

Limitation of Firm Fixed Effects Models and the Missing R&D-Patent Relation: New Methods and Evidence

Hui-Ching Chuang National Taipei University
Po-Hsuan Hsu National Tsing Hua University
Chung-Ming Kuan National Taiwan University
Jui-Chung Yang National Taiwan University

SFS Cavalcade Asia-Pacific
December 15, 2024

The prevailing use of firm fixed effects

- A common practice in economics, finance, and accounting studies is to include **firm fixed effects in regression models**.
- Researchers choose fixed effects (FE) regressions and claim that such absorb the influences of individual-specific, unobservable, and time-invariant effects – which is the advantage.
- In this paper, we argue that, without theoretical modeling or appropriate econometric designs, such prevailing use of fixed effects may have disadvantage –failing to identify the effect of the persistent variable of interest.
- The R&D-patent relation had been extensively examined by Hausman et al. (1984), Griliches (1990), etc.
 - It is perhaps the most intuitive relation in economics: **more R&D input, more patent output**

Findings in the literature

- Several studies document the FE model **weakens/eliminates** the explanatory power of persistent variables (such as R&D):
 - Aghion, Reenen, and Zingales (2013, AER) “*In the specifications where we include fixed effects, the coefficient on the R&D stock **falls significantly.***”
 - Balsmeier, Fleming, and Manso (2017, JFE) “*Alternative regressions with R&D investments scaled by total assets reveal a significant positive effect only in specifications without firm fixed effects. Inclusion of controls for time-invariant firm heterogeneity leads to statistically **insignificant** results.*”

Example: Luong et al. (2017, JFQA)

TABLE 2
Baseline Regressions

Table 2 reports the regressions of firm innovation on institutional ownership. Columns 1 and 2 (3 and 4) show the pooled ordinary least squares (OLS) (Firm fixed effects) regression results. The dependent variable is shown as the column heading in columns 1–4. The main independent variable is foreign institutional ownership (FIO). All explanatory variables are lagged by 1 year. Variable definitions are in Appendix B. Standard errors are clustered at the firm level and reported in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Variables	ln(PATENT)	ln(CITEPAT)	ln(PATENT)	ln(CITEPAT)
	1	2	3	4
FIO	0.010*** (0.003)	0.014*** (0.004)	0.008*** (0.003)	0.011*** (0.004)
DIO	-0.010*** (0.002)	-0.012*** (0.003)	-0.001 (0.001)	-0.001 (0.002)
INSIDE	-0.072 (0.063)	-0.054 (0.070)	0.062* (0.032)	0.084* (0.044)
ln(AGE)	0.062** (0.029)	0.062* (0.032)	0.086** (0.037)	0.118** (0.049)
HHI	0.396 (0.292)	0.400 (0.326)	-0.170 (0.274)	0.152 (0.347)
HHI ²	-0.277 (0.277)	-0.280 (0.307)	0.152 (0.241)	0.050 (0.290)
RD	2.267*** (0.238)	2.637*** (0.288)	0.054 (0.132)	-0.232 (0.215)
CAPEX	2.313*** (0.269)	2.913*** (0.308)	0.378*** (0.134)	0.519*** (0.184)
PPE	-0.232** (0.116)	-0.192 (0.129)	-0.082 (0.086)	-0.095 (0.114)
LEV	-0.365*** (0.097)	-0.453*** (0.104)	-0.132** (0.057)	-0.185*** (0.070)
ROA	-0.616*** (0.144)	-0.850*** (0.165)	-0.037 (0.079)	-0.169 (0.108)
Year fixed effects	Yes	Yes	Yes	Yes
Firm fixed effects	No	No	Yes	Yes
Industry fixed effects	Yes	Yes	No	No
Country fixed effects	Yes	Yes	No	No

Dependent variables: innovation output

ln(PATENT): Natural logarithm of the number of patents filed by each firm in a year plus 1.

ln(CITEPAT): Natural logarithm of the number of citations received by each firm's patents in a year plus 1.

