
金融基础设施激活数据要素潜能： 来自央行信用报告的证据

周 博 梁蕴佳*

内容提要 金融基础设施作为我国数据基础设施制度建设的一部分,是畅通数据资源大循环的重要因素。本文以征信系统这一金融基础设施为例,基于拍拍贷金融科技平台 2015–2016 年发生的 285 603 笔月度信贷交易数据,利用该平台于 2015 年 4 月 27 日开通征信认证(即允许借款人向平台提供央行信用报告)的外生冲击构造断点回归,考察征信数据要素的经济价值及其作用机制。研究发现,该平台开通征信认证显著缓解了借贷行为中的逆向选择问题,进一步使该平台的借款利率下降了 0.59 个百分点,同时借款金额提高了 15.24%,从而促进了信贷资金的配置效率。福利分析发现,征信数据要素可得性增加使社会福利水平提升了 0.0927%,并实现了价值 1614 万元的福利改进。本研究为统筹数据资源整合共享和开发利用以及推进数字中国建设提供了重要的政策启示。

关键词 信息不对称 断点回归 充分统计量 福利分析

一 引言

做强做优做大数字经济是我国经济发展的新动能。2022 年年底,《中共中央 国务院关于构建数据基础制度更好发挥数据要素作用的意见》指出“充分发挥我国海量

* 周博:北京大学经济学院;梁蕴佳(通讯作者):四川大学经济学院 成都市一环路南一段 24 号 610065;电子信箱:sczhoubo@pku.edu.cn(周博);yunjia_liang@163.com(梁蕴佳)。

作者感谢四川省哲学社会科学基金青年项目(SCJJ24ND306、SCJJ24ND232)的资助,感谢杜鹏程和匿名审稿专家的宝贵意见。当然,文责自负。

数据规模和丰富应用场景优势”^①。要激活数据要素潜能首先需要弄清什么是制约我国发挥数据规模和应用场景优势的关键因素,这样才能有针对性地采取相应的政策措施。在畅通数据资源大循环上,我国需解决数据发展和数据安全相互协调的问题。一方面,数据发展要求各种数据资源的开放和共享。但另一方面,数据资源的开放共享又面临数据安全和数据隐私泄露的担忧。如何在保障数据安全的前提下实现数据发展、开放和共享是当前的政策难题。对此,中共中央、国务院在2023年2月印发的《数字中国建设整体布局规划》提出“构建国家数据管理体制机制,健全各级数据统筹管理机构”这一解决思路。随后的《党和国家机构改革方案》提出“组建国家数据局”,即政府机构统筹数据资源整合共享和开发利用,由此在保证数据安全的基础上充分发挥我国海量数据的规模优势。

实际上,早在2003年由中共中央和国务院授权中国人民银行建立的覆盖全社会的征信体系,就是一次国家数据管理体制机制的创新试验。我国征信体系作为金融基础设施^②的重要组成部分(李稻葵等,2016),由中国人民银行征信中心建设、运行和维护,并负责为各类金融活动提供基础性公共服务。具体地,征信系统通过采集、整理、保存、加工个人的基本信息、信贷信息和反映其信用状况的其他信息,建立个人信用信息共享机制,有效解决金融交易中的信息不对称问题,从而防范金融风险、推动信贷市场发展和支持实体经济高质量发展。目前,我国征信系统已成为全球最大的信用信息数据库。根据中国人民银行征信管理局数据显示,截至2024年3月末,个人征信系统接入商业银行等各类从事信贷业务的法人机构6124家,累计收录11.6亿自然人信息^③,2024年全年提供信用报告查询服务67亿次^④。那么,如此庞大的征信体系作为数字经济时代下我国重要的金融基础设施,如何激活其蕴含的数据要素潜能,从而有效解决金融市场上的信息不对称问题?其带来的经济价值又有多大?这正是本文尝试回答的问题。

本文基于拍拍贷金融科技平台2015–2016年发生的285 603笔日度信贷交易数据,利用该平台于2015年4月27日开通征信认证功能(即允许借款人向平台提供其

① 资料来自 https://www.gov.cn/zhengce/202212/content_6720768.htm。

② 国家发展和改革委员会针对《中华人民共和国国民经济和社会发展第十四个五年规划和2035年远景目标纲要》名词解释中提到,金融基础设施指为各类金融活动提供基础性公共服务的系统和制度等,基础征信系统则属于我国6类金融基础设施之一。资料来自 https://www.ndrc.gov.cn/fggz/fzzlgh/gjfgzh/202112/t20211224_1309524_ext.html。

③ 数据来自 https://www.gov.cn/zhengce/202404/content_6944522.htm。

④ 数据来自 https://www.gov.cn/lianbo/bumen/202504/content_7016963.htm。

央行信用报告)这一外生冲击构造断点回归设计,探究征信报告可得性能否解决信息不对称问题及其经济效应。研究发现:第一,该平台开通征信认证功能增加了征信大数据的可得性,为该平台在贷前审核阶段提供了重要的数据要素,使得每笔贷款的初始信用等级显著提高了0.31个单位,这在一定程度上解决了事前信息不对称的问题。但由于开通征信认证功能这一冲击本身并没有改变借款人事后违约成本以及按时还款收益,因而每笔贷款的事后违约率没有显著变化,这表明在本文研究场景下征信大数据能缓解逆向选择问题,但不能解决道德风险问题。第二,在开通征信认证后,该平台的借款利率下降了0.59个百分点,同时借款金额提高了15.24%,这表明征信报告可得性提高有利于贷方识别和评估借款人风险以收取更符合借款人真实风险水平的利率并增加了信贷供给,从而促进了信贷资金的配置效率。第三,本文进一步基于Jansen *et al.* (2025)的框架构建充分统计量进行福利分析,发现征信大数据可得性提升因缓解逆向选择问题而使社会福利改进了0.0927%。从社会福利价值看,该金融科技平台开通征信认证后,在2015年4月27日到2016年12月31日这一时间段累计提高了社会福利1614万元,这体现了征信数据要素的经济价值。

同现有文献相比,本文边际贡献主要体现在以下两方面:第一,本文通过利用金融科技平台开通征信认证功能这一外生冲击构造断点回归设计的识别策略,有效克服了内生性问题,较好地检验了征信数据要素可得性提高能缓解借贷双方的事前信息不对称问题。这表明征信报告等大数据要素是降低信贷市场上信息不对称的重要金融基础设施工具。本研究结论为当前我国数字经济时代背景下如何激活大数据要素以解决信贷市场上的信息不对称问题提供了重要启示,拓展了解决信息不对称问题的研究思路。

