中国制造业企业的研发创新:基本事实、常见误区与合适计量方法讨论

龙小宁　　林志帆

内容提要:当前,学术界对研发创新的关注持续升温,本文对研究中最常用的四个数据来源(工业企业数据库、上市公司数据、世界银行企业调查数据、私营企业调查数据)中制造业企业的研发创新变量进行分析,为相关研究提供基础信息支持与计量方法参考。主要发现包括:1. 中国制造业企业研发创新的发展态势较为乐观,常规年度工业企业数据库中研发支出与新产品产出为正的企业比例仅约 10% 的"创新荒漠"很可能是统计纰漏造成的假象。跨国比较揭示,中国企业的研发创新领先于其他新兴发展中国家;2. 近年,进行研发创新的制造业上市公司比例高于 80%、投入强度持续上升,专利活动实现了数量与质量的双重提升;3. 当服从离散、混合、截断分布的研发创新数据作为被解释变量时,OLS 估计不一致,适用方法应为 Probit 与 Logit 模型、Tobit 模型、Truncation 模型与 MLE 估计。

关键词:研发创新;工业企业数据库;上市公司数据;世界银行中国企业调查数据;中国私营企业调查数据;计量模型选择

DOI:10.19365/j.issn1000-4181.2018.02.10

一、引言

技术进步是经济发展的长期动力。在中国经济行至增长模式切换的"十字路口"之际,学术界对研发创新的关注持续升温,学者们多以研发支出、新产品产值、专利的申请与授权等可量化的研发创新指标为被解释变量,探索财税要素、市场竞争、政策的激励与扭曲、知识产权保护、FDI 技术溢出等因素的影响;同时,许多关注中国经济发展的非专业研究者也借助各种新闻载体与自媒体表述了他们关于企业研发创新的看法。这些资料为洞悉中国企业研发创新的症结与困难提供了有益的启示,但本文梳理总结后发现:一方面,受限于数据可获得性与统计质量问题,许多研究者未能对中国企业的研

收稿日期:2017-07-24

基金项目:本文受国家自然科学基金面上项目"产业政策的微观基础和我国产业结构转型研究:基于产品空间理论的考察"(71273217)、国家自然科学基金应急管理项目"中国应对'双反'调查的策略研究与政策建议"(71741001)、中央高校基本科研业务费专项资金项目"大数据时代知识产权与创新发展研究"(2072015100)、马克思主义理论研究和建设工程重大项目"中国特色社会主义政治经济学研究"(2015MZD006)的资助。

作者简介:龙小宁,厦门大学经济学院、王亚南经济研究院,博士,教授,博士生导师;林志帆,厦门大学王亚南经济研究院,博士研究生。

致　　谢:作者感谢厦门大学经济学科应用微观午餐会参与师生的讨论,感谢两位审稿专家富有建设性的建议,但文责自负。

发创新状况形成全面的认识,往往在基本事实上陷入误区,观点过于悲观、甚至有"危言耸听"之嫌[①];另一方面,部分实证研究忽视了工业企业数据库、上市公司数据、企业调查数据等数据来源中研发创新变量呈现的离散选择、零值堆积、统计截断等分布特征,在将这些变量作为被解释变量时未能采取正确的计量方法进行处理,降低了实证研究的可信度。

因此,有必要对中国企业研发创新的基本事实进行系统整理,破除可能存在的误区,并尝试提供关于计量模型选择的建议。基于现有数据来源覆盖的行业范围与研发创新数据的可得性,本文仅对制造业企业进行研究。具体而言,本文对研究中最常用的四个数据来源(即 1998-2007 年的工业企业数据库、2007-2015 年的上市公司数据、2003 年/2005 年/2012 年世界银行企业调查数据、2002-2012 年间六轮中国私营企业调查数据)进行对比分析,尝试勾勒中国制造业企业研发创新的全幅图景。本文的一个重要发现是,常规年度工业企业数据库呈现的研发支出与新产品产出为正的企业比例仅约 10% 的"创新荒漠"现象很可能是统计纰漏造成的假象,严重低估了企业从事研发创新活动的比例——其他企业规模可比的数据来源(例如 2004 年全国经济普查、世界银行中国企业调查、中国私营企业调查)显示该比例至少在 35% 以上。此外,尽管国有部门较低的创新效率与资源错配问题仍不容忽视,但整体而言中国制造业企业的研发创新呈现出较为乐观的态势,研发投入、新产品产出、专利的申请与授权在广延与强度边际上都实现了较快的增长,各项研发创新指标基本都领先于其他新兴发展中国家。最后,本文指出,当研发创新数据作为被解释变量时,其服从的离散、混合、截断分布将使线性模型的 OLS 估计不一致,进而讨论 Probit 与 Logit 模型、Tobit 模型、Truncation 模型与 MLE 估计的适用性。

当然,本文仅是数据的使用者、而非发布者,不能保证掌握这些数据来源的所有信息,也难以洞悉企业研发创新活动的所有细节。因此,本文仅"抛砖引玉"地呈现一些在"干中学"的研究中发现的重要现象、数据事实与相应的推测,其中可能会包含一些主观倾向。但本文希望,本文能够提供企业层面研发创新研究的背景信息支持,也为学者们提供计量模型选择的参考,以推动这一领域的研究发展。

二、中国制造业企业研发创新状况分析

当前,与研发创新相关的学术研究中最常用的四个数据来源分别是:工业企业数据库、上市公司数据、世界银行企业调查数据、私营企业调查数据。这些数据来源中包含的与研发创新相关的变量如表 1 所示。以下本文分节对相关信息进行分析讨论。

表 1　　常用数据来源中研发创新变量信息一览表

数据库	工业企业数据库	上市公司数据	世界银行企业调查数据	私营企业调查数据
研发支出	√	√	√	√
研发/技术人员	×	√	√	√
新产品产出	√	×	√	√
专利	×	√	√	√

(一)工业企业数据库:1998-2007 年

来自国家统计局的工业企业数据库是学者们进行研发创新相关研究最为常用的数据来源。尽管未能包含更近年度的信息,但作为国家统计局收集的基础性数据,工业企业数据库具有样本量大、变

① 例如,在微信传播不到半个月阅读量超过十万次、被点赞两千余次的《中国制造已经穷途末路》就是这类文章的典型代表(链接:http://mp.weixin.qq.com/s/dRI03RZ83nNPmfa0Zq_3Gw)。

量多、时间跨度长等优点(聂辉华等,2012),基于这套数据研究得到的结论具有广泛的代表性。基于本文的研究目的,本文仅保留了 CIC 二位行业代码介于 13 至 43 之间的制造业企业,对研发支出与新产品产值这两个变量进行分析。

本节的核心发现为:除 2004 年第一次全国经济普查的数据外,常规年度工业企业数据库中研发支出与新产品产出变量约 90% 的观测值为零的现象很可能是统计纰漏的结果,大大低估了中国制造业企业进行研发创新活动的比例。本文将结合多方面证据证明,常常见诸报端或部分研究的背景陈述中类似于"中国制造业企业迷失于'创新荒漠'、创新乏力"等论断实际上是明显的误区。并且,针对零值观测值性质的讨论将对实证研究中的计量模型设定与估计方法选择产生重大影响。

1. 总体状况

研发支出与新产品产值这两个变量在各年度的数据可获得性如表 2 所示:

表 2　　工业企业数据库研发创新变量各年度可获得性信息

	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
研发支出	×	×	×	√	×	×	√	√	√	√
新产品产值	√	√	√	√	√	√	×	√	√	√

进而,表 3 和表 4 报告了这两个变量在各数据可得年份的统计信息,可以发现:

(1)除 2004 年(第一次全国经济普查数据,下文另行讨论)以外,2001 年报告研发支出的企业比例仅约 12.5%,而在 2005–2007 年间该比例下降至约 10%。查看正值样本,研发支出强度从 2001 年的 0.8% 逐年稳定上升至 2007 年的 1.4%;

(2)在近十年的样本跨度中,新产品产值为正的企业比例从 1998 年的近 8% 稳定地下降至 2003 年的 6.5% 左右,而在 2005–2007 年间则在 10% 左右波动,十年间该比例仅提高了两个百分点。但查看正值样本,可以发现新产品产值占总产值的平均比重从 1998 年的 33.64% 稳定攀升至 2007 年的 43.68%,上升了十个百分点①。

表 3　　工业企业数据库研发支出统计信息:分年度

年度	2001	2004	2005	2006	2007
正值企业比例(%)	12.4778	37.0069	10.0270	10.4122	10.9721
(研发支出/年产值)平均值	0.0081	0.0038	0.0125	0.0135	0.0140

说明:"(研发支出/年产值)平均值"统计的是正值样本以 0.5 为上限进行右侧缩尾后的平均值。

表 4　　工业企业数据库新产品产值统计信息:分年度

年度	1998	1999	2000	2001	2002
正值企业比例(%)	7.8894	7.5576	7.5440	7.5516	7.1205
(新产品产值/年产值)平均值	0.3364	0.3494	0.3640	0.3706	0.3780
年度	2003	2005	2006	2007	
正值企业比例(%)	6.5659	9.9885	10.3441	9.0178	
(新产品产值/年产值)平均值	0.3909	0.3610	0.3761	0.4368	

说明:"(新产品产值/年产值)平均值"统计的是正值样本信息。

① 学者们使用的工业企业数据库可能有诸多来源、数据清洗流程也各异、样本量也各不相同,但研发创新变量数值为正的企业比例大致都在 8%–15% 之间。

表3和表4在某种程度上是一种“怪象”:经济增长理论揭示,一国经济的发展将使其向世界技术前沿靠近,可供模仿学习的技术空间越来越小,经济增长的动力将逐渐转向研发创新。在1998-2007年间,中国的人均GDP从不到一千美元快速增长至约三千美元,世界排名攀升了数十位,大大缩短了与世界技术前沿的距离,因此可以预期中国企业的研发创新力度将逐渐增大——然而,工业企业数据库显示,企业研发创新仅有强度边际上的进步(研发支出强度与新产品产值占比提升),在广延边际上反而出现了轻微的下降。换言之,更少的企业做了更多的研发创新,研发创新活动似乎变得更为“集中”了。

基于以下两方面的证据,本文认为,常规年度工业企业数据库中研发创新变量为零的企业样本出现了统计纰漏,许多实际上进行了研发创新活动的企业被误报为零,造成了“从事研发创新活动的企业比例长期在低水平徘徊甚至下降”的“创新荒漠”假象:

(1)工业企业数据库中来自2004年全国第一次经济普查的数据①显示:约37%的企业研发支出为正,这与表3中其他年份仅10%出头的比例形成了巨大的反差;同时,2004年企业的平均研发支出强度仅约0.38%,远低于其他年份的平均水平——这说明,常规年度的统计工作可能对研发创新不够重视,将许多实际中有进行研发创新的企业的相关变量记录为零;而且,被误报为零的企业大多是研发创新强度较小的企业,因此,当它们的研发支出在经济普查中被正确记录时,企业的平均研发创新强度就下降了;

(2)2008年,国家统计局发布了“全国工业企业创新调查数据”报告,尽管具体抽样方案与统计口径没有披露,但该报告也提供了一些常规年度工业企业数据库可能存在统计纰漏的证据②:根据该报告中“表1-有创新活动的企业分布情况”,在2004-2006年期间,规模以上制造业企业有研发创新活动的比例为30%,其中63.3%的企业有R&D活动(研发支出为正)、83.5%的企业有产品创新(新产品产值为正),因此研发支出为正的企业比例应约为19%(即30%×63.3%)、新产品产值为正的企业比例应约为25%(即83.5%×63.3%),这些数字均远高于常规年度工业企业数据库所显示的水平。

此外,后文对三轮世界银行中国企业调查数据以及六轮全国私营企业调查数据的统计分析将提供更多关于常规年度的工业企业数据库存在统计纰漏的证据。

2. 分所有制分析

不同所有制间的差异也值得关注。首先,本文关注研发支出:表5呈现了工业企业数据库中国有企业、民营企业、外资企业③三个分样本的统计信息。

表5揭示,除2004年,国有企业报告研发支出的比例介于14-17%,外资企业的比例在12%左右,民营企业的比例在2001年约为11.5%、其余年份都低于10%;2004年的普查数据显示,约45%的国有企业与约52%的外资企业的研发支出为正,民营企业的比例仅约26%——总体而言,国有企业开展研发创新的比例最高,外资企业次之,民营企业最低。在研发支出强度方面:除2001年,国有企业的研发支出强度最大,从2001年的0.9%逐步上升至2007年的约2.1%,而民营企业与外资企业不相上下,研发支出强度从2001年的0.95%左右上升至2007年约1.7%——总体而言,国有企业的研发创新强度也是最大的。

尽管Hart et al.(1997)、Shleifer(1998)等经典研究指出,由于信息不对称、委托代理冲突等激励问题,国有企业进行研发创新的动机不足。但表5却呈现了一副“国有企业研发创新投入最为活跃”的图景,本文认为有三种可能的解释:

---

① 该年的数据与往年的主要区别为将规模以下的小型企业纳入统计范畴、对更多变量进行统计、数据质量也更高。本文仅保留规模以上企业的数据使之与其他年份可比。

② 数据来源:http://www.stats.gov.cn/ztjc/ztsj/2006gysj/200802/t20080222_61467.html。

③ 本文将工业企业数据库中的国有企业和数量极少的集体企业合称为国有企业,将港澳台资企业与来自其他国家(与地区)的外资企业统称为外资企业。

(1)国有企业的确是中国研发创新最为活跃的群体:在1998-2007年间,对于民营企业而言,积极争取国内外订单、利用低廉的要素成本进行粗放投资和生产是成本最低、利润最大的选择,同时民营经济也缺乏正规金融与知识产权保护的支持,故研发创新的动机不足;同时,外资企业研发投入较低的现象也易于理解——在发展中国家经营的外资公司可以利用母公司技术,无需进行独立的研发创新(Almeida & Fernandes,2008);相比之下,对于国有企业而言,中央政府一直强调"科学技术是第一生产力"的"政治任务"必须由它们承担,且在信贷、土地等资源获取上也更有优势,从而创新投入具有数量优势;

(2)国有企业平均规模更大引致统计偏误:始于1997年的"抓大放小"改革使国有企业的数量占比从1995年的24%持续下降至2015年的2.3%。在政策引导与市场竞争的"大浪淘沙"中,仅规模大、实力强的国有企业得以生存。Wei et al.(2017)指出,这些规模较大的国有企业往往在研发创新上具有数量优势,因此在不对企业规模进行控制的情况下,直接对比不同所有制企业的研发创新变量可能得到误导性的结论;

(3)统计纰漏在不同所有制间的差异:第一,民营企业规模较小,内部分工与财务制度可能相对于国有企业与外资企业较不规范,不设专门的研发岗位,难以界定哪些人员的工资支出与相关费用应记为研发支出,或没有将研发支出与管理费用科目下的其他项目区分开来;第二,出于偷税漏税避税或躲避监管等原因,民营企业在面对统计人员时往往不愿提供充分的信息。因此,如果本文相信常规年度的工业企业数据库存在统计纰漏,那么民营企业的误差很可能最为严重,民营企业的研发创新活动被低估的幅度可能最大。

表5　　工业企业数据库研发支出统计信息:分所有制与分年度

年度	2001	2004	2005	2006	2007
1. 国有企业					
正值企业比例(%)	13.9968	45.6028	13.9516	15.3561	16.3338
(研发支出/年产值)平均值	0.0090	0.0115	0.0181	0.0208	0.0208
2. 民营企业					
正值企业比例(%)	11.5759	26.3652	8.4809	8.5410	9.0696
(研发支出/年产值)平均值	0.0093	0.0024	0.0145	0.0163	0.0169
3. 外资企业					
正值企业比例(%)	11.6287	52.6492	10.3293	11.6282	12.7278
(研发支出/年产值)平均值	0.0097	0.0020	0.0147	0.0161	0.0167

说明:"(研发支出/年产值)平均值"统计的是正值样本信息。

为对解释(1)与(2)进行辨析,本文在表6中将各种所有制企业的观测值混在一起后按照销售额分为五个从小到大的分位段,在每个区间内对不同所有制企业的研发支出变量进行统计与对比。表6显示:在越大的规模区间,企业进行研发的比例也越高,说明企业规模与研发创新的广延边际间的确存在明显的正相关。在每个区间,国有企业与外资企业研发支出为正的比例较为接近①,而民营企业的比例垫底;进而查看研发支出强度,可发现国有企业均高于民营企业与外资企业——这便说明,即便对企业规模进行分组控制,无论从有研发投入的企业比例还是从投入强度来看,国有企业都是研发创新投入最为活跃的群体。因此,本文在解释(1)与(2)间倾向于接受(1);不过,现有数据无法对解释(3)进行直接的检验,本文只能基于其他方面的证据对统计纰漏问题进行推断。

① 在较小的规模区间,外资企业的比例稍高;在较大的规模区间,国有企业的比例稍高。

表 6　　工业企业数据库研发支出统计信息:分所有制与分企业规模

研发支出	企业类型	企业销售额分位数区间				
		0-20%	20-40%	40-60%	60-80%	80-100%
正值企业比例(%)	国有企业	13.0788	14.9730	17.1327	20.8769	36.4659
	民营企业	8.8680	10.1700	10.9531	12.4965	19.8179
	外资企业	14.4016	16.5742	16.5233	17.9084	26.8938
正值样本的(研发支出/年产值)平均值	国有企业	0.0161	0.0102	0.0100	0.0094	0.0094
	民营企业	0.0088	0.0084	0.0085	0.0080	0.0078
	外资企业	0.0079	0.0067	0.0058	0.0062	0.0067

进而,本文关注研发创新活动的产出:表 7 呈现了国有企业、民营企业、外资企业的新产品产值的统计信息。首先,查看新产品产值为正的企业比例:在 1998-2007 年间,国有企业该比例从近 9% 上升至 2006 年约 14% 的顶点,民营企业从约 5% 上升至 2006 年约 9.6% 的顶点,外资企业则从 1998 年的近 7% 上升至 2007 年的 10%——在所有年度,国有企业新产品产值为正的比例都是最高的,这与它们进行研发投入的比例最高的事实相对应。然而,查看新产品产值占总产值的比重却发现:在所有年度,国有企业的新产品产值比重都最低,稳定在 33% 上下;而外资企业这一比例最高,基本都在 45% 以上;民营企业平均而言也在 40% 左右——结合研发支出强度的信息可知,国有企业以最高的研发投入强度产出了最低的创新产出密度。从投入产出关系来看,国有企业可能存在创新效率较低的问题。

表 7　　工业企业数据库新产品产值统计信息:分所有制与分年度

年度	1998	1999	2000	2001	2002	2003	2005	2006	2007
1. 国有企业									
正值企业比例(%)	8.9034	8.6906	8.8348	9.0426	9.2320	8.9410	13.9535	14.1398	12.6356
(新产品产值/年产值)平均值	0.3049	0.3157	0.3260	0.3346	0.3362	0.3488	0.3321	0.3489	0.4077
2. 民营企业									
正值企业比例(%)	5.1633	5.0064	5.6279	5.3209	5.6169	5.1975	9.0935	9.6343	7.8178
(新产品产值/年产值)平均值	0.4053	0.3977	0.4155	0.4249	0.4203	0.4195	0.3447	0.3574	0.4376
3. 外资企业									
正值企业比例(%)	6.8416	6.9108	6.8639	8.0063	6.1794	5.9558	8.6435	9.3691	9.9941
(新产品产值/年产值)平均值	0.4765	0.4815	0.4731	0.4216	0.4413	0.4432	0.4500	0.4651	0.4627

说明:"(新产品产值/年产值)平均值"统计的是正值样本信息。

与表 6 相似,为排除企业规模的干扰,表 8 给出了三种所有制企业的新产品产值变量分规模区间对比的信息。可以发现,除了在最小的区间,国有企业在其余区间内新产品产值为正的比例基本都高于民营企业与外资企业;但在新产品产值比重的统计上,在所有的规模区间内国有企业都是最低的,尤其与外资企业的差距大于十个百分点;另外,Wei et al.(2017)匹配国家知识产权局的专利数据与工业企业数据库发现,在所有的规模区间,国有企业拥有的专利数都是最少的,这些证据进一步确认了国有企业创新效率最低的事实。

一言以蔽之,表 5-8 揭示,尽管国有企业的研发投入最多,但创新产出最少,揭示出国有企业的关键问题是低下的创新效率。从工业企业数据库中本文还发现,无论是绝对量还是控制企业规模影响的相对量,国有企业也得到了最多的政府补贴,而且它们在银行贷款的可得性与贷款利率、土地资源等方面也具有明显优势——这意味着,大量的资源流向了研发创新效率最低的国有部门,这一严重的资源错配问题值得关注。