Our focus:

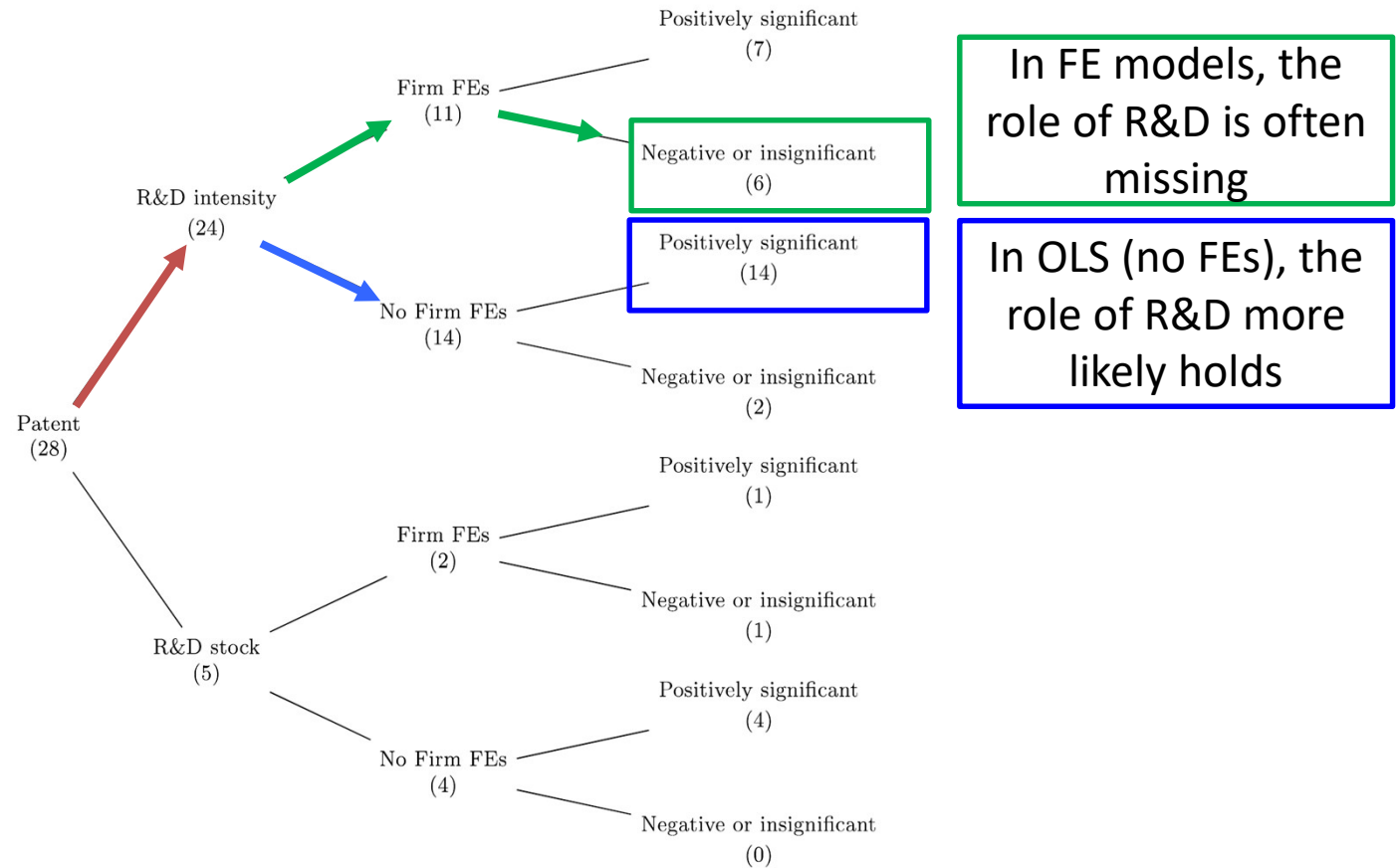
RD: Research and development expenditures scaled by total assets.

How come R&D does not explain patents?

Our survey of the corporate innovation literature

- We start our screening of corporate innovation papers from the **200+** papers based on the survey papers of Ederer and Manso (2011), He and Tian (2018), Lerner and Seru (2022) and authors' reading list.
- **36** papers include the patent/citation-R&D equations at least in one column of their tables. They are published on: JFE(10), JFQA(8), MS(7), RFS(4), JF(2), AER, Econometrica, JAE, JLE, and REStat from 2007 to 2021.
- **28** specify the dependent variable as patent count and use linear regression model to investigate the patent/citation-R&D relationship.
- **6** use Poisson (negative binominal) model to investigate the patent/citation-R&D relationship.
- Others use hierarchical linear model, change to change, Tobit, and Fama-Macbeth approach.