第二,本文从经验研究上首次估计了数据要素的经济价值,这是对现有文献的重要补充。现有文献讨论了大数据作为一种新型生产要素能直接参与企业生产活动并促进经济增长(杨俊等,2022),相关研究还表明大数据可得性的提高能促进企业投资(Begenau *et al.*, 2018; Farboodi *et al.*, 2019)新产品的开发和创新(Beraja *et al.*, 2023; 田利辉等, 2025)。另一支文献研究了大数据作为重要的信息中介有利于加大信贷市场的竞争(He *et al.*, 2023),并提高市场运行效率(Farboodi and Veldkamp, 2020; 曹光宇等, 2022)。但目前鲜有文章直接测量大数据的经济价值(李稻葵等, 2016)。而本文量化估计了征信大数据可得性提高在信贷市场上产生的经济价值,结果显示其创造了1614万元的福利改进,这为进一步发展和支持大数据产业提供了经济学的成本收益分析框架。

本文后续安排为:第二部分对相关文献进行回顾并梳理政策背景;第三部分介绍本文数据;第四部分构建本文计量模型并报告基准回归结果;第五部分进一步分析征信认证开通对借款合同产生的影响;第六部分基于 Jansen *et al.* (2025) 的模型估计福利改善的大小;最后是本文主要结论与政策启示。

二 文献综述与政策背景

(一)文献综述

信息不对称理论由 Arrow (1963) 首次提出,他认为信息不对称会使完全竞争市场的假设不成立。后续研究从理论和经验上证明了信息不对称在二手车市场 (Akerlof, 1970)、劳动力市场 (Michael, 1973)、保险市场 (Rothschild and Stiglitz, 1976; Finkelstein and Poterba, 2004; Einav and Finkelstein, 2018) 和信贷市场 (Stiglitz and Weiss, 1981; Karlan and Zinman, 2009; Stroebel, 2016; Gupta and Hansman, 2022) 中的存在性。

信息不对称的存在还会导致市场失灵和福利损失 (Stiglitz and Weiss, 1981)。例如贷方因无法区分出高风险和低风险借款人而不能针对不同风险水平的借款人进行差别定价,由此导致了低效率的信贷配给和福利损失 (Dell'Ariccia and Marquez, 2004; Djankov *et al.*, 2007; 王正位等, 2020)。在经验研究上, Einav *et al.* (2010) 估计了保险市场上由于逆向选择导致的无效定价所产生的福利损失大小, DeFusco *et al.* (2022) 基于金融科技贷款机构的随机实验发现信息不对称会对均衡价格产生较大的扭曲,但整体福利损失不大。

如何解决信息不对称是经济学领域研究的核心问题之一。在信贷市场上, Stiglitz and Weiss (1981) 与 Chan and Thakor (1987) 认为提供抵押品能减少逆向选择问题,同时抵押品也有助于降低借款人的违约动机,从而降低信贷市场的道德风险 (Igawa and Kanatas, 1990; Menkhoff *et al.*, 2006)。因此,以抵押为主要形式的担保广泛运用于传统商业银行的贷款业务中 (尹志超和甘犁, 2011)。另有研究发现声誉机制可以降低道德风险, Bu and Liao (2022) 的乡村小额信贷随机实验表明,在村里公示借款人的违约信息有利于改善其还款表现; Liao *et al.* (2023) 基于互联网信贷平台的随机实验也发现征信违约的提醒可以显著降低用户的违约概率。

在如今的信息科技时代下,丰富的数据要素为解决信息不对称提供了新的工具。Pagano and Jappelli (1993) 指出信息的获取本身就是解决信息不对称的直接方法。例如,银行等金融机构之间共享借方的资信状况能显著缓解信息不对称并降低违约率

(Jappeli and Pagano, 2002; Doblas-Madrid and Minetti, 2013)。但如何收集数据并从海量数据中提取有价值的信息是发挥数据要素潜能的难点之一。在数据收集过程中,隐私保护和数据的可信性问题制约了私人部门进行数据采集。此外,相较于公开数据和调查数据,政府部门中积累的海量行政数据覆盖面更全、信息维度更多、质量也更可信(Beraja *et al.*, 2023; 方锦程等, 2023)。因此,由政府主导构建国家数据管理体制机制是实现海量数据资源统筹共享,从而充分利用我国数据规模优势的重要路径。

那么我国央行征信系统作为国家数据管理体制机制建设的关键尝试应如何激活其数据要素潜能以有效发挥其价值?目前鲜有研究对此进行回答。本文将基于金融科技平台开通征信认证这一事件冲击考察征信数据要素能否解决信息不对称问题及其创造的经济价值,对现有文献进行有益补充。

(二)政策背景

在保证数据安全的前提下充分发挥我国海量数据的规模优势,意味着由政府部门主导相应的数据体系建设是更高效的。2003年,党中央、国务院决定授权中国人民银行建立健全覆盖全社会的征信体系。2004年,中国人民银行开始组织商业银行建立全国集中统一的个人征信系统,并实现15家全国性商业银行和8家城市商业银行在全国7个城市的成功联网试运行。2005年8月底,个人征信系统完成与全国所有商业银行和部分有条件的农村信用联社的联网运行。根据《征信业管理条例》(国务院令 第631号)规定,从事信贷业务的机构应按规定向金融信用信息基础数据库(即央行征信系统)提供信贷信息。央行征信系统对各授信机构采集的个人信贷交易信息以及与个人信贷相关的个人基本信息、抵质押物信息等进行汇总、加工、整合形成个人信用信息基础数据库和基础核心产品个人信用报告。在过去的十几年,作为我国金融市场重要数据要素供给之一的征信体系蓬勃发展,目前已成为世界上规模最大、覆盖人口最多、收集信贷信息种类最全的征信系统。

个人信用报告^①作为个人征信系统的基础核心产品详细记录了个人信息主体的信贷历史交易信息,以及当前负债、逾期历史、电信、水电气等后付费业务欠费信息和欠税记录、民事判决记录、强制执行记录、行政处罚记录等负面历史信息,是反映个人信用行为的“经济身份证”。个人信用报告主要包括8个部分:个人基本信息、信息概要、信贷交易信息明细、非信贷交易信息明细、公共信息明细、本人声明、异议标注及查询记录,同时标注了对应的信息来源。个人信用报告中的不良信息自不良行为或