表 8　　工业企业数据库新产品产值统计信息:分所有制与分企业规模

研发支出	企业类型	企业销售额分位数区间				
		0-20%	20-40%	40-60%	60-80%	80-100%
正值企业比例(%)	国有企业	3.6814	5.8167	7.6548	11.4572	25.0284
	民营企业	4.7544	5.6064	6.4691	8.1496	13.8843
	外资企业	5.1014	5.2017	5.6705	6.9087	12.6855
正值样本的(新产品产值/年产值)平均值	国有企业	0.3984	0.3579	0.3425	0.3245	0.3216
	民营企业	0.4841	0.4375	0.3975	0.3701	0.3463
	外资企业	0.5465	0.4901	0.4678	0.4512	0.4399

3. 分行业分析

本文按照二位数制造业行业代码对研发支出与新产品产值进行分行业统计与对比。虽然工业企业数据库中研发创新变量可能存在统计纰漏,但各行业中研发创新变量为正的企业比例以及正值样本的统计数据仍可以反映一些有效信息(见表 9)。①

表 9　　工业企业数据库分行业研发创新概况

研发支出正值比例最高的五个行业	新产品产值正值比例最高的五个行业
烟草制品业　医药制造业　仪器仪表及文化、办公用机械制造业　通信设备、计算机及其他电子设备制造业　专用设备制造业	医药制造业　仪器仪表及文化、办公用机械制造业　通信设备、计算机及其他电子设备制造业　专用设备制造业　烟草制品业
研发支出正值比例最低的五个行业	新产品产值正值比例最低的五个行业
木材加工及木、竹、藤、棕、草制品业　废弃资源和废旧材料回收加工业　黑色金属冶炼及压延加工业　农副食品加工业　造纸及纸制品业	废弃资源和废旧材料回收加工业　印刷业和记录媒介的复制　纺织服装、鞋、帽制造业　造纸及纸制品业　农副食品加工业
研发支出强度最高的五个行业	新产品产值比重最高的五个行业
仪器仪表及文化、办公用机械制造业　通信设备、计算机及其他电子设备制造业　医药制造业　专用设备制造业　交通运输设备制造业	通信设备、计算机及其他电子设备制造业　纺织服装、鞋、帽制造业　仪器仪表及文化、办公用机械制造业　皮革、皮毛、羽毛(绒)及其制品业　文教体育用品制造业
研发支出强度最低的五个行业	新产品产值比重最低的五个行业
皮革、皮毛、羽毛(绒)及其制品业　造纸及纸制品业　黑色金属冶炼及压延加工业　纺织服装、鞋、帽制造业　木材加工及木、竹、藤、棕、草制品业	烟草制品业　农副食品加工业　饮料制造业　黑色金属冶炼及压延加工业　造纸及纸制品业

从这些行业排名来看,研发创新较为活跃的行业基本都是新兴的资本技术密集行业,而创新性较低的行业基本为传统的劳动密集型行业,与本文的常识认知一致。

(二)上市公司数据:2007-2015 年

中国的上市公司财务报表与附注数据也提供了研发创新数据,本文得以了解这些精英企业的研发创新状况。本文从国泰安 CSMAR 数据库的"公司专利与研发创新"子库获取上市公司从 2007-2015 年②间研发投入(包括研发支出与研发人员数)、专利申请和授权的数据,并根据证监会的《上市

① 由于篇幅限制,相应的统计表格略去,感兴趣的读者可以联系作者索取。

② 财政部于 2006 年发布 38 项《企业会计准则》具体准则,要求自 2007 年 1 月 1 日起在上市公司范围施行(http://kjs.mof.gov.cn/zhengwuxinxi/zhengcefabu/200805/t20080522_33653.html)。

公司行业分类指引(2001 年)》选择制造业企业进行分析。

1. 研发投入

表 10 提供了制造业上市公司研发支出数据的统计信息。

表 10　　制造业上市公司研发支出统计信息:分年度

年度	2007	2008	2009	2010	2011
正值企业比例(%)	11.0980	19.2888	23.5234	12.9348	34.5440
(研发支出/营业收入)平均值	0.0202	0.0288	0.0323	0.0377	0.0388
年度	2012	2013	2014	2015	
正值企业比例(%)	82.1333	84.9469	85.8302	87.4704	
(研发支出/营业收入)平均值	0.0356	0.0376	0.0382	0.0408	

说明:(1)"(研发支出/营业收入)平均值"统计的是正值样本以 0.5 为上限进行右侧缩尾后的平均值;(2)历年制造业上市公司数量为 847、928、982、1041、1294、1500、1601、1602、1692 个。

从表 10 可以发现一些有趣的现象:

(1)在 2012 年以前,报告了研发支出的企业比例在 11-35% 间震荡——但这一期间的数据不应理解为制造业上市公司开展研发创新活动的真实比例,而更可能反映了企业财务人员对 2006 年底财政部公布的《企业会计准则》仍在进行学习与适应。上市公司的专利申请数据可为上面观点提供侧面证据:假如将研发支出视为创新投入,而专利申请与授权视为创新产出,那么具有创新投入的企业比例应大于或等于具有创新产出的企业比例。而在 2007-2011 年间,进行了专利申请的企业比例从约 42% 上升至 65%(见表 11),远高于报告研发支出的企业比例。将专利申请数据与企业研发支出数据匹配也可以发现,有相当多进行了专利申请的企业在当年与之前年份都没有报告研发支出,图 1 左边的 Venn 图对这一现象进行了简洁的描绘。这种投入-产出"倒挂"的异象说明:2007-2011 年间上市公司的研发支出数据披露不完全,低估了制造业上市公司进行研发活动的比例①。在财务数据没有披露报告研发支出的上市公司中,实际上有相当比例的企业进行了研发创新活动;

(2)在 2012 年后,报告研发支出的制造业上市公司比例陡然超过 80% 并逐年上升②,略高于同期这些企业进行专利申请的比例。更为重要的是,进行了专利申请的企业基本都报告了研发支出数据(如图 1 所示),创新的投入产出关系符合逻辑。因此,2012 年后的数据应当较为接近该期间制造业上市公司开展研发创新活动的真实比例。

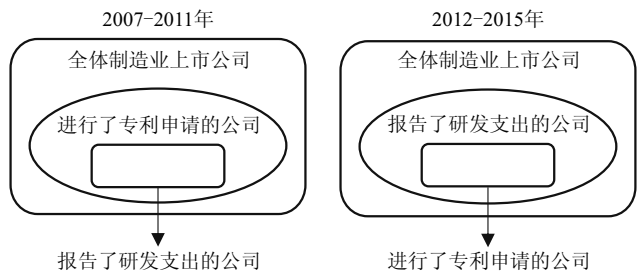

图 1　2007-2015 年间制造业上市公司创新投入与产出关系

表 10 也报告了制造业上市公司的研发支出强度:在"财力"方面,在正值样本中,研发支出强度从 2007 年的 2.02% 逐年上升至 2015 年的 4.08%,近十年间翻了一番,发展态势良好。在"人力"方面:在 2015 年,约 86% 的企业报告了研发人员数,这一比例与报告了研发支出的企业比例非常接近,研发人员占企业雇员人数比例的均值约为 12.65%。

① 这一问题并非中国独有:Koh & Reeb(2015)发现,Compustat 数据库中也存在一些有专利活动但 R&D 支出一直为零的"伪空白"(pseudo-blank)公司。

② 本文猜测,企业的研发支出占销售收入之比超过某些门槛就可能得到"高新技术企业认定"而获得税收优惠和政府补助(杨国超等,2017)是企业主动披露研发支出意愿增强的重要动机之一。

2. 专利申请与授权

表 11 报告了制造业上市公司的专利活动信息，本文还分别对发明、实用新型、外观设计进行统计，从中对专利结构与质量的演变情况进行分析。表中首先报告了制造业上市公司各类专利加总的逐年统计数据，可以发现，进行了专利申请的企业比例从 2007 年的 42% 逐年上升至近年接近 70% 的水平；此外，表中还显示，进行了发明、实用新型、外观设计申请的企业比例分别从 34%、25%、15% 上升至约 60%、50%、20%，体现出明显的进步，说明企业的知识产权保护意识逐渐加强。从数量上看，企业年均专利申请和授权量从 2007 年的约 25 个和 22 个上升至近年的约 50 个和 30 个，增长明显；分类来看，发明专利申请量实现了最大幅度的增长，实用新型次之，但外观设计增长不明显——这说明，制造业上市公司的专利结构逐渐偏向创新含量高、经济价值大、授权难度大、保护期限长的发明专利。在中国整体专利结构在近年因大量“灌水”的实用新型与外观设计而持续恶化的大背景下（龙小宁和王俊，2015），“精英阶层”的制造业上市公司在专利质量上出现了“逆势”上升，呈现出乐观的发展态势。

表 11　　制造业上市公司专利申请与授权统计信息：分年度与专利类型

年度	有专利申请的企业比例（%）	企业平均专利申请数	企业平均专利授权数	专利申请授权率（%）	有发明申请的企业比例（%）	企业平均发明专利申请数	企业平均发明专利授权数	发明申请授权率（%）
2007	42.1488	24.7031	21.7591	88.0825	34.3566	9.9691	6.3574	63.7711
2008	48.1681	26.7987	23.1163	86.2593	38.7931	11.4028	6.8306	59.9026
2009	52.7495	29.4170	25.3436	86.1531	44.9084	12.1225	7.4626	61.5600
2010	55.3314	33.4931	28.7344	85.7920	47.3583	13.6024	8.0426	59.1262
2011	64.3740	37.2737	31.2317	83.7901	56.0278	15.0579	8.1159	53.8976
2012	66.4667	41.8084	33.2538	79.5384	58.6000	17.9647	8.2628	45.9946
2013	68.4572	42.6150	29.4462	69.0982	61.0244	19.4166	4.6448	23.9220
2014	66.4170	49.4624	30.4107	61.4825	58.3645	22.2064	0.5316	2.3937
2015	61.7612	44.4124	18.6469	41.9857	52.8960	22.1441	0	0
年度	有实用新型申请的企业比例（%）	企业平均实用新型专利申请数	企业平均实用新型专利授权数	实用新型申请授权率（%）	有外观设计申请的企业比例（%）	企业平均外观设计专利申请数	企业平均外观设计专利授权数	外观设计申请授权率（%）
2007	25.8560	15.7717	15.7717	100	15.5844	18.6667	18.6667	100
2008	32.8663	16.7475	16.7475	100	17.4569	17.0741	17.0741	100
2009	38.3910	18.1804	18.1804	100	17.1079	18.0833	17.7560	98.1896
2010	40.6340	21.2222	21.2222	100	17.2911	20.0500	20.0500	100
2011	48.9954	22.7697	22.7697	100	20.0927	21.9077	21.9077	100
2012	52.1333	25.6535	25.6522	99.9950	20.2000	19.2442	19.2442	100
2013	54.2786	24.6133	24.6122	99.9953	20.6121	19.2333	19.2333	100
2014	52.7466	29.0923	29.0888	99.9878	22.0974	20.5706	20.5650	99.9725
2015	45.8629	26.4459	19.2320	72.7220	18.0260	19.9016	14.9574	75.1565