Our survey of the corporate innovation literature: Least square approach on Patent (paper #)



Our survey of the corporate innovation literature

- Our literature review suggests a surprising and puzzling pattern.
 - 40% to 50% estimates for R&D in the literature show insignificant or even negative coefficients on R&D input.
- **OLS** allows us to understand R&D's explanatory power for **total variations of patents** (= cross-sectional/between-firm variations + time-series/within-firm variations)
- **FE** models absorb all cross-sectional/between-firm variations in patents
 - An analogy: a **high (low) tech** firm's R&D and patents are **persistently high (low)**. Thus, cross-sectional variation could be **more important** than time-series variation (Hausman et al., 1984; Hall et al., 2005).
 - However, FE models eliminate all cross-sectional variations in firms' patents – **so R&D role is missing**
 - So, the estimation results of FE models only tell us R&D's explanatory power for a firm's time-series variations in patents

Issues with FE model results

- If R&D input **cannot** explain patent output in a fixed effect regression model, then to what extent can we **trust prior estimation results** for the explanatory power of those new factors?
 - Instead, it may only capture the time-series variation of patent output, which is relatively small when compared with the cross-sectional variation (Hausman et al., 1984; Hall et al., 2005).
 - In addition, statistical inferences based on firm fixed effects regressions may be **dominated by firms** that are featured with larger within-firm (time-series) variation (deHaan, 2021).
- **These issues also apply to many persistent variables in corporate finance, such as ownership structure, culture, executive talents, etc.**

Other persistent explanatory variables

- A surge of studies on **new factors** that can explain corporate innovation performance.
 - Those new factors include managerial style, compensation design, institution ownership, board structure, accounting standard, banking policy and others (Ederer and Manso (2011), He and Tian (2018), and Lerner and Seru (2022))
- These models often use firm fixed effects regressions.
- However, if these new variables are also persistent, just like R&D, then they likely correlate with firm unobservable heterogeneities.
 - Coles, Daniel, and Naveen (2006, JFE) *“One possible reason for slightly **weaker** results **using firm fixed effects** is that the relation between firm investment policy and **vega** is strong in the cross-section but not very prominent in the time series.”*
- A critical issue: do those new factors have sufficient explanatory power (both cross-section and time series)? Those factors may only explain (limited) time-series variation but **not the full picture**

Econometric Tools

- A lack of appropriate econometric tools to address the issue for more reliable statistical inferences.
- Not to include firm fixed effects (Baltagi et al., 2000; Hall et al., 2005; Noel and Schankerman, 2013; Pesaran and Zhou, 2018) may introduce alternative biases.
- **Our propositions and contributions:**
 1. Adjusted Hausman and Taylor (“adj-HT” 1981) method
 2. Machine learning
 - Post-Regularization LASSO (PRL)
 - Double-machine learning (DML)

Overview: OLS, FE, HT, PRL, and DML

$$Innov_{i,t+1} = \beta_0 + \beta_{R\&D} R\&D_{i,t} + \beta_2 X_{i,t} + \sum_{s \in \mathcal{S}} \alpha_s \text{dummy}_{s,i} + u_{i,t}$$

- OLS includes none of the firm dummies, i.e., $\mathcal{S} = \emptyset$.
- FE includes all of the firm dummies, i.e., $\mathcal{S} = \{1, \dots, N\}$.
- Adjusted HT uses the **demeaned** $X_{i,t}$ and **demeaned** R&D to construct the moment conditions in GMM estimation for $\beta_{R\&D}$
- PRL and DML **select** some of the firm dummies, i.e., $\mathcal{S} \in \{1, \dots, N\}$ while keep the valid inference of $\beta_{R\&D}$.
 - Intuitively, since important dummies have been selected to control for, **we prevent the omitted-variable bias**.
 - On the other hand, since unimportant dummies are not selected, **we have better power in identifying the role of persistent R&D**.