① 限于篇幅,个人信用报告的样例未展示,详见本刊网站(www.jweonline.cn)本文补充材料附录A。

事件终止之日起保留5年,正面信息是个人信息主体的信用财富,并一直展示在信用报告中。

自2006年中国人民银行个人征信系统在全国正式联网运行开始,所有持牌金融机构均可以免费访问,但金融科技公司不能访问该系统。而自2013年以来,金融科技消费信贷市场快速发展,2017年其贷款额已达到2.8万亿元人民币^①。为了防范系统性金融风险,原中国人民银行和中国银行保险监督管理委员会开始对金融科技平台实施严格监管,2019年9月起要求金融科技平台全面纳入央行征信系统^②,即各平台应向央行征信系统依法合规报送该平台的所有信贷信息。与此同时,金融科技平台也开始可以访问央行个人征信系统,从而实现信用信息共享。但在本文研究区间内(2015–2016年),金融科技平台均无法访问央行征信系统,即无法获取个人征信报告的相关信息。

央行征信系统的访问受限极大限制了金融科技平台的发展。现有文献指出机器学习模型等算法技术有利于提高金融科技平台筛选和监督借款人的能力与效率(Buchak *et al.*, 2018; Fuster *et al.*, 2019; Di Maggio and Yao, 2021; 黄益平和邱晗, 2021),但如果传统银行拥有金融科技贷款机构无法访问的客户数据如央行征信系统,即使金融科技贷款机构拥有更好的数据处理算法,其筛选借款人的准确性仍会较低(He *et al.*, 2023)。因此,为获取央行征信系统提供的个人信用报告这一重要的数据要素,本文研究的拍拍贷金融科技平台(后文简称A平台)于2015年4月27日开通了征信认证功能,即用户在申请贷款阶段可以自主选择上传由中国人民银行征信中心开具的个人信用报告,这为该平台在贷前风险评估阶段提供了大量关于借款人的征信数据。

A平台是我国首家纯信用无担保线上贷款平台。基于大数据、云计算和人工智能等科技的快速发展,A平台采用了先进的大数据风控系统对每一笔借款标的进行风险评估,以根据风险评级对每笔借款形成定价,保证收益和风险相匹配。为更加精准地评估每个借款标的的风险概率,A平台在2015年4月27日开通了征信认证功能,从而把个人信用报告中涵盖的多渠道、多维度的用户数据纳入风控算法模型。至此,A平台的大数据风控系统使用的信息维度可分为3类:一是用户的基本信息,如性别、年龄、学历等;二是用户在该平台的历史借款和还款记录以及登录该平台的各种行为

① 数据来自 <https://www.brookings.edu/articles/whats-happening-with-chinas-fintech-industry/>。

② 资料来自 <http://credit.tianshui.gov.cn/118/14157.html>。

数据等;三是用户自主上传的个人信用报告中的各种信用情况,如个人当前负债、逾期历史、合同履行、后付费业务欠费、欠税记录、奖励记录等。这些海量数据信息为A平台有效评估借款人的风险评级提供了重要数据要素。

三 数据与描述性统计

本文主要使用的数据来自A平台2015年1月1日到2016年12月31日发生的日度信贷交易全样本数据的10%随机抽样,共包含285 603笔贷款数据。A平台于2007年在上海创立,是国内第一家由工商部门特批获得“金融信息服务”资质从而得到政府认可的互联网金融平台,并于2017年在美国纽约证券交易所上市。A平台是国内领先的金融科技公司,将集大数据、云计算和人工智能等科技为一体的大数据风控系统应用于贷前审批、风险定价、贷后风险管理等阶段,运用科技助力金融服务。截至2024年3月底,A平台累计注册用户数达1.5亿,累计放款金额约8377.6亿元。此外,A平台的借款人特征^①也与全国网贷市场的用户群体属性一致,该平台借款人在全国31个省(市、自治区)均有分布,主要为20-40岁的年轻男性,教育水平为大专学历的人群占比最高,职业类型多为工薪阶层。因此,本文使用的A平台信贷数据具有良好的样本代表性。

央行征信报告是金融科技平台对借款人进行风险评估的重要数据要素,自2019年9月监管部门要求金融科技平台全面纳入央行征信系统起,各平台陆续开始正式接入。A平台于2019年12月11日宣布其已获准接入中国人民银行征信中心,可访问并更新平台借款人在征信中心的信用记录。但在本文研究区间内(2015-2016年),A平台还无法访问央行征信系统,即无法获取个人征信报告的相关信息。为获取个人信用报告这一重要的征信数据要素,A平台在2015年4月27日开通了征信认证功能,即用户在申请贷款阶段可以自主选择上传由中国人民银行征信中心开具的个人信用报告。这意味着在2015年4月27日开通征信认证功能后,A平台虽然仍无法访问央行征信系统,但用户自主上传的由中国人民银行征信中心开具的个人信用报告为A平台提供了大量关于借款人的征信信息,这是之前其不具有的重要数据要素。

需要强调的两点事实是:第一,A平台并没有提前向用户公告将于2015年4月27

^① 本文描述了A平台借款人所在地区、受教育程度、收入水平和职业类型4个方面的特征,但限于篇幅未在正文中报告,详见本刊网站本文补充材料附录B。

日开通征信认证的消息,而是在2015年4月27日上线征信认证功能后向用户提示相关操作指南;第二,2015年4月27日就是A平台正式开通征信认证的日期,并不存在试运行阶段。因此,用户针对A平台在2015年4月27日开通征信认证这一事件并不存在提前反应的可能,所以该事件冲击日期相对于用户个体的贷款时间是外生的。正如图1所示,2015年4月27日之前A平台进行征信认证的贷款比例一直为0。4月27日在A平台开通征信认证功能后,该平台上传征信报告进行认证的贷款比例开始明显增加,且之后一直维持在6%左右的水平,提高了该平台获得借款人征信数据的可得性。在贷款审批阶段,A平台会读取借款人的个人信用报告内容,对多维度的信用信息进行分析、挖掘和整合,并将整合后的信用信息纳入信用评价算法模型计算用户的初始信用等级。因此,在开通征信认证功能后,A平台依据征信大数据能更准确地预测借款人的风险概率,且能进一步基于更准确的风控评级制定风险定价。