说明：每年的专利申请授权率以当年的申请截止至数据提取时点（2017 年 5 月）被授权的比例计算。

表 11 还显示，专利总体的申请授权率从 2007 年超过 88% 的水平逐年下降至 2015 年仅约 42% 的水平，这是否说明中国制造业上市公司的专利申请质量发生了恶化呢？

本文认为，这一现象需要综合发明、使用新型、外观设计三类专利的授权率差异进行理解：（1）发明专利的授权率在 2007–2010 年间稳定在 60% 左右，但近年的授权率却出现急速下降。这很可能是发明专利从申请到授权具有较长时滞造成的——在中国现行的专利制度下，发明从申请到授权需要经历形式审查与实质审查两个环节，Xie & Zhang（2015）对 1985–2009 年间工业企业申请的专利进行统计发现，发明从申请到授权的平均时滞为 47 个月。因此，从 2012 年起上市公司发明专利的授权率

出现大幅下降的主要原因应该是有相当部分的申请尚未完成实质审查;(2)相比之下,实用新型与外观设计专利只需要企业提交完整的申请材料,无需实质审查即可授权,时滞仅约 12 个月(Xie & Zhang,2015)。除样本末期有部分申请尚未完成形式审查流程导致申请授权率下滑外,这两类专利的申请授权率接近 100%。综上可知,近十年间制造业上市公司的专利申请质量应当是较为稳定的,专利总体申请授权率持续下降的现象应当是申请结构逐年偏向授权率较低的发明专利以及较近年份的各类申请尚未完成审核流程这两大原因造成的。

3. 分所有制分析

进而,本文根据国泰安 CSMAR 数据库提供的实际控制人信息,将制造业上市公司分为国有企业与非国有企业两组,对研发支出、专利申请与政府补贴进行统计分析。

表 12　制造业上市公司研发支出、专利申请与政府补贴统计信息:分所有制与分年度

分表 A:国有企业	研发支出正值比例(%)	研发支出强度	有专利申请的企业比例(%)	专利申请强度	有政府补贴的企业比例(%)	政府补贴强度
2007	11.0874	0.0208	43.0704	0.9836	80.1706	0.0101
2008	15.3551	0.0269	49.5202	1.2213	89.8273	0.0106
2009	18.6047	0.0286	54.1861	1.3338	95.1163	0.0113
2010	16.5179	0.0322	56.9196	1.1212	95.5357	0.0109
2011	18.8034	0.0374	63.6752	1.3264	96.3675	0.0118
2012	79.8403	0.0303	64.0719	1.4944	96.4072	0.0166
2013	82.6962	0.0303	62.3743	1.2482	96.3783	0.0148
2014	83.5010	0.0319	62.3743	1.0824	97.7867	0.0143
2015	86.9650	0.0360	57.7821	0.9208	98.8327	0.0181
分表 B:非国有企业	研发支出正值比例(%)	研发支出强度	有专利申请的企业比例(%)	专利申请强度	有政府补贴的企业比例(%)	政府补贴强度
2007	11.8980	0.0195	42.7762	1.7170	79.6034	0.0113
2008	26.0526	0.0302	49.4737	3.4330	83.9474	0.0128
2009	28.9575	0.0343	54.0541	3.5040	90.7336	0.0127
2010	33.9318	0.0400	57.2711	3.3859	92.1005	0.0162
2011	45.3739	0.0391	67.4271	2.9956	94.9303	0.0140
2012	86.5323	0.0381	69.9894	2.4389	96.1824	0.0164
2013	89.0267	0.0402	73.4733	2.2441	97.1374	0.0149
2014	89.9232	0.0401	70.6334	2.2164	97.3129	0.0135
2015	90.7776	0.0421	65.3707	1.6377	97.8300	0.0143

说明:(1)研发支出强度与政府补贴强度是正值样本的(研发支出/营业收入)与(政府补贴/营业收入)在以 0.5 为上限进行右侧缩尾后的平均值;(2)专利申请强度是正值样本的(专利申请数/营业收入(单位:亿元))在以 500 为上限进行右侧缩尾后的平均值。

表 12 揭示:(1)在所有年度,非国有企业报告研发支出的比例一直高于国有企业①,且除 2007 年外,非国有企业的研发支出强度都高于国有企业。这说明,在研发投入方面,非国有企业领先于国有企业②;(2)自 2011 年起,非国有企业进行专利申请的比例高于国有企业,且在控制企业规模后,非国有企业的专利申请强度在所有年份都高于国有企业——非国有企业的每亿元营业收入对应 1.6–3.5 个专利申请,而国有企业的每亿元营业收入仅对应 0.9–1.5 个专利申请——这说明,在创新产出方

① 不过,仍然需要注意:2007–2011 年间的数值不能代表真正具有研发支出企业的比例。

② 此外,在 2015 年,非国有企业超过 90% 拥有研发人员,研发人员占比平均值约为 13.4%;而国有企业约 83% 拥有研发人员,研发人员占比的平均值约为 11.3%,非国有企业同样领先。

面,非国有企业同样领先于国有企业①;(3)在政府补贴方面,在近年几乎所有的上市公司都获得了补贴。获得政府补贴的国有企业比例微微高于非国有企业比例,但在大多数年度非国有企业的政府补贴强度微微高于国有企业,总体而言不同所有制企业在政府补贴资源上几乎没有差异。

表 12 反映的信息与工业企业数据库(绝大多数企业都没有上市)的情况既有区别又有相似:工业企业数据库显示,国有企业中得到政府补贴的比例和补贴强度都明显领先,充裕的资源优势使得国有企业在研发支出方面领先、但创新产出垫底,不同所有制间凸显出明显的创新效率差异与资源错配问题;在上市公司群体中,非国有企业的研发支出强度大致为国有企业的 1.2 倍,而专利申请强度接近国有企业的 2 倍,可见创新效率差异仍然存在。但关键的区别在于,上市公司群体中非国有企业在政府补贴上不再落后于国有企业,资源约束被明显放松,一定程度上减轻了资源错配。而一旦资源错配减轻,非国有企业即表现出更高的研发支出强度,进一步反映了金融资源错配在研发方面的负面影响。这说明,在当前如火如荼进行中的供给侧改革背景下,打破广大非上市民营企业的资源约束瓶颈、降低资源错配程度将是激励企业创新的重要举措。

(三)世界银行企业调查数据:2000-2011 年

1. 中国企业状况

迄今为止,世界银行在中国进行了三轮企业调查,分别是在 2003 年与 2005 年发布的 Investment Climate Survey 和在 2012 年发布的 Enterprise Survey,均提供了与研发创新相关的信息,覆盖的年度范围分别为 2000-2002 年、2002-2004 年、2009-2011 年——前两轮调查与工业企业数据库的样本期有较长的重合,因此信息具有一定的可比性;第三轮调查则有助于本文了解近年来中国经济新一轮转型中企业研发创新行为的变化。

世界银行企业调查最重要的特征之一是数据质量高——大部分问题的选项都包含"不知道(don't know (spontaneous))"和"不适用(not applicable (spontaneous))",供受访者拒绝或无法回答时选择——去掉这一小部分样本后,剩余样本基本是真实数据,这与常规年度的工业企业数据库中研发创新变量可能存在统计纰漏的情况大不相同。本文基于删去回答"不知道"和"不适用"观测值后的样本展开统计分析。需要注意的是,这三轮调查数据中绝大部分样本为民营企业,因此,本文梳理得到的信息主要反映民营企业的研发创新状况。

表 13-15 分别报告了这三轮企业调查的统计信息,可以发现:(1)在 2000-2002 年间,约 40% 的企业有"财力"和"人力"投入,研发支出强度在 5% 以上,研究开发人员占总雇员数的比例在 9% 以上,约 40% 的企业有新产品产出,新产品销售额占比约为 35%。此外,约有 25% 的企业拥有专利;(2)在 2002-2004 年间,研发支出为正的企业比例超过了 50%,相对于第一轮调查约 40% 的比例进一步提升,研发支出强度在 2% 上下②;(3)在 2009-2011 年间,研发支出和新产品产值为正的企业比例都超过了 40%,与前两轮调查较为接近。其中,研发支出强度平均值高于 5%,仍然维持了较高的水平;新产品销售额占比的平均约为 25%,略低于前两轮调查的水平。

表 13　　2003 年世界银行中国企业调查数据研发创新变量信息

问卷问题	正值企业比例(%)	研发创新强度
在 2000 年,是否有研发支出?	36.6357	0.0561
在 2001 年,是否有研发支出?	38.9691	0.0508

① 本文还计算了国有企业与非国有企业的专利申请中发明专利的比例以及发明专利的申请授权率,没有发现显著的差异,说明不同所有制企业的专利申请质量较为接近。囿于篇幅限制没有报告。

② 由于世界银行第二轮调查没有提供销售额的数据,本文仅能以核心业务收入作为分母对研发支出进行标准化,故研发支出强度与其余两轮不可比。

续表

问卷问题	正值企业比例(%)	研发创新强度
在2002年,是否有研发支出?	45.2929	0.0534
在2000年,是否有专职的研究开发人员?	39.7104	0.0911
在2001年,是否有专职的研究开发人员?	42.0455	0.0927
在2002年,是否有专职的研究开发人员?	46.2098	0.0974
在过去三年间,是否引入了新产品?	44.1324	–
在2000年,是否引入了新产品?	42.5490	0.3721
在2001年,是否引入了新产品?	36.2745	0.3474
在2002年,是否引入了新产品?	32.6471	0.3233
企业是否有专利?	24.2157	–

原始数据来源:http://microdata.worldbank.org/index.php/catalog/591。

说明:(1)该轮调查包含1548个企业(其中制造业企业1020个),分布于北京市、天津市、上海市、广州市、成都市共5个城市;(2)研发创新强度是正值样本的(研发支出/年销售额)、(新产品销售额/年销售额)、(研究开发人员数/总雇员数)在进行以0.5为上限进行右侧缩尾后的平均值。