Our proposition-1

Adjusted Hausman-Taylor methods

- Consider the simplified HT (1981, Econometrica) model

$$Y_{i,t+1} = \beta Z_i + \beta_2 X_{i,t} + \alpha_i + \epsilon_{i,t}.$$

- HT allow arbitrary correlation between Z_i and α_i , and use moment conditions:

$$E[(X_{i,t} - \bar{X}_i)'(Y_{i,t+1} - \beta Z_i - \beta_2 X_{i,t})] = \mathbf{0}.$$

$$E[X_{i,t}'(Y_{i,t+1} - \beta Z_i - \beta_2 X_{i,t})] = \mathbf{0}.$$

- Treat rarely time-varying $R\&D$ as Z_i , and add an extra moment condition:

- The correlation between firm fixed effects (FEs) and R&D mainly arises from the firm's population-level R&D

- Deviations from this level are exogenous to the FEs.

$$E[(R\&D_{i,t} - \overline{R\&D}_i)(Y_{i,t+1} - \beta_{R\&D} R\&D_{i,t} - \beta_2 X_{i,t})] = 0.$$

- Thus, similar to HT, we can identify $\beta_{R\&D}$ by GMM, using $(X_{i,t} - \bar{X}_i)$, $X_{i,t}$, and $(R\&D_{i,t} - \overline{R\&D}_i)$ to construct the moment conditions.

Sample

- We first collect the financial and accounting data of all publicly-listed firms in the U.S. from CRSP and Compustat.
- We exclude financial and utility firms (SIC in 6000-6999, and 4900-4999), and firms with negative and missing total asset and sales.
- We then collect the patent and citation data of all public firms from the PatentsView patent database that is organized by the USPTO.
- As a result, we have 86,341 firm-year observations during 1976-2000. (We also consider sample of firms with at least one patent during the sample period.)

Our baseline regressions

$$Innov_{i,t+1} = \beta_0 + \beta_{R\&D} R\&D_{i,t} + \beta_2 X_{i,t} + \sum_{s \in S} \alpha_s dummy_{s,i} + u_{i,t}$$

- $Innov_{i,t+1}$ is one of innovation measures: $\ln(1+Patent)$, $\ln(1+Citation)$, and $\ln(1+AdjCitation)$.
- $R\&D_{i,t}$ is the past five years R&D expenditures divide by total asset. We also consider R&D/ME, or $\ln(1+R\&D)$ for five years for robustness.
- $X_{i,t}$ denotes firm characteristic controls: *R&D missing dummy, capital level, $\ln(1+Firm\ age)$, $\ln(K/L)$, Tobin's Q, ROA, leverage, cash divide by the total asset, Institutional ownership ratio, KZ index, Herfindahl-Hirschman index, and Herfindahl-Hirschman index square.*
- We will discuss the **Poisson** regression later.

OLS and Fixed Effects

	OLS (Year Dummies Only)	Fixed Effects (All Firm and Year Dummies)
Patent regression		
R&D/Assets	0.593***	0.041
	(0.042)	(0.028)
Citation regression		
R&D/Assets	1.396***	-0.051
	(0.084)	(0.068)
AdjCitation regression		
R&D/Assets	0.590***	0.033
	(0.045)	(0.031)

Firm cluster standard errors in parentheses. *p<0.1, **p<0.05, and ***p<0.01.
We suppress the year FEs and firm characteristics variables to save space.

Adjusted HT, OLS and Fixed Effects

	OLS (Year Dummies Only)	adjHT (Year Dummies Only)	Fixed Effects (All Firm and Year Dummies)
Patent regression			
R&D/Assets	0.593***	0.220***	0.041
	(0.042)	(0.027)	(0.028)
Citation regression			
R&D/Assets	1.396***	0.551***	-0.051
	(0.084)	(0.065)	(0.068)
AdjCitation regression			
R&D/Assets	0.590***	0.231***	0.033
	(0.045)	(0.030)	(0.031)

Firm cluster standard errors in parentheses. *p<0.1, **p<0.05, and ***p<0.01.
We suppress the year FEs and firm characteristics variables to save space.