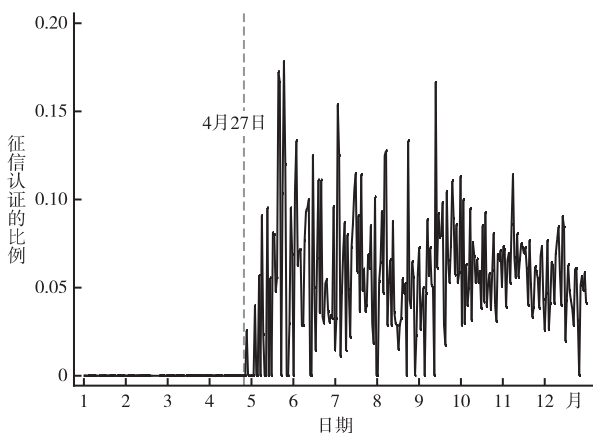


图1 2015年平台开通征信认证前后征信认证比例的变化情况

本文主要使用的每笔贷款数据的各项特征变量信息见表1。其中,初始信用等级有1-6共6个等级类别,数值越大表示初始信用等级越好。事后违约率是对每笔贷款进行追踪的结果变量,由于有少部分贷款在数据获取时还未进入还款阶段,故这部分样本的事后违约率是缺失的。在该平台申请贷款阶段,用户需要报告自己的性别、年龄、手机号、户口、脸部识别、教育学历等个人基本信息。需要说明的是,手机认证、户口认证、视频认证和教育认证这4项功能早在A平台成立初期就已经开通使用。此外,如果借款人为淘宝店主则其在贷款申请阶段还可以进行淘宝认证,需要注意的是,淘宝认证功能的开通时间大约是在征信认证功能开通的前1周。因此,为排除开

通淘宝认证功能对本文结果的干扰,在后文的稳健性检验部分我们将剔除借款人为淘宝店主这部分样本以排除淘宝认证的影响。但从表1可以看出,借款人为淘宝店主这部分样本占比很小,仅0.37%左右,因此淘宝认证功能开通对本文研究结果的影响并不大。

表1	描述性统计结果					
	变量	样本数	均值	方差	最小值	最大值
	征信认证	285 603	0.0311	0.1737	0	1
	初始信用等级	285 603	3.6420	0.8754	1	6
	事后违约率	279 687	0.1269	0.2429	0	1
	女性	285 603	0.3192	0.4662	0	1
	年龄	285 603	29.0439	6.6168	17	56
	手机认证	285 603	0.3974	0.4894	0	1
	户口认证	285 603	0.0333	0.1795	0	1
	视频认证	285 603	0.0608	0.2389	0	1
	教育认证	285 603	0.3505	0.4771	0	1
	淘宝店主	285 603	0.0037	0.0611	0	1
	历史成功借款次数(对数)	285 603	0.9026	0.7170	0	6.4770
	历史正常还款期数(对数)	285 603	1.6388	1.2482	0	7.8272
	借款利率(%)	285 603	20.5878	1.8132	6.5	24
	借款金额(对数)	285 603	8.0503	0.8249	4.6052	13.1224

说明:征信认证、女性、手机认证、户口认证、视频认证、教育认证及淘宝店主变量均为0-1变量,是为1,否则为0。

四 经验分析

(一)模型设定

本文研究目标是识别央行信用报告提供的征信数据要素是否有利于A平台缓解信息不对称问题,若简单比较上传信用报告(处理组)和不上传信用报告(对照组)的个体贷款特征的差异,则会存在较大的内生性问题,即信用资质状况更好的借款人选择上传个人信用报告的概率本身就更高。因此,为克服内生性问题,本文利用A平台开通征信认证功能这一外生冲击,采用断点回归设计(RDD)进行经验分析。虽然黄炜等(2023)指出年度数据在时间维度上的颗粒度较粗,能否适用于断点回归有待商

椎,但当数据在时间维度的颗粒度更细时则能很好地满足断点回归所需的驱动变量连续性假设。例如,比年度数据颗粒度更细且更高频的日度数据则能满足时间断点的使用条件。随着数据可得性的提高,越来越多的文献利用颗粒度更细的数据构造时间断点设计进行经验研究。周沂等(2021)基于日度空气质量数据使用时间断点回归发现,在中央生态环境保护督察组进驻时,被督查地的PM2.5浓度相较进驻前显著降低。Greenstone *et al.*(2022)同样基于日度空气污染数据构造时间断点回归,研究发现引入自动空气污染监测系统有利于空气污染的治理。马述忠等(2024)则基于详细的产品-周度层面数据,通过时间断点设计研究发现产品层面的质量认证因传递质量信号而显著缓解了商品交易中的信息不对称。基于此,本文参考上述文献的做法,基于日度信贷交易数据并使用时间断点回归的模型设定检验征信数据要素能否缓解信贷市场上的信息不对称问题。

A平台开通征信认证功能的准确日期为本文提供了理想的政策断点规则。从图1可知,2015年4月27日之前A平台进行征信认证的贷款比例一直为0,即在A平台开通征信认证功能之前,借款人上传信用报告的概率为0。在2015年4月27日A平台开通征信认证功能后,进行征信认证的贷款比例开始逐渐增加,即2015年4月27日之后借款人上传信用报告的概率远大于2015年4月27日之前。这一断点规则可以表示为:

$$Cut_i = s_i - Credit \quad (1)$$

$$Pr(Treat_i = 1 | Cut_i) = \begin{cases} g_1(Cut_i), & \text{如果 } Cut_i \geq 0 \\ g_0(Cut_i), & \text{如果 } Cut_i < 0 \end{cases}, 1 > g_1(Cut_i) > g_0(Cut_i) = 0 \quad (2)$$