表14　　2005年世界银行中国企业调查数据研发创新变量信息

问卷问题	正值企业比例(%)	研发创新强度
在2002年,企业的研发支出有多少?	52.2903	0.0206
在2003年,企业的研发支出有多少?	55.2339	0.0189
在2004年,企业的研发支出有多少?	56.9677	0.0191

原始数据来源:http://microdata.worldbank.org/index.php/catalog/602。

说明:(1)该轮调查包含12400个企业(全部为制造业企业),分布于30个省区共120个城市;(2)研发创新强度是正值样本的(研发支出/核心业务收入)在以0.5为上限进行右侧缩尾后的平均值。

表15　　2012年世界银行中国企业调查数据研发创新变量信息

问卷问题	正值企业比例(%)	研发创新强度
在过去三年间,是否有研发支出?	41.2746	0.0523
在过去三年间,是否引入了新产品?	43.3234	0.2488

原始数据来源:http://microdata.worldbank.org/index.php/catalog/1559。

说明:该轮调查包含2700个企业(其中制造业企业1692个),分布于11个省区共25个城市;(2)研发创新强度对于新产品是正值样本的(新产品销售额/年销售额)的平均值,对于研发支出是正值样本的(研发支出/年销售额)在以0.5为上限进行右侧缩尾后的平均值。

本文注意到,这些结果与常规年度工业企业数据库中研发创新变量正值比例仅约10%的情况形成了巨大的反差,可能的原因有两个:(1)世界银行第一轮调查仅在北京市、天津市、上海市、广州市、成都市这五个经济发展水平最高的城市进行,第二轮调查所在的120个城市也是经济发展水平较高的样本——这些城市在交通水电等基础设施、行政审批、政策支持、法制环境、人力资本供给等方面为企业提供了良好的支持,因此当地企业进行研发创新的比例可能会明显高于全国平均水平;(2)如前所述,常规年度工业企业数据库的研发创新变量出现了统计纰漏,大量为零的样本中包含了许多实际上进行了研发创新活动的企业,从而大大低估了企业进行研发创新活动的比例。

为辨析这两种可能性,本文从工业企业数据库中提取1998–2002年间北京市、天津市、上海市、广州市、成都市这五个城市的企业样本进行统计,发现研发支出和新产品产值为正的企业比例也仅略微超过10%,远低于世界银行第一轮调查显示的水平;进而提取第二轮调查的120个城市的企业样本进

行统计,发现研发支出为正的企业比例也远低于世界银行数据,从而第一种可能性可以基本排除,进一步确认了常规年度工业企业数据库存在统计纰漏的事实。简而言之,中国企业的研发创新活动早在本世纪初就已经相当活跃,工业企业数据库所呈现的"创新荒漠"实际上是统计纰漏引发的假象。

2. 跨国比较:新兴发展中国家状况

世界银行企业调查数据的另一重要特征为跨国可比性——世界银行于 2000 年前后开始在全球范围对发展中国家企业进行的投资环境调查(Investment Climate Survey)、营商环境与企业表现调查(Business Environment and Enterprise Performance Survey)等采取了较为一致的调查问卷;在 2006 年后整合形成的企业调查(Enterprise Survey)项目对问卷进行了统一,统计口径具有完全的跨国可比性。本文从中挑选一些较有代表性且发展程度与中国相近的新兴发展中国家(包括金砖国家中的巴西、印度、南非,以及阿根廷、韩国、墨西哥、马来西亚、印度、越南①),对这些国家企业的新产品、研发投入、专利进行统计分析,有助于对中国企业研发创新的国际地位与发展趋势进行判断②。

本文发现,中国企业的研发创新总体而言领先于其他新兴发展中国家,具体而言:

(1)对于同期经济发展水平稍高于中国的阿根廷、巴西、南非,研发支出为正的企业比例在 50% 左右,与中国相当,但中国约 5% 的研发支出强度明显领先于这些国家约 3% 的水平;不过,这些国家有新产品的企业比例稍高一些。对比韩国发现,中国在引入新产品的企业比例与韩国大致相当,具有研发支出的企业比例与研发支出强度则超过韩国;

(2)对于同期经济发展水平与中国相当或稍低的马来西亚、墨西哥、印度、越南,引入新产品、研发支出为正、具有专职研究开发人员的企业比例基本都低于中国。

但值得注意的是,越南企业的研发创新在其于 2006 年加入 WTO 后取得了极为瞩目的进步:研发支出为正的企业比例从 2004 年不足 10% 上升至 2015 年的 25%,研发支出强度也从约 2.6% 上升至 4.65%;引入新产品的企业比例约为 35%,接近中国 2012 年的水平,新产品产值占年销售额的比重约 40%,比中国 2012 年的水平还高出近 15%。这说明,部分当前经济发展水平较低的发展中国家同样进行了相当活跃的研发创新活动。印度、越南、马来西亚等经济开放、自然资源丰裕、人口年龄结构年轻、劳动力成本具有明显优势的国家不仅在逐渐承接低端制造业的转移,在未来还将对中国制造业大国的地位形成挑战。

(四)中国私营企业调查数据:2001-2011 年

由中共中央统战部、中华全国商业联合会、中国(民)私营经济研究会等部门自 1993 年起联合进行的十余轮中国私营企业调查(Chinese Private Enterprise Survey,CPES)也有部分轮次的数据提供了与企业研发创新相关的信息。该调查每轮在中国 31 个省市(不含港澳台地区)内按照约 0.05% 的比例对不同规模、不同行业的私营企业进行分层随机抽样,具有广泛的代表性。中国私营企业调查数据的问卷内容虽然历经变更,但较多问题与备择项保持了较高的一致性,因此信息在时间维度上的可比性较强。2002 年、2004 年、2006 年、2008 年、2010 年、2012 年共 6 轮数据提供了研发创新的相关信息,从中可以探究私营制造业企业研发创新活动的演变情况,并与其它数据来源进行对比。与世界银行企业调查相似,中国私营企业调查也具有数据质量的优势,本文可以清晰地区分绝大部分变量的缺失值与零值。表 16 呈现了基于这 6 轮调查的统计数据。

可以发现:(1)在 2001-2011 年间,中国私营制造业企业报告研发支出的比例在 50%-65% 之间,与世界银行企业调查较为接近,同样远高于常规年度工业企业数据库的水平,而研发支出强度在 3.3%-10% 之间;(2)超过 95% 的企业拥有技术人员,占总雇员人数的平均比在于 11%-13.3% 间,研

① 这些国家在过去一二十年间实现了较为稳定的经济社会发展,发展制造业的成本与资源优势较为明显,部分国家已逐渐取代中国成为新的全球中低端制造业基地。

② 由于篇幅限制,相应的统计表格在正文略去。

发创新的“人力”投入强度也较大;(3)仅 2006 年的数据提供了新产品的信息,近 60% 的企业报告了新产品销售额,占总销售额的比重约为 46%,稍高于世界企业调查,更远高于常规年度的工业企业数据库;(4)拥有专利的企业比例在 2001 年约为 24%,与 2003 年的世界银行企业调查显示的水平极为接近,该比例在后续轮次逐步上升至 2007 年的近 36%,企业平均专利数也从约 4.16 个上升至 8.92 个,翻了一番有余,体现出快速的进步,说明私营企业保护技术和知识产权的意识逐渐加强。

表 16　　2002–2012 年间六轮私营企业调查数据研发创新变量信息

调查年度(制造业企业数/总调查企业数)	正值企业比例(%)	研发创新强度
2002 年(1186/3258)		
2001 年是否有研发支出?	65.6331	0.0997
2001 年是否有技术人员?	98.6388	0.1090
企业是否有专利(至 2001 年)?	23.742	4.1561
2004 年(988/3012)		
2003 年是否有研发支出?	64.3011	0.0940
2003 年是否有技术人员?	97.7726	0.1231
企业是否有专利(至 2003 年)?	27.2321	5.2582
2006 年(1452/3837)		
2005 年是否有研发支出?	63.5714	0.0344
2005 年是否有技术人员?	95.9620	0.1197
近三年是否推出了新产品?	57.3818	0.4635
企业是否有专利(至 2005 年)?	27.3460	8.8499
2008 年(1460/4098)		
2007 年是否有研发支出?	63.8596	0.0325
2007 年是否有技术人员?	97.2384	0.1325
企业是否有专利(至 2007 年)?	35.8571	8.9223
2010 年(1695/4614)		
2009 年是否有研发支出?	58.5479	0.0358
2012 年(1786/5073)		
2011 年是否有研发支出?	51.4322	0.0637

说明:(1)前五轮的研发支出强度是正值样本的(研发支出/年销售额)在以 0.5 为上限进行右侧缩尾后的平均值,对于最后一轮是正值样本的(研发支出/营业收入)在以 0.5 为上限进行右侧缩尾后的平均值;(2)技术人员的投入强度是正值样本的(技术人员数/企业雇员数)在以 0.5 为上限进行右侧缩尾后的平均值;(3)新产品的产出强度是正值样本的(新产品销售额/年销售额)的平均值;(4)专利强度是正值样本的专利数以 100 个为上限进行右侧缩尾后的平均值;(5)前两轮的研发支出数据为“企业的新产品、新技术、新项目的研发投资”,第三至第五轮的研发支出数据为“企业的研发费用”,最后一轮的研发支出数据为“企业的新产品研发经费”与“企业的技术创新、工艺改造费用”相加。

总体而言,中国私营企业调查数据库说明,在 2001–2011 年间,私营制造业企业的研发创新活动较为活跃并持续进步,反映的信息与世界银行企业调查高度一致。

(五)解读多个数据来源的一致与冲突

以上四个小节整理分析了工业企业数据库、上市公司数据、世界银行企业调查数据、中国私营企业调查数据中制造业企业的研发创新状况,表 17 对各个数据来源的关键信息进行了总结,本文尝试从中分析多个数据来源的一致与冲突。需要注意的是,研发创新活动往往与企业规模正相关①。因

① 在与研发创新相关的实证研究中,企业规模往往作为解释变量引入,许多研究都发现企业规模与研发创新的概率间存在显著且稳定的正相关关系(如聂辉华等,2008 等)。

此,在对比不同数据来源的信息时,需要考虑企业规模差异的影响。本文使用销售额的平均值和中位数作为代理变量,发现企业平均规模从大到小分别为上市公司、工业企业数据库、世界银行企业调查、中国私营企业调查。因此,合理的预期是,企业开展研发创新的比例将在上市公司、工业企业数据库、世界银行企业调查、中国私营企业调查间呈现大致递减的趋势。