OLS, Adjusted HT, and Fixed Effects

- Our OLS regression results suggest a robust pattern that R&D input positively explains patent output when regression models do not include firm fixed effects.
- The magnitude of this positive relation is severely weakened when firm fixed effects are included in regressions.
- Our adjusted HT GMM results suggest a robust pattern that R&D input positively explains patent output.
- The magnitude of this positive relation is slightly weakened when compared to the OLS results while still significant positive (such as 0.593 => 0.220)

Our proposition-2

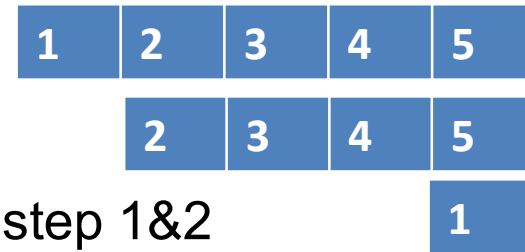
- Unobserved heterogeneity exists **in some firms but not others**.
 - Some managers are aggressive in investing in R&D and pursuing patent output, but others are not.
 - Some firms have a strong, innovation-oriented culture, while others do not.
- A smarter methodology that can **select** which individual firm dummies to be included is called for.
- In this paper, we proposed the second advanced machine learning method:
 1. Post-regularization LASSO (**PRL**, Chernozhukov et al., 2015)
 2. Double machine learning (**DML**, Chernozhukov et al., 2018)
 - to select individual firm dummies (and explanatory variables) in explaining firm-level patent outputs.

Post-Regularization LASSO (PRL)

- PRL proceeds in the following 3 steps:
 - **Step1:** LASSO of $Innov_{i,t+1}$ on firm dummies and force small coefficients of some dummies to 0. (estimate step) Then, Post LASSO: OLS of $Innov_{i,t+1}$ on selected firm dummies, obtain the residuals, \hat{r}_y . (get residual step)
 - **Step2:**
 - a) LASSO of $R\&D_{i,t}$ on firm dummies and force small coefficients of some dummies to 0. Then, Post LASSO: OLS of $R\&D_{i,t}$ on selected firm dummies, obtain the residuals, $\hat{r}_{R\&D}$.
 - b) LASSO of $X_{i,t}$ on firm dummies and force small coefficients of some dummies to 0. Then, Post LASSO: OLS of $X_{i,t}$ on selected firm dummies, obtain the residuals, \hat{r}_X .
 - **Step3:** OLS of \hat{r}_y on $\hat{r}_{R\&D}$, \hat{r}_X and obtain the coefficient $\hat{\beta}_{R\&D, PRL}$.
- If a firm dummy is selected in either Step 1 or Step 2 (partialing-out/residualizing), it is informative to $Innov_{i,t+1}$ and $R\&D_{i,t}$.

Double Machine Learning (DML)

- Chernozhukov et al. (2018) propose the DML which generalizes the PRL to a general model selection (LASSO, random forests, gradient boosting, neural nets, etc.) and add the cross-fitting procedures to PRL.
- DML proceeds in the following steps:
 - splits sample into random K folds,
 - use leave- k -out sample in the estimate step 1&2
 - use the k th-fold sample to obtain the residuals for Y and R&D
 - stake all K folds residuals, use OLS to obtain $\hat{\beta}_{R\&D,DML}$.
- DML uses sample splitting to eliminates the dependence between the estimation steps, reduce the **post-model-selection bias (or, errors in estimated variables)** of PRL. However, as the cross-fit procedure reduces the sample size, DML also reduces the estimation efficiency.
- Yang, Chuang, and Kuan (2020, JoEcts) use DML to examine the Big N audit quality effect in the accounting literature.



Patent regression: PRL and DML results

PRL

	OLS (Year Dummies Only)	PRL (Firm and Year Dummies)	Fixed Effects (All Firm and Year Dummies)
R&D/Asset	0.593***	0.199***	0.041
	(0.042)	(0.018)	(0.028)
Number of dummies		11,570	
Number of selected dummies		1,241 (10.73%)	

DML

	OLS (Year Dummies Only)	DML (Firm and Year Dummies)	Fixed Effects (All Firm and Year Dummies)
R&D/Asset	0.593***	0.213***	0.041
	(0.042)	(0.014)	(0.028)
Number of dummies		1,1570	
Number of selected dummies		1,737 (15.01%)	