其中, s_i 表示每笔贷款*i*合同签订日期; $Credit$ 表示A平台开通征信认证的日期即2015年4月27日; Cut_i 是驱动变量,表示该笔贷款合同签订日期发生于2015年4月27日之前或之后; $Treat_i$ 为处理变量,表示该笔贷款*i*是否进行征信认证,即借款人是否上传个人信用报告; $g_1(Cut_i)$ 表示发生于2015年4月27日及之后的贷款*i*进行征信认证的概率, $g_0(Cut_i)$ 表示发生于2015年4月27日之前的贷款*i*进行征信认证的概率,且 $1 > g_1(Cut_i) > g_0(Cut_i) = 0$, (2)式表明 $Pr(Treat_i = 1 | Cut_i)$ 是 Cut_i 的非连续函数,以 $Cut_i = 0$ 为临界点。

以上断点规则符合断点回归的设计原理,因此本文选择此方法,并通过工具变量的两阶段最小二乘(2SLS)方法进一步阐释本文计量模型设定的思路。本质上,A平台开通征信认证功能这一外生冲击就是处理变量 $Treat_i$ (是否上传个人信用报告进行征信认证)的工具变量,即:

$$Treat_i = \alpha_1 + \theta_1 I(s_i \geq Credit) + f(s_i - Credit) + \varepsilon_i \quad (3)$$

$$Y_i = \alpha_2 + \theta_2 Treat_i + f(s_i - Credit) + \varepsilon_i \quad (4)$$

其中, $I(s_i \geq Credit)$ 为指示变量, 如果贷款 i 合同签订日期发生于 2015 年 4 月 27 日及之后则为 1, 否则为 0, 是处理变量 $Treat_i$ 的工具变量。一方面, A 平台开通征信认证功能这一事件冲击会影响个体上传信用报告的概率, 即满足工具变量的相关性条件; 另一方面, A 平台开通征信认证功能的时间是由平台决定而不是用户个体决定的, 因此该事件冲击日期相对于用户个体的贷款时间是外生的, 即满足工具变量的外生性条件。 Y_i 表示被解释变量, 即该笔贷款 i 的一系列特征变量; $f(s_i - Credit)$ 表示驱动变量 Cut_i 的多项式函数; ε_i 为残差项。

因此, 利用 A 平台开通征信认证功能这一外生冲击的准确日期构造 RDD, 可以有效识别征信数据要素可得性是否有利于 A 平台解决信息不对称问题。我们参考 Greenstone *et al.* (2022) 的做法, 基于以下模型设定进行参数估计:

$$Y_i = \gamma_1 I(s_i \geq Credit) + \gamma_2 f(s_i - Credit) + \gamma_3 I(s_i \geq Credit) \times f(s_i - Credit) + \mu_i \quad (5)$$

其中, Y_i 表示该笔贷款 i 的一系列特征变量, 比如是否征信认证、初始信用等级、事后违约率等; s_i 表示贷款 i 合同签订日期; $Credit$ 表示 A 平台开通征信认证的日期即 2015 年 4 月 27 日; $I(s_i \geq Credit)$ 为指示变量, 如果贷款 i 合同签订日期发生于 2015 年 4 月 27 日及之后则为 1, 否则为 0; 我们还加入了控制函数 $f(s_i - Credit)$, 以及指示变量与控制函数的交互项以灵活控制开通征信认证前后天数的影响; μ_i 为误差项。由于 A 平台开通征信认证的日期相对于用户个人在该平台申请贷款的日期是外生的, 我们在上述回归模型中没有加入额外的控制变量。考虑到本文使用的冲击是在日度层面, 为了与冲击的维度保持一致, 我们将误差项在日度层面进行聚类。本文基准回归采用带宽为断点前后的 90 天进行估计, 后面会进一步使用邻近的其他带宽进行再次检验, 以保证基准回归结果的稳健性。

(5) 式中待估参数 γ_1 是本文关心的核心系数, 其估计开通征信认证后 A 平台的各项贷款特征是否立即出现断点跳跃。此外, 本文使用的日度数据在时间维度上具有颗粒度更细、更高频的优势, 能较好满足断点回归所需的驱动变量平滑性假设。

(二) 基准回归结果

本文首先绘制了断点回归结果的可视化图形, 见图 2。从图 2 可以直观地看到, 在 A 平台开通征信认证后, 进行征信认证的贷款比例显著增加, 且每笔贷款的初始信用等级也显著提高, 而事后违约率有所下降但并不显著。对应的断点回归系数估计结果见表 2, 其中贷款的征信认证比例增加了 2.7 个百分点且在 1% 水平上显著, 初始信用也显著提高了 0.31 个等级, 而事后违约率的估计系数并不显著。

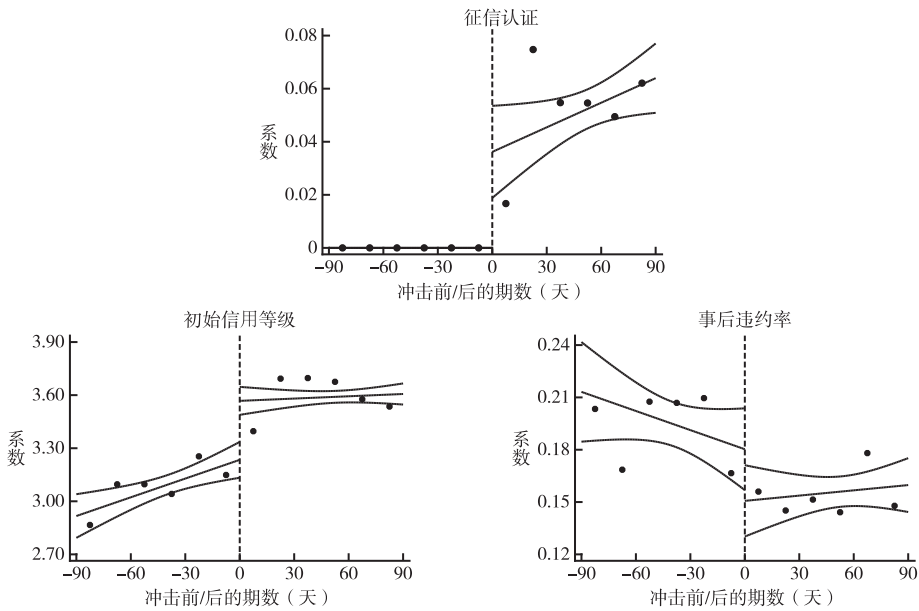


图2 断点回归结果的图形证据

说明:圆点表示每个6分箱体(bin)内部的因变量均值;直线为局部线性拟合;曲线表示95%显著性水平下的置信区间。后图4、5、7-9同。

表2	断点回归基准结果		
	(1)	(2)	(3)
	征信认证	初始信用等级	事后违约率
指示变量	0.0270*** (0.0083)	0.3056*** (0.1059)	-0.0232 (0.0248)
带宽(天数)	90	90	90
样本量	5731	5731	5486