表 17　　中国制造业企业研发创新状况关键信息:基于四个常见数据来源的总结

数据库	工业企业数据库 (常规年度)	上市公司数据	世界银行 中国企业调查数据	中国私营企业 调查数据
覆盖年度范围	1998-2007 年	2007-2015 年	2000-2011 年	2001-2011 年
企业销售额平均值	7.3467 亿元	52.5721 亿元	4.4738 亿元	0.9775 亿元
企业销售额中位数	1.7199 亿元	14.0358 亿元	0.4481 亿元	0.1300 亿元
研发支出情况	正值比例:10%-13% 平均强度:0.8%-1.4%	正值比例: 11%-35%(2007-2011 年) 82%-88%(2012-2015 年) 平均强度:2%-4%	正值比例:35%-57% 平均强度:2%-6%	正值比例:约 63% 平均强度:3.5%-10%
研发/技术人员情况	-	正值比例:约 86% 平均强度:约 12.7%	正值比例:约 43% 平均强度:约 9.5%	正值比例:约 95% 平均强度:约 12%
新产品情况	正值比例:6%-11% 平均强度:33%-44%	-	正值比例:32%-44% 平均强度:24%-38%	正值比例:约 57% 平均强度:约 46%
专利情况	-	正值比例:42%-70% 平均申请数:25-50 平均授权数:22-34	正值比例:约 25%	正值比例:23%-36% 平均持有数:4-9

首先,查看研发支出的情况,可以发现:基于 2012 年后上市公司充分披露的研发支出数据,具有研发支出的企业比例高于 80%;而平均规模小得多的世界银行企业调查和中国私营企业调查则显示,具有研发支出的企业比例大致在 35%-63% 之间。不过,对于企业调查数据,本文有两点需要注意:(1)由于调查往往涉及腐败、寻租、逃税等敏感问题,或企业担心商业机密泄露,多数企业会选择拒访——以世界银行 2012 年中国企业调查为例,其执行报告显示企业的调查应答率仅约 20%。如果愿意接受调查的企业是管理者较为开明、信息相对透明、发展前景较好的企业,而这些企业也进行更为活跃的研发创新,则调查过程就可能引入了样本选择偏误;(2)对于"研发创新"这些意义积极且判断标准较为主观的问题,受访企业往往会给出夸大的回答。从而,世界银行企业调查和私营企业调查都可能会高估中国企业开展研发创新活动的比例,对其解读需要谨慎。但至少这两个数据来源中开展研发创新活动的企业比例低于上市公司,仍符合逻辑。相比之下,常规年度的工业企业数据库令人吃惊:尽管平均企业规模仅次于上市公司样本,但研发支出为正的企业比例仅 10% 出头,远低于世界银行企业调查和中国私营企业调查中的比例,也低于 2004 年普查数据中的比例。综上,本文认为常规年度的工业企业数据库很可能存在严重的统计纰漏,大量研发支出实际为正的企业数据被误报为零,大幅低估了企业进行研发投入的比例。

其次,新产品产值变量也存在同样的问题:在企业平均规模较小的世界银行企业调查和中国私营企业调查中,新产品产值为正的企业比例在 32%-57% 之间;但企业平均规模更大的常规年度工业企业数据库中,仅约 10% 的企业为正。这说明,大量实际中产出了新产品的企业数据被误报为零,大幅低估了企业具有创新产出的比例。但在这三个数据来源的正值样本中,新产品产值占年产值或年销售额的比重却大致相当——这说明,工业企业数据库中研发创新变量的正值样本数据可能仍是可信的。

本文着重分析常规年度工业企业数据库中可能存在统计纰漏的意义在于:(1)在规范层面上,工业企业数据库是研发创新相关研究中最常用的数据来源,如果学者们将研发创新变量异常高的零值

比例误解为现实情况,很可能做出过于悲观的判断,无法针对真实情况给出“对症下药”的政策建议;(2)在实证层面上,如果变量的零值不真实,而是包含了大量实际为正、但被误报为零的样本,那么在将研发创新变量作为被解释变量时,合适的计量方法应当为 Truncation 模型与 MLE 估计,而非 Tobit 模型,更遑论线性模型与 OLS 估计①,下文将对合适计量方法的选择进行讨论。

三、合适计量模型与估计方法的讨论

前文利用工业企业数据库、上市公司数据、世界银行企业调查数据、中国私营企业调查数据呈现了中国制造业企业研发创新的基本事实。本文发现,由于研发创新变量往往具有离散选择、零值堆积与正值连续分布并存的混合形式、截断形式等分布特征,在实证研究中作为被解释变量时往往使模型扰动项违背关键经典假设,线性模型的 OLS 估计不一致、常用的统计检验失效,从而需要对合适计量方法的选择进行讨论。

前文利用工业企业数据库、上市公司数据、世界银行企业调查数据、中国私营企业调查数据呈现了中国制造业企业研发创新的基本事实。本文发现,由于研发创新变量往往具有离散选择、零值堆积与正值连续分布并存的混合形式、截断形式等分布特征,在实证研究中作为被解释变量时往往使模型扰动项违背关键经典假设,线性模型的 OLS 估计不一致、常用的统计检验失效,从而需要对合适计量方法的选择进行讨论。

(一)Probit 与 Logit 模型

在研究中,本文经常关注企业是否进行研发创新,此时被解释变量服从“0-1”两点分布。而对于线性概率模型(linear probability model) $y_i = x'_i\beta + \varepsilon_i$,OLS 估计不一致,这是因为:在 $P(y_i = 1 \mid x_i) = x'_i\beta + \varepsilon_i$ 中,β 估计的一致性取决于严格外生性条件 $E(\varepsilon_i \mid x_i) = 0$ 能否满足。但在线性概率模型中,要么 $\varepsilon_i = 1 - x'_i\beta$(当 $y_i = 1$),要么 $\varepsilon_i = -x'_i\beta$(当 $y_i = 0$),ε_i 与 x_i 明显相关,$E(\varepsilon_i \mid x_i) = 0$ 不成立,OLS 估计量不一致。而且,尽管 y_i 的实际取值只能为 0 或 1,线性模型得到的预测值 $\hat{y}_i$ 却可能小于 0 或大于 1。此外,线性概率模型还存在异方差问题($Var(\varepsilon_i) = Var(1 - x'_i\beta) = Var(-x'_i\beta) = Var(x'_i\beta) \neq c$),系数显著性检验失效。

因此,对于被解释变量服从“0-1”两点分布的情况,可考虑如下的分布概率:

$$P(y = 1 \mid x) = F(x,\beta) \qquad P(y = 0 \mid x) = 1 - F(x,\beta) \tag{1}$$

式(1)中的 $F(x,\beta)$ 称为连接函数(link function),如果选择随机变量的累积分布函数作为连接函数,就可以保证被解释变量取 0 或 1 的概率在[0,1]区间内。

1. Probit 模型与 MLE 估计

若选取标准正态分布 $N(0,1)$ 的累积分布函数作为连接函数,式(1)称为 Probit 模型,被解释变量取值为 1 的概率可以表示为:

$$P(y = 1 \mid x) = F(x,\beta) = \Phi(x'\beta) = \int_{-\infty}^{x'\beta} \varphi(t)\,dt \tag{2}$$

对于样本 $\{x_i, y_i\}_n$,观测值 (x_i, y_i) 的概率密度函数可以写为:

$$\begin{aligned} &f(y_i \mid x_i,\beta) = \Phi(x'\beta) \text{ if } y_i = 1 \\ &f(y_i \mid x_i,\beta) = 1 - \Phi(x'\beta) \text{ if } y_i = 0 \end{aligned} \tag{3}$$

---

① 本文梳理文献发现,大量基于工业企业数据库的实证研究使用了线性模型与 OLS 估计(如林炜,2013;李文贵和余明桂,2015 等)或 Tobit 模型(如周亚虹等,2012;杨洋等,2015 等),仅极少数的研究如董晓芳和袁燕(2014)对零值的成因进行分析并使用了 Truncation 模型与 MLE 估计。

可以紧凑地写为：

$$f(y_i \mid x_i,\beta) = [\Phi(x'\beta)]^{y_i}[1-\Phi(x'\beta)]^{1-y_i} \tag{4}$$

根据式(4)可以写出全样本(观测数为 n)的似然函数：

$$L(y_i,x_i) = \prod_{i=1}^{n} f(y_i \mid x_i) \tag{5}$$

从而，MLE 估计量 $\hat{\beta}_{MLE} = \arg\max L(y_i,x_i)$ 为一致估计量，证明从略。

2. Logit 模型与 MLE 估计

若选取 logistic 分布的累积分布函数作为连接函数，式(1)称为 Logit 模型，被解释变量取值为 1 的概率可以表示为：

$$P(y=1 \mid x) = F(x,\beta) = \Lambda(x'\beta) = \frac{\exp(x'\beta)}{1+\exp(x'\beta)} \tag{6}$$

对于样本 $\{x_i,y_i\}_n$，观测值 (x_i,y_i) 的概率密度函数可以紧凑地写为：

$$f(y_i \mid x_i,\beta) = [\Lambda(x'\beta)]^{y_i}[1-\Lambda(x'\beta)]^{1-y_i} \tag{7}$$

根据式(7)可以写出全样本(观测数为 n)的似然函数：

$$L(y_i,x_i) = \prod_{i=1}^{n} f(y_i \mid x_i) \tag{8}$$

同样，MLE 估计量 $\hat{\beta}_{MLE} = \text{argmax}L(y_i,x_i)$ 为一致估计量，证明从略。

3. Probit 与 Logit 模型的适用性分析①

当反映企业是否有研发创新行为的变量作为被解释变量时，如“有无研发支出”、“有无新产品产值”、“有无专利申请与授权”等，应使用 Probit 或 Logit 模型；具体而言，对应的是世界银行中国企业调查与中国私营企业调查数据在剔除“不清楚”与“拒绝作答”观测值后的样本，以及上市公司的专利数据(是否有专利申请与授权)、2012 年及以后的研发支出数据(是否有研发支出)。需要注意的是，本文不建议研究者根据常规年度的工业企业数据库中的研发支出、新产品产出变量以及 2012 年前制造业上市公司研发支出变量构造企业研发创新决策的“0-1”变量——这是因为，常规年度工业企业数据库中的研发创新变量很可能存在统计纰漏，而 2012 年前制造业上市公司的研发支出数据披露不充分；从而，在常规年度工业企业数据中研发创新变量数值为零、以及 2012 年前没有报告研发支出数据的制造业上市公司中，很可能包含了大量实际上进行了研发创新的企业，构造出来的离散选择变量不能准确地反映企业研发创新的广延边际决策。