Citation regression: PRL and DML results

PRL

	OLS (Year Dummies Only)	PRL (Firm and Year Dummies)	Fixed Effects (All Firm and Year Dummies)
R&D/Asset	1.396***	1.397***	-0.051
	(0.084)	(0.083)	(0.068)
Number of dummies		11,570	
Number of selected dummies		525 (4.54%)	

DML

	OLS (Year Dummies Only)	DML (Firm and Year Dummies)	Fixed Effects (All Firm and Year Dummies)
R&D/Asset	1.396***	1.364***	-0.051
	(0.084)	(0.050)	(0.068)
Number of dummies		11,570	
Number of selected dummies		947 (8.18%)	

Adjusted-Citation regression: PRL and DML results

PRL

	OLS (Year Dummies Only)	PRL (Firm and Year Dummies)	Fixed Effects (All Firm and Year Dummies)
R&D/Asset	0.590***	0.210***	0.033
	(0.045)	(0.019)	(0.031)
Number of dummies		11,570	
Number of selected dummies		1,194 (10.32%)	

DML

	OLS (Year Dummies Only)	DML (Firm and Year Dummies)	Fixed Effects (All Firm and Year Dummies)
R&D/Asset	0.590***	0.198***	0.033
	(0.045)	(0.015)	(0.031)
Number of dummies		11,570	
Number of selected dummies		1,882 (16.27%)	

PRL and DML results

- The coefficients on R&D input are statistically significant and that their economic magnitude is much closer to those from OLS models without firm fixed effects (than those with firm fixed effects).
- PRL and DML select about **10% to 20% of firm dummies** to be included in regression models -- the bias from adding all firm dummies **overpowers** the bias from not adding any at all (the consequence is an insignificant R&D coefficient)
- These results, together with prior analyses, suggest that **most firm dummies** do not play a crucial role.
- To recap: FE model = **0.041 (insig.)**
- OLS = 0.593,
- adj-HT = 0.220,
- PRL = 0.199, DML = 0.213

STATA code

- To implement adjusted Hausman and Taylor:

```
ivregress gmm y z x (z = demean_z demean_x),  
          wmatrix(cluster firmID)
```

- To implement PRL

```
poregress y z x, controls(i.firmID)  
          vce(cluster firmID)
```

- To implement DML

```
xporegress y z x, controls(i.firmID)  
          vce(cluster firmID) xfolds(#folds)
```

Hui-Ching Chuang, Po-Hsuan Hsu, Chung-Ming Kuan, Jui-Chung Yang (2024). *Limitation of Firm Fixed Effects Models and the Missing R&D-Patent Relation: New Methods and Evidence*.

Robustness

- Alternative R&D measures
 - Tested R&D/ME, and $\ln(1+R\&D)$ in addition to R&D/AT,
- Patenting firms
 - Excluded firms without any patent for during its sample period.
- Handling missing R&D values
 - Remove firm-year observations with missing R&D
- Alternative specifications in HT, PRL, and DML methods
 - Different fold count from two to five in DML method

Poisson regression

$$E(\text{Innov}_{i,t+1} | R\&D_{i,t}, x_{i,t}) = \exp(\beta_0 + \beta_{R\&D} R\&D_{i,t} + \beta_2 X_{i,t} + \sum_{s \in \mathcal{S}} \alpha_s \text{dummy}_{s,i})$$

- Poisson regression includes none of the firm dummies, i.e., $\mathcal{S} = \emptyset$.
- Poisson fixed effect regression includes all of the firm dummies, i.e., $\mathcal{S} = \{1, \dots, N\}$.
- Adjusted Hausman-Taylor uses demeaned $X_{i,t}$ and demeaned $R\&D_{i,t}$ in GMM to identify $\beta_{R\&D}$ of the rarely time-varying R&D.
- PRL Poisson (Belloni, Chernozhukov and Wei, 2016, JBES) and DML select some of the firm dummies, i.e., $\mathcal{S} \in \{1, \dots, N\}$.
 - PRL Poisson proceeds in the similar fashion as PRL, except it uses the post LASSO Poisson regression in Step 1 and use GMM in Step 3.
- DML follows the PRL Poisson steps with cross-fitting.