说明:括号内的值为日度层面聚类的标准误,*、**、***分别代表10%、5%及1%的显著性水平,后表同。本文所有断点回归结果均使用局部线性回归点估计、三角核加权函数估计,故后表不再重复汇报。

上述结果表明,A平台开通征信认证后用户上传个人信用报告的比例上升,使得征信大数据的可得性提高。每笔贷款的初始信用等级提升说明个人信用报告确实为A平台进行贷前审核提供了重要的数据要素,即A平台将个人信用报告的信息纳入算法系统对用户进行风险评估,以筛选出信用评级更高的借款人群向其发放贷款。这表明征信认证功能的开通有利于A平台筛选出更优质的客户即缓解了逆向选

择问题。但对于事后违约率没有显著影响说明并没有解决道德风险问题,这可能是因为在申请贷款阶段的征信认证冲击本身并没有改变借款人事后违约成本以及其按时还款收益。这意味着开通征信认证功能为A平台在贷前审核阶段提供了更多的信息从而缓解了事前信息不对称问题,但其本身并没有发挥对借款人贷后行为的监督功能。理论上,只有通过加强贷后违约惩罚机制或提高贷后按时偿还激励等方式才可能解决事后信息不对称问题,例如将借款人在A平台的贷后行为纳入央行征信系统并更新其信用信息。Liao *et al.* (2023)在互联网无抵押消费信贷平台进行的随机实验发现,收到如下短信内容:“你的贷款还款情况和违约信息会同时上传央行征信系统”的提醒可以显著降低借款人的违约概率,但对于缓解逆向选择问题没有影响。Bu and Liao (2022)在中国西部某省份的乡村社会中关于小额信贷的随机对照实验也发现,在村里公示借款人的违约信息这一公开“羞辱处罚”的机制有利于改善其还款表现。上述两项随机实验的研究结果共同表明,建立事后惩罚机制有利于解决道德风险问题,这与本文研究结果形成互补,共同为如何激活数据要素潜能以解决信贷市场上的信息不对称问题提供了重要启示。即数据要素本身只能缓解逆向选择问题,必须同时实施配套的贷后惩罚机制才可能解决道德风险问题。

(三)平滑性检验

断点回归设计的一个关键识别假设是平滑性假设(Jia *et al.*, 2021),即除处理变量外,驱动变量以及其余变量在断点处(2015年4月27日)都应该是平滑变化的。我们首先检验驱动变量的平滑性,以证明样本在断点的左右两侧是随机分配的,这样才能保证断点回归模型的估计是无偏的。本文(1)式中的驱动变量为 Cut_i ,由于借款人无法提前知道A平台开通征信认证的日期,所以A平台开通征信认证的冲击是相对外生的,每笔贷款的合同签订日期不会受到这一外生冲击的影响。因此该驱动变量在断点处应该是连续变化的,即样本是随机分配的。我们使用McCrary (2008)检验以直观地判断驱动变量在断点处是否连续变化,如图3所示,断点两侧的驱动变量密度函数是连续的,并且断点处估计的对数

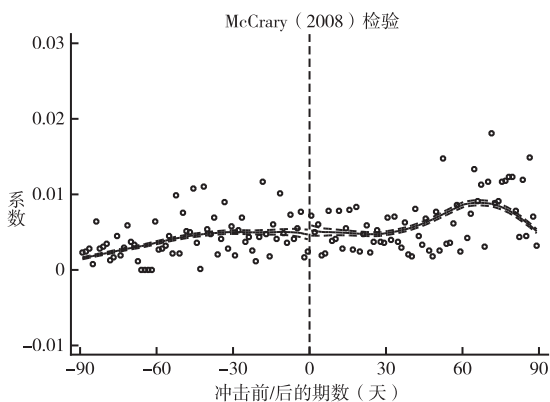


图3 驱动变量的平滑性检验

样本数差 $[r(\theta)=0.1188, r(se)=0.1010]$ 不显著。这证明了驱动变量的平滑性,表明样本是随机分配的,即说明本文使用的日度信贷数据满足断点回归设计所需的驱动变量连续性假设。

此外,我们还进一步检验了其他变量的平滑性。在A平台申请贷款阶段,用户需要填报自己的性别和年龄以及通过手机认证、户口认证、视频认证和教育认证核实其手机号、户口、脸部识别、教育学历等个人基本信息。一方面,征信认证这一事件冲击不应改变用户群体的性别和年龄结构,另一方面,手机认证、户口认证、视频认证和教育认证这4项功能早在A平台成立初期就已经开通使用,因此个体特征变量和其他认证功能在断点处应该也是平滑连续的。从图4和图5可知,性别、年龄、手机认证、户口认证、视频认证和教育认证在断点处均不存在显著跳跃。平滑性检验对应的系数回归结果^①也与图形结果一致,性别、年龄、手机认证、户口认证、视频认证和教育认证等变量在开通征信认证功能前后均没有显著变化。平滑性假设的成立表明A平台开通征信认证功能这一事件冲击是相对外生的。

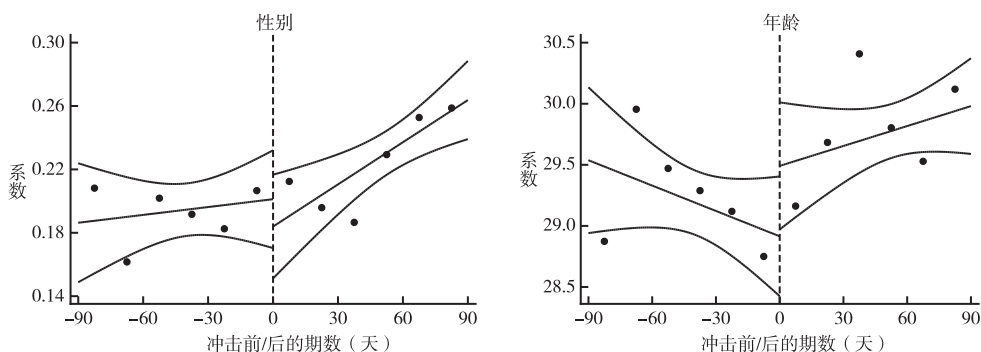


图4 个体特征变量的平滑性检验

(四)更换不同的带宽

Imbens and Lemieux(2008)指出在断点回归设计中,估计系数的灵敏度与带宽的选择密切相关,故本部分将利用邻近带宽进行稳健性检验。图6显示了邻近带宽的估计结果,虽然改变带宽对各项贷款特征变量的估计系数大小略有影响,但系数方向以及显著性并未发生改变。即在不同带宽设定下的估计结果均显示,A平台开通征