(二) Tobit 模型

除研发创新的广延边际决策之外，本文往往还关心其强度边际决策，即企业的研发支出有多少(或研发支出强度有多大)、新产品产值占总产值或年销售额的比重有多大、企业申请与获得授权的专利有多少个等。对于广延边际决策为 0 的企业，强度边际决策自然为 0，这是企业基于自身目标函数与约束条件做最优化决策得到的“边角解”(corner solution)；对于广度边际决策为 1 的企业，强度边际决策为连续分布正值，从而形成零值堆积(pile-up)与正值连续分布并存的混合分布(mixed distribution)。本文第二部分揭示，在多个数据来源中都有相当比例的企业不进行研发创新，零值比例不可忽视(non-trivial)。此时，y_i 的概率分布不连续，称为归并数据(censored data)。当以归并变量为线性模型 $y_i = x'_i\beta + \varepsilon_i$ 的被解释变量，无论使用正值子样本或全样本数据，OLS 估计均不一致，以下进行证明：

1. 分布设定

假设 $y_i^* = x'_i\beta + \varepsilon_i$，$y_i^*$ 为潜变量(latent variable)，代表企业不可观测的最优研发创新投入或产出

① 在 STATA 中，Probit 与 Logit 模型的命令分别为 *probit* 和 *logit*，面板数据命令分别为 xtprobit 和 xtlogit；此外，*Probit* 模型还可以结合工具变量法处理内生变量的问题，命令为 ivprobit。

量,假设扰动项 $\varepsilon_i \mid x_i \sim N(0,\sigma^2)$。当企业的最优研发创新量 $y_i^* > 0$ 时,本文可以观测到"现实"的研发创新投入或产出 $y_i = y_i^*$;当企业的最优研发创新量 $y_i^* \leq 0$ 时,本文只能观测到边角解 $y_i = 0$。

2. 被解释变量大于 0 的子样本估计

对于研发创新量为正的子样本而言,被解释变量的条件期望为:

$$\begin{aligned} E(y_i \mid x_i; y_i > 0) &= E(y_i^* \mid x_i; y_i^* > 0) = E(x'_i\beta + \varepsilon_i \mid x_i; x'_i\beta + \varepsilon_i > 0) \\ &= x'_i\beta + E(\varepsilon_i \mid x_i; x'_i\beta + \varepsilon_i > 0) = x'_i\beta + E(\varepsilon_i \mid x_i; \varepsilon_i > -x'_i\beta) \\ &= x'_i\beta + \sigma \times \lambda(-x'_i\beta/\sigma) \end{aligned} \tag{9}$$

其中,$\lambda(\cdot)$ 为"逆米尔斯比率"(inverse Mill's ratio,定义为 $\lambda(\cdot) = \frac{\varphi(\cdot)}{1-\Phi(\cdot)}$,$\varphi(\cdot)$ 与 $\Phi(\cdot)$ 分别为标准正态分布的概率密度函数与累积分布函数)。对子样本进行 OLS 估计:

$$y_i = x'_i\hat{\beta}_{OLS} + u_i \tag{10}$$

在这一线性模型中,式(9)中的非线性项 $\sigma \times \lambda(-x'_i\beta/\sigma)$ 被纳入扰动项 u_i 中,导致扰动项与解释变量集 x_i 相关,不满足严格外生性条件 $E(\varepsilon_i \mid x_i) = 0$,故 OLS 估计不一致。

3. 全样本估计

对于全样本而言,被解释变量的条件期望为:

$$\begin{aligned} E(y_i \mid x_i) &= 0 \times P(y_i = 0 \mid x_i) + E(y_i \mid x_i; y_i > 0) \times P(y_i > 0 \mid x_i) \\ &= E(y_i \mid x_i; y_i > 0) \times P(y_i > 0 \mid x_i) \end{aligned} \tag{11}$$

其中,$E(y_i \mid x_i; y_i > 0)$ 已由式(9)给出。进而,计算被解释变量为正的概率:

$$\begin{aligned} P(y_i > 0 \mid x_i) &= P(y_i^* > 0 \mid x_i) = P(x'_i\beta + \varepsilon_i > 0 \mid x_i) = P(\varepsilon_i > -x'_i\beta \mid x_i) \\ &= P\left(\frac{\varepsilon_i}{\sigma} > \frac{-x'_i\beta}{\sigma} \mid x_i\right) = 1 - \Phi\left(\frac{-x'_i\beta}{\sigma}\right) = \Phi\left(\frac{x'_i\beta}{\sigma}\right) \end{aligned} \tag{12}$$

从而,被解释变量的条件期望可最终写为:

$$E(y_i \mid x_i) = \Phi\left(\frac{x'_i\beta}{\sigma}\right) E(y_i \mid x_i; y_i > 0) = \Phi\left(\frac{x'_i\beta}{\sigma}\right)[x'_i\beta + \sigma \times \lambda(-x'_i\beta/\sigma)] \tag{13}$$

可以发现,式(13)是一个关于被解释变量 x_i 的非线性函数,基于线性模型的 OLS 估计无法得到一致的系数估计量。

4. Tobit 模型与 MLE 估计

Tobin(1958)提出了归并数据的 MLE 估计法,该方法被称为 Tobit 模型或归并回归(censored regression)。对于一个离散点与连续分布组成的混合分布,概率密度函数为:

$$\begin{aligned} f(y_i \mid x_i) &= f_1(y_i \mid x_i) 1(y_i > 0) \times P(y_i = 0 \mid x_i) 1(y_i = 0) \\ &= \left[\frac{1}{\sigma}\varphi\left(\frac{y_i - x'_i\beta}{\sigma}\right)\right] 1(y_i > 0) \times \left[1 - \Phi\left(\frac{x'_i\beta}{\sigma}\right)\right] 1(y_i = 0) \end{aligned} \tag{14}$$

其中,$1(\cdot)$ 为示性函数(indicator function),括号内的条件成立时函数值为 1,反之则取值为 0。当 $y_i = y_i^* > 0$ 时,被解释变量的条件概率密度函数为:$f_1(y_i \mid x_i) = \frac{1}{\sigma}\varphi(\frac{y_i - x'_i\beta}{\sigma})$;当 $y_i^* \leq 0$ 时 y_i 集聚于 $c = 0$ 上,该点概率为:$P(y_i = 0 \mid x_i) = 1 - P(y_i > 0 \mid x_i) = 1 - \Phi(\frac{x'_i\beta}{\sigma})$。

根据式(14)这一概率密度函数可以写出全样本的似然函数:

$$L(y_i, x_i) = \prod_{i=1}^{n} f(y_i \mid x_i) \tag{15}$$

从而,MLE 估计量 $\hat{\beta}_{MLE} = \text{argmax} L(y_i, x_i)$ 为一致估计量,证明从略。

5. Tobit 模型的适用性分析①

2004 年的工业企业数据库(普查数据)、世界银行中国企业调查数据、中国私营企业调查数据、上市公司的专利申请与授权数据、以及 2011 年以后充分披露的研发支出数据包含了企业研发创新强度边际信息,相应的变量数据服从“零值堆积与正值连续分布共存”的混合分布,当这些变量作为被解释变量时,本文应使用 Tobit 模型进行回归估计。在实际应用中,为削弱被解释变量数值过大引致的异方差和估计系数过小等问题的干扰,本文往往对原始数据加一后取自然对数作为被解释变量(即 $\ln(1+y_i)$),原数据中处于归并点 0 的数据经过变换后仍等于 0,新生成的混合分布仍以 0 为归并点(但正值分布更为集中,更接近正态分布的“钟形”特征),Tobit 模型仍然适用。但此时估计的条件期望函数不再是 $E(y_i \mid x_i)$,而是 $E[\ln(1+y_i) \mid x_i]$,对系数的解读需要更为小心。

但对于常规年度工业企业数据库中的研发支出与新产品产出数据而言,尽管这两个变量看似也服从“离散点堆积与连续正值分布共存”的混合分布,但严重的统计纰漏使得零值部分包含了大量实际为正、但被误报为零的观测值,故 Tobit 模型不适用,下一小节将对这些数据的处理与合适的计量模型进行讨论。

(三) Truncation 模型

除前述的归并数据外,受限被解释变量的另一种情况是,只有当 $y_i \geq c$(左侧截断,left truncation)或 $y_i \leq c$(右侧截断,right truncation)时,(y_i, x_i) 才能被观测到。基于截断样本与线性模型进行关于全样本信息的研究,OLS 估计不一致,证明如下:

1. 分布设定

为简化分析,本文仅对左侧截断的情形进行讨论。假设有线性模型 $y_i = x'_i\beta + \varepsilon_i$,扰动项 $\varepsilon_i \mid x_i \sim N(0,\sigma^2)$,故有 $y_i \mid x_i \sim N(x'_i\beta,\sigma^2)$;基于 $y_i \geq c$ 本文得到截断样本 $\{y_i, x_i\}_n$。

2. 样本条件期望函数的估计

被解释变量的条件期望可以写为:

$$\begin{aligned} E(y_i \mid x_i; y_i \geq c) &= E(x'_i\beta + \varepsilon_i \mid x_i; x'_i\beta + \varepsilon_i \geq c) = x'_i\beta + E(\varepsilon_i \mid x_i; x'_i\beta + \varepsilon_i \geq c) \\ &= x'_i\beta + \sigma \times \lambda((c - x'_i\beta)/\sigma) \end{aligned} \tag{16}$$

与 Tobit 模型相似,上式中的 $\lambda(\cdot)$ 为“逆米尔斯比率”。如果对式(16)进行 OLS 估计 $y_i = x'_i\hat{\beta}_{OLS} + u_i$,非线性项 $\sigma \times \lambda((c - x'_i\beta)/\sigma)$ 将被纳入扰动项 u_i 中,导致扰动项与解释变量集 x_i 相关,不满足严格外生性条件 $E(\varepsilon_i \mid x_i) = 0$,OLS 估计不一致。

3. Truncation 模型与 MLE 估计

如果能够写出截断样本的概率密度函数,本文便能够使用 MLE 方法对总体参数进行估计。基于 $y_i \mid x_i \sim N(x'_i\beta,\sigma^2)$,样本能够被观测到的概率为:

$$\begin{aligned} P(y_i \geq c \mid x_i) &= 1 - P(y_i < c \mid x_i) \\ &= 1 - P\left(\frac{y_i - x'_i\beta}{\sigma} < \frac{c - x'_i\beta}{\sigma} \mid x_i\right) = 1 - \Phi\left(\frac{c - x'_i\beta}{\sigma}\right) \end{aligned} \tag{17}$$

而全样本的概率密度函数为:

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2}\left(\frac{y_i - x'_i\beta}{\sigma}\right)^2\right] = \frac{1}{\sigma}\varphi\left(\frac{y_i - x'_i\beta}{\sigma}\right) \tag{18}$$

从而,截断样本的概率密度函数为:

---

① 在 STATA 中,Tobit 模型的命令为 *tobit*,面板数据命令为 *xttobit*;Tobit 模型还可以结合工具变量法处理内生解释变量的问题,命令为 *ivtobit*。

$$f(y_i \mid y_i \geq c, x_i) = \frac{f(y_i)}{P(y_i \geq c \mid x_i)} = \frac{\varphi[(y_i - x'_i\beta)/\sigma]}{\sigma\{1 - \Phi[(c - x'_i\beta)/\sigma]\}} \tag{19}$$

进而,可以写出截断样本的似然函数:

$$L(y_i \mid y_i \geq c, x_i) = \prod_{i=1}^{n} f(y_i \mid y_i \geq c, x_i) \tag{20}$$

从而,MLE 估计量 $\hat{\beta}_{MLE} = \text{argmax} L(y_i \mid y_i \geq c, x_i)$ 为一致估计量,证明从略。

4. Truncation 模型的适用性分析①

在使用 2007-2011 年间制造业上市公司的研发支出数据作为被解释变量时,由于数据披露不完全,呈现典型的截断分布特征,应使用 Truncation 模型进行研究。但由于有效样本量较小,回归可能出现参数估计不稳定等问题;此外,披露了研发支出数据的企业可能不是总体的代表性抽样,使用这些数据还可能引入样本选择问题。因此本文认为,2012 年前制造业上市公司的研发支出数据在实证研究中应谨慎使用。

工业企业数据库的情况也需要讨论:由于统计纰漏,常规年度工业企业数据库中研发创新变量的零值很可能有相当比例是数值实际为正的企业。本文建议,在使用这套数据进行研究时,应剔除零值部分,运用 Truncation 模型进行估计——这并非因为研发创新变量数值为零的样本不可观测(即典型的截断情形),而是因为零值数据不可信,从而应主动放弃这部分样本,将数据可信度较高的正值样本视为一个截断样本进行处理。

Tobit 模型与 Truncation 模型的另一关键区别是,前者利用了全样本信息,后者仅利用了被解释变量可观测的企业信息——以常规年度的工业企业数据库为例,这意味着失去约 90% 的样本。但所幸工业企业数据库的样本量极大,即便仅保留研发创新变量为正的企业,仍有数以万计的观测值,本文无需担心小样本回归的效率低、参数不稳定等问题。

(四)不同数据来源中研发创新变量合适计量方法总结

基于前边三个小节的讨论,本文将工业企业数据库、上市公司数据、世界银行企业调查数据、私营企业调查数据中研发创新变量所对应的合适计量方法总结如下:

表 18　　各数据来源中研发创新变量合适计量方法总结

数据来源 / 研发创新变量	工业企业数据库(常规年度)	上市公司数据	世界银行中国企业调查数据	中国私营企业调查数据
是否有研发支出	不适用	2012 年之前:不适用 2012 年及之后:Probit/Logit	Probit/Logit	Probit/Logit
研发支出数量/强度	Truncation	2012 年之前:Truncation 2012 年及之后:Tobit	Tobit	Tobit
是否有新产品	不适用	无相关信息	Probit/Logit	Probit/Logit
新产品产值/销售额/比重	Truncation	无相关信息	Tobit	Tobit
是否有专利	无相关信息	Probit/Logit	Probit/Logit	Probit/Logit
专利数量②	无相关信息	Tobit	Tobit	Tobit

① 在 STATA 中,Truncation 模型的命令为 *truncreg*。

② 尽管专利数据的原始数值是非负整数,但由于其取值众多,故可将数据加 1 后取自然对数视为连续变量结合 Tobit 模型进行研究。面板数据固定效应 Tobit 模型的相关内容可见于 Honore(1992)与 STATA 非官方命令 *pantob*。如果直接使用专利的原始数值进行研究,应考虑泊松回归、负二项回归等计数模型,详见 Wooldridge(2010)相关章节的讨论。但这些模型的 MLE 估计往往收敛较慢或难以收敛,且在运用于面板数据时控制个体固定效应将丢弃被解释变量观测值在样本期一直为零的个体,应谨慎使用。

四、结论与启示

在中国经济增长出现结构性减速的大背景下,中国企业研发创新的现状如何、症结何在是关系到“新常态经济”下中国能否顺利切换至以技术进步为核心驱动力的发展模式,以及如何为“以创新促发展,以改革促创新”制定相关政策的关键问题。当前,学术界对企业研发创新的研究方兴未艾,本文对研究最为常用的多个数据来源中制造业企业的研发创新变量进行梳理总结,详细解析不同数据来源中的一致与冲突,尝试破除既有文献与相关媒体报道中的常见误区,为相关研究提供基础信息支持与计量方法参考。

本文得到双重维度的发现:在规范层面上,本文发现,在 1998-2015 年间,中国企业的研发创新状况与发展态势较为乐观,研发支出、新产品产出、专利申请与授权等指标在广延边际上持续拓宽、在强度边际上稳定进步,基本领先于其他经济发展水平相近的新兴发展中国家;常规年度工业企业数据库呈现研发创新变量为正的企业比例仅约 10% 的“创新荒漠”很可能是统计纰漏造成的假象。在技术层面上,本文证明,当研发创新变量作为被解释变量时,离散、混合、截断分布特征将使线性模型的 OLS 估计不一致,故需要选取 Probit 与 Logit、Tobit、Truncation 模型与 MLE 方法才能得到可信的实证研究结论。

在政府将鼓励研发创新与推动技术进步上升至“国策”高度的时代背景下,中国企业的研发投入可能已不存在“总量不足”问题,关键症结在于不同所有制间存在明显的创新效率差异与资源错配。本文发现,国有企业的研发创新效率明显偏低;同时,在非上市企业群体中,大量的政府补贴与金融资源偏向了国有企业,资源瓶颈明显制约了私营企业研发创新的发展。而在上市公司群体中,资源错配问题得到一定的缓解,但在当前多层次市场发展不成熟、机制设计尚不完善、投机炒作屡见不鲜的资本市场中,如何激励产业资本持续投向研发创新、避免“脱实向虚”也是亟待解决的难题。对这些重要问题的回答与相关政策的制定都将是未来深入研究的方向。

参考文献:

[1] Almeida R. and A. M. Fernandes, 2008, “Openness and Technological Innovations in Developing Countries: Evidence from Firm-level Surveys,” *The Journal of Development Studies*, 44(5): 701-727.

[2] Hart O., A. Shleifer and R. W. Vishny, 1997, “The Proper Scope of Government: Theory and an Application to Prisons,” *Quarterly Journal of Economics*, 112(4): 1127-1161.

[3] Honore B. E., 1992, “Trimmed LAD and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects,” *Econometrica*, 60(3): 533-565.

[4] Koh P., and D. Reeb, 2015, “Missing R&D,” *Journal of Accounting and Economics*, 60(1): 73-94.

[5] Shleifer A., 1998, “State versus Private Ownership,” *Journal of Economic Perspectives*, 12(4): 133-150.

[6] Tobin J., 1958, “Estimation of Relationships for Limited Dependent Variables,” *Econometrica*, 26(1): 24-36.

[7] Wei S., Z. Xie and X. Zhang, 2017, “From ‘Made in China’ to ‘Innovated in China’: Necessity, Prospect, and Challenges,” *Journal of Economic Perspectives*, 31(1): 49-70.

[8] Wooldridge J. M., 2010, Econometric Analysis of Cross-Section and Panel Data. Cambridge MA: MIT Press.

[9] Xie Z. and X. Zhang, 2015, “The Patterns of Patents in China,” *China Economic Journal*, 8(2): 122-142.

[10] 董晓芳、袁燕,2014,“企业创新、生命周期与聚集经济”,《经济学(季刊)》,第 1 期,第 767-792 页。

[11] 李文贵、余明桂,2015,“民营化企业的股权结构与企业创新”,《管理世界》,第 4 期,第 112-125 页。

[12] 林炜,2013,“企业创新激励:来自中国劳动力成本上升的解释”,《管理世界》,第 10 期:第 95-105 页。

[13] 龙小宁、王俊,2015,“中国专利激增的动因及其质量效应”,《世界经济》,第 6 期,第 115-142 页。

[14] 聂辉华、江艇、杨汝岱,2012,“中国工业企业数据库的使用现状和潜在问题”,《世界经济》,第 5 期,第 142-158 页。

[15] 聂辉华、谭松涛、王宇峰,2008,“创新、企业规模和市场竞争”,《世界经济》,第 7 期,第 57-66 页。

[16] 杨国超、刘静、廉鹏、芮萌,2017,“减税激励、研发操纵与研发绩效”,《经济研究》,第 8 期,第 110-124 页。

[17] 杨洋、魏江、罗来军,2015,“谁在利用政府补贴进行创新?——所有制和要素市场扭曲的联合调节效应”,《管理世界》,第 1 期,第 75-86 页。
[18] 周亚虹、贺小丹、沈瑶,2012,“中国工业企业自主创新的影响因素和产出绩效研究”,《经济研究》,第 5 期,第 107-119 页。

R&D and Innovations of Chinese Manufacturing Firms: Basic Facts, Common Misunderstandings, and Discussions on the Appropriate Choice of Econometric Methods

LONG Xiaoning, LIN Zhifan
Xiamen University, Xiamen, 361005

Abstract: This paper attempts to provide background information and econometric reference by summarizing on four of the most popular data sources: the Chinese industrial enterprise database, listed company database, World Bank enterprise survey data, and Chinese private enterprise survey data. Main findings include: 1. Only about 10 percent of firms in the Chinese industrial enterprise database (regular years) report positive R&D input and new product output, but this is probably due to statistical flaws which leads to the underestimation of the proportion of innovation-active firms. Cross-country comparisons indicate that China leads other developing countries in innovations; 2. More than eighty percent of listed manufacturing companies engage in R&D with increasing intensity, and they are experiencing a patent boom with improving quality; 3. The discrete, mixed, and truncated distributions of the innovation-related variables would result in inconsistent OLS estimates when they are the dependent variable of linear models. Probit and Logit, Tobit, Truncation models and MLE are proposed as feasible solutions.

Key Words: R&D and innovation; Chinese industrial enterprise database; listed company database; WorldBank enterprise survey data; Chinese private enterprise survey data; choice of econometric model

〔责任编辑:柏培文〕〔校对:张相伟〕