes |
| Sector×Year FE | yes | yes | yes | yes |
| Region×Year FE | yes | yes | yes | yes |
| $N$ | 6467 | 6467 | 6467 | 6467 |
| Hansen-$J$ | | 9.168 | | 9.403 |
| $J$ $p$-value | | 0.935 | | 0.923 |
| KP $F$ | | 17.70 | | 34.87 |

Table 7: Alternative mesures for $S$ based on the ratio between the number of publications and, respectively, the number of employees (Publ/Emp) and R&D expenditures (Publ./RD); IV estimate of the effect of the time-invariant variable $S_i$; homogenous variables: $\ln(RD/L)$ and $\ln(RD/L) \times S$.

| Variable | (11) FEF-IV | (12) GMM | (13) FEF-IV | (14) GMM |
|---|---|---|---|---|
| Measure for $S$ | Publ/Emp | Publ/Emp | Publ/RD | Publ/RD |
| $\ln(K/L)$ | 0.572*** | 0.572*** | 0.571*** | 0.573*** |
| | (0.029) | (0.028) | (0.029) | (0.028) |
| $\ln(RD/L)$ | 0.272*** | 0.289*** | 0.196*** | 0.291*** |
| | (0.090) | (0.076) | (0.076) | (0.062) |
| $S$ | -2.670 | -0.807 | -3.705** | -1.961** |
| | (1.701) | (1.052) | (1.493) | (1.701) |
| $\ln(RD/L) \times S$ | 0.259 | 0.082 | 0.398*** | 0.216** |
| | (0.163) | (0.101) | (0.154) | (0.098) |
| Firm FE | yes | yes | yes | yes |
| Time FE | yes | yes | yes | yes |
| Sector×Year FE | yes | yes | yes | yes |
| Region×Year FE | yes | yes | yes | yes |
| $N$ | 6467 | 6467 | 6467 | 6467 |
| Hansen-$J$ | | 14.79 | | 12.77 |
| $J$ $p$-value | | 0.611 | | 0.752 |
| KP $F$ | | 34.86 | | 41.17 |

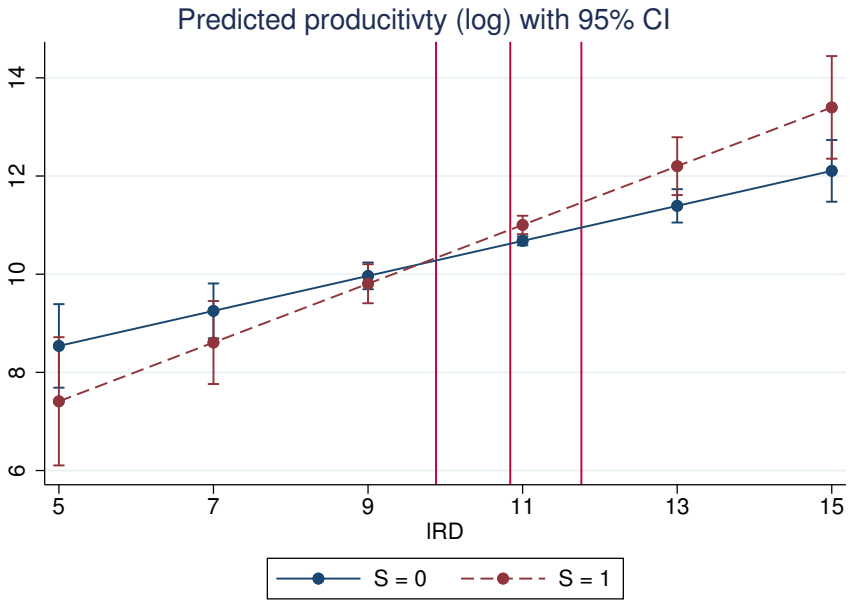


Figure 1: Predicted productivity (from Model 10 in Table 6) as a function of R&D, $S = 0$ and $S = 1$; vertical lines identify Q25, median, and Q75 in the distribution of $\ln(RD/L)$