^① 限于篇幅,系数估计结果未报告,详见本刊网站本文补充材料附表C1和C2。

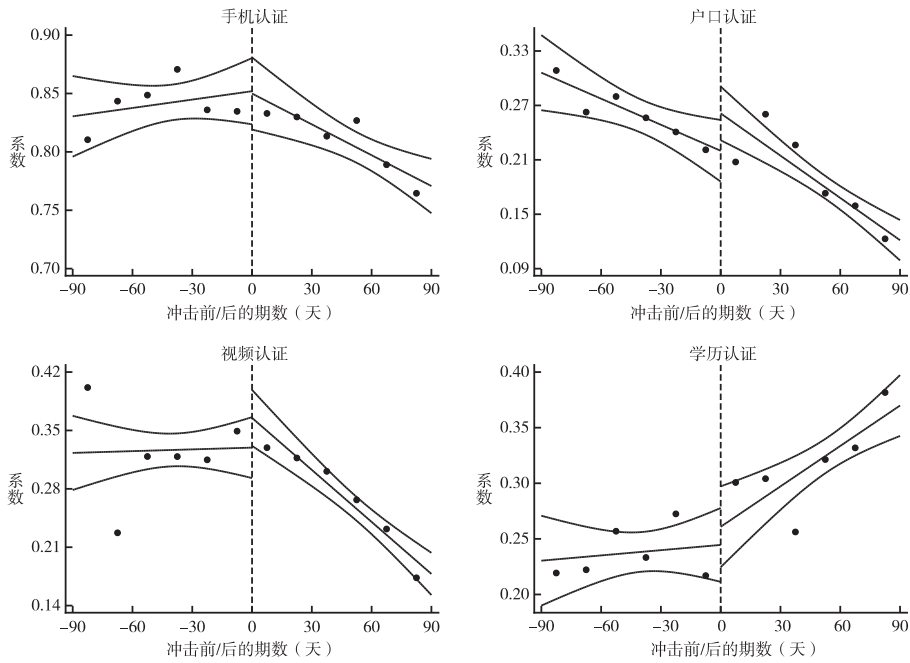


图5 其他认证的平滑性检验

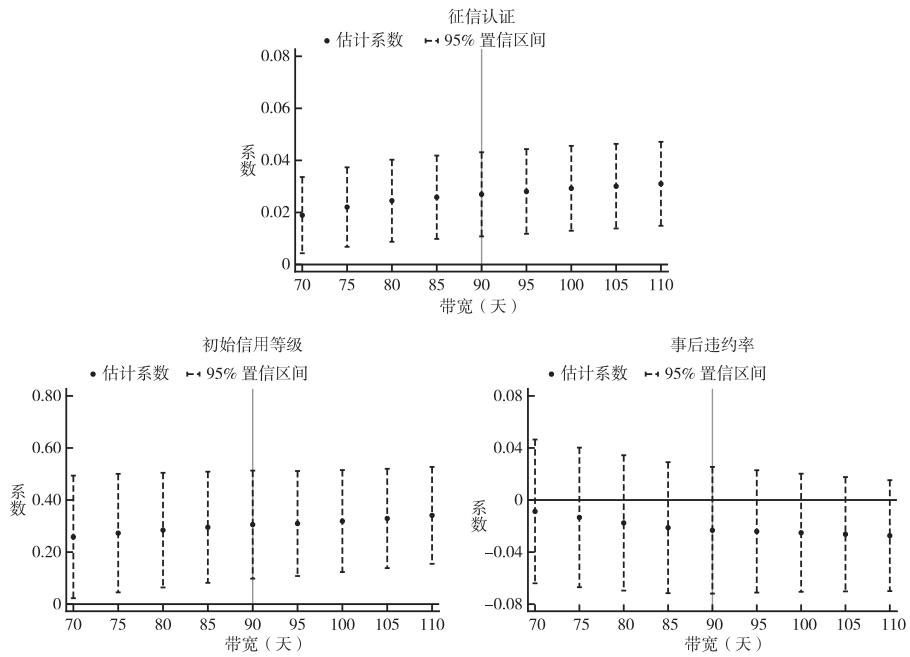


图6 不同带宽下的估计结果

信认证后增加了每笔贷款进行征信认证的概率,同时提高了每笔贷款的初始信用等级,但事后违约率没有显著变化。更换不同带宽的稳健性检验结果表明本文基准结果对于带宽的依赖性并不大。

(五)排除淘宝认证政策的影响

由于A平台淘宝认证功能的开通时间大约是在征信认证功能开通的前1周,为排除淘宝认证功能可能对本文基准回归结果的干扰,我们剔除了借款人为淘宝店主的这部分样本后再次回归,以排除淘宝认证功能开通的影响。图7直观地表明在剔除借款人为淘宝店主的样本后,断点回归的图形结果依然稳健,即A平台在开通征信认证功能后每笔贷款进行征信认证的概率提高,且贷款的初始信用等级也得到提升,但事后违约率没有显著变化^①。上述结果排除了淘宝认证功能开通对本文研究结果可能存在的影响。