Adjusted Hausman-Taylor and Poisson regression

Patent Counts

	Poisson (Year Dummies only)	FE Poisson (All Firm and Year Dummies)	adjHT (Year Dummies only)
R&D/Asset	2.305***	-0.248	1.679***
	(0.187)	(0.255)	(0.215)

Citation Counts

	Poisson (Year Dummies)	FE Poisson (Firm and Year Dummies)	adjHT (Year Dummies)
R&D/Asset	2.094***	-0.125	1.400***
	(0.219)	(0.252)	(0.229)

Firm cluster standard errors in parentheses. *p<0.1, **p<0.05, and ***p<0.01.
We suppress the year FEs and firm characteristics variables to save space.

PRL Poisson and DML

Patent Counts

	Poisson	PRL Poisson	DML
	Year Dummies only	Firm and Year Dummies	Firm and Year Dummies
R&D/Asset	2.305***	2.407***	2.312***
	(0.187)	(0.161)	(0.120)
Number of dummies		11,570	11,570
Number of selected dummies		1,218 (10.53%)	2,484 (21.47%)

Citation Counts

	Poisson	PRL Poisson	DML
	Year Dummies only	Firm and Year Dummies	Firm and Year Dummies
R&D/Asset	2.094***	2.299***	2.262***
	(0.219)	(0.256)	(0.113)
Number of dummies		11,570	11,570
Number of selected dummies		1,226 (10.51%)	2,426 (20.97%)

Poisson regression results

- Poisson regressions are common for count-based dependent variables in econ / finance literature (Cohn et al., 2022).
- When we implement adjusted HT, PRL Poisson, and DML, we again find that the estimated coefficients on R&D input are significantly positive **and closer to** those from Poisson regressions without firm fixed effects (than those with firm fixed effects).
- Like our results in OLS regressions, the existence of firm fixed effects also prevents Poisson models from delivering a positive R&D-patent relation. The bias that we attempt to highlight and solve in this paper is a general issue across different econometric models.

Our recommendations

1. Instead of only reporting regressions with firm fixed effects, please also present the results **without** firm fixed effects and discuss why the coefficient estimates vary
2. The results from no firm FEs can serve as a “**second opinion**” for the effect you would like to examine
3. If the results from 1 and 2 go the opposite directions. Consider our adjusted Hausman and Taylor, PRL and DML methods as “**third opinion**”.
 - easy to implement by **STATA** (or R/Python). We make our codes available online:
 - ✓ <https://github.com/hcchuang/Limitation-of-Firm-Fixed-Effects-Models-and-the-Missing-R-D-Patent-Relation>
 - ✓ handle omitted variable issues without strict assumptions
 - ✓ enable researchers to decide exactly which firm dummies should be added in regressions.

Our contributions

- Corporate finance studies tend to solve firm-specific, time-invariant unobservables issues by using fixed effects models (e.g., Angrist and Pischke, 2009; Imbens and Wooldridge, 2009; Roberts and Whited, 2013)
- We illustrate the potential **biases** of such a practice by using the intuitive R&D-patent relation as our lab.
 - More importantly, we offer two **feasible and ready-to-use** methodologies to enable corporate finance researchers to analyze the effects of economic variables that are persistent in time, such as ownership structure and managerial capability.
 - In particular, we provide explanations that they may use to **justify** their choice of regression specifications without firm fixed effects (or with only a limited set of firm dummies).

Our contributions (Cont.)

- We add to modern machine learning techniques in corporate finance research, for the selection of relevant covariates (e.g., Chinco et al., 2019; Feng et al., 2020; Gu et al., 2020; Erel et al., 2021).
- This study also adds to the economics literature by supporting and justifying prior studies' choice of not including firm fixed effects to estimate knowledge production functions (Pakes and Griliches, 1984; Blundell et al., 1995; Hall et al., 2007; Noel and Schankerman, 2013).

Thank you!

Questions? Comments?

hcchuang@gm.ntpu.edu.tw (Hui-Ching Chuang)

Chuang, Hui-Ching and Hsu, Po-Hsuan and Kuan, Chung-Ming and Yang, Jui-Chung, Limitation of Firm Fixed Effects Models and the Missing R&D-Patent Relation: New Methods and Evidence (2024).

Available at SSRN: <https://ssrn.com/abstract=4636846>

<https://github.com/hcchuang/Limitation-of-Firm-Fixed-Effects-Models-and-the-Missing-R-D-Patent-Relation>