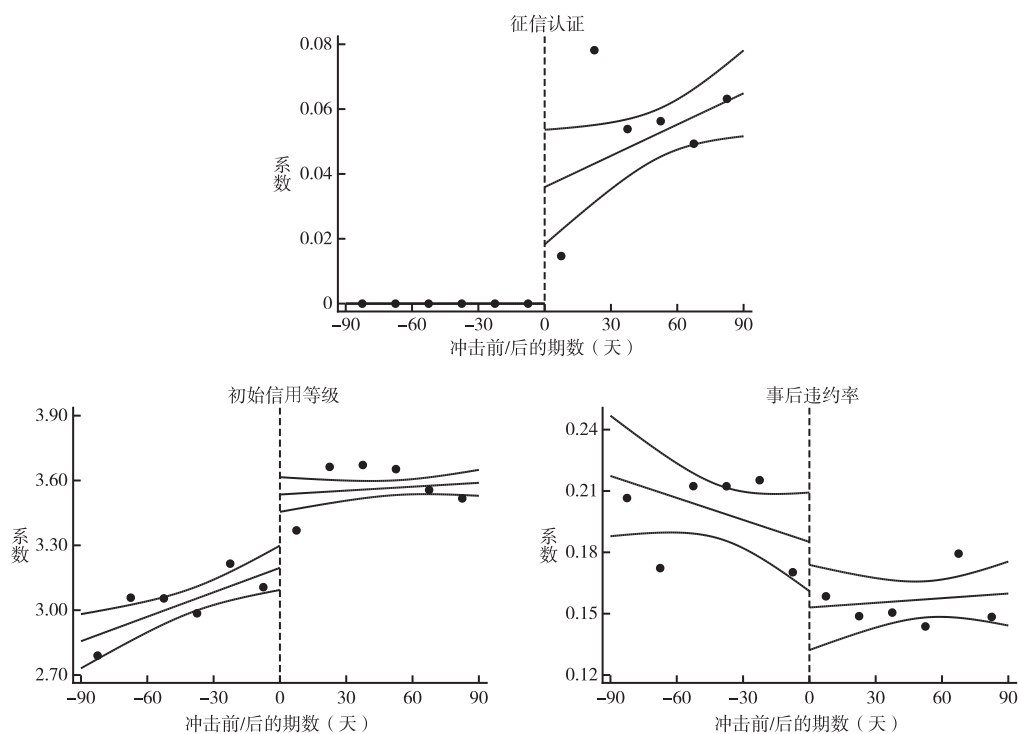


图7 排除淘宝认证功能开通的影响

① 限于篇幅,对应的系数估计结果未报告,详见本刊网站本文补充材料附表C3。

(六)加入月度和周度固定效应

为进一步排除可能存在的同时期宏观因素的干扰,我们在(5)式模型设定的基础上加入了月度和周度固定效应,以控制宏观层面的经济政策或其他冲击可能产生的影响。加入月度和周度固定效应后的回归结果见表3。与表2基准回归结果相比,现有回归结果除系数大小略有不同外,系数的方向和显著性与基准回归结果基本保持一致。这表明在剔除可能存在的同时期宏观层面的经济因素或冲击后,本文基本结论依然成立。

表3 加入月度和周度固定效应的断点回归结果

	(1)	(2)	(3)
	征信认证	初始信用等级	事后违约率
指示变量	0.0185** (0.0078)	0.2806*** (0.1022)	-0.0180 (0.0246)
月度固定效应	控制	控制	控制
周度固定效应	控制	控制	控制
样本量	5731	5731	5486

五 进一步分析

(一)借款人的历史表现

前文基准回归分析发现,在A平台开通征信认证后,每笔贷款的初始信用等级显著提升。但存在一种可能的解释是只要上传了征信报告的借款人,其初始信用等级都会更高,即A平台并没有真正地将个人征信报告中丰富的信用数据要素纳入信用评价算法模型,而只是简单地基于借款人是否上传征信报告对用户进行初始信用评级。为考察A平台是否在贷前审核阶段就真正利用了个人征信报告提供的重要数据要素,并筛选出了信用等级更高的借款人群,我们进一步使用借款人的历史表现衡量借款人的信用优质程度以证明其是否真的排除了风险水平更高的借款人群,即检验是否在一定程度上缓解了逆向选择问题。

由图8可以看出,在A平台开通征信认证功能后获得贷款的借款人群其历史成功借款次数和历史正常还款期数显著高于开通征信认证前的借款人群。其中,历史成功借款次数是指该借款人在本次借款之前于A平台成功借款的次数,历史正常还款期数是指该借款人在本次借款之前于A平台按期还款的期数。与初始信用等级相

比,这两个历史表现指标更为直接、客观地衡量了借款人的信用水平。从表4结果可知,通过比较A平台征信认证开通前后借款人群的历史表现差异可以发现,借款人的历史成功借款次数提高了17.64%,同时历史正常还款期数增加了33.14%,二者均在1%的水平上显著。这表明A平台确实筛选出了信用水平更为优质的借款人,征信认证的开通确实具有缓解逆向选择问题的作用,这也是征信数据的价值所在。

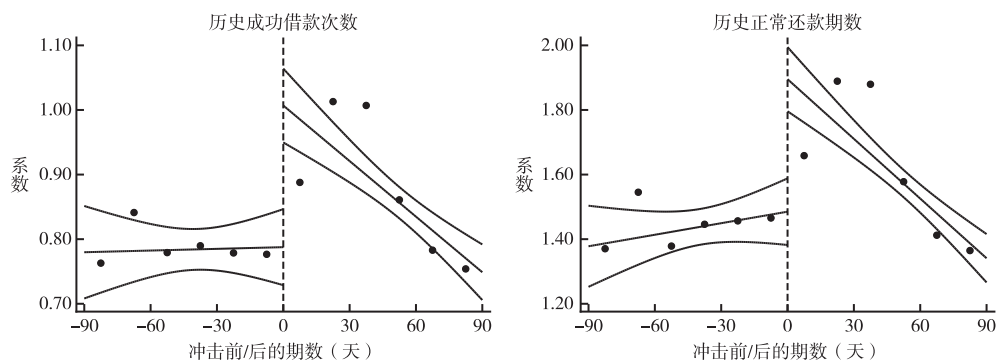


图8 借款人的历史表现

表4 借款人历史表现的参数回归结果

	(1)	(2)
	历史成功借款次数	历史正常还款期数
指示变量	0.1764*** (0.0584)	0.3314*** (0.1037)
样本量	5731	5731

(二)借款利率和金额的影响

在开通征信认证功能后,A平台纳入征信大数据的风控系统能更准确地预测借款人的风险概率,从而缓解了逆向选择问题。那么,A平台是否进一步基于对借款人准确的风控评级进行风险定价?为此,我们进一步考察了A平台征信认证的开通对借款利率和借款金额的影响。如图9所示,在开通征信认证后,A平台的借款利率显著降低同时借款金额显著上升,表明A平台确实根据借款人群的不同信用等级进行了风险定价。表5的回归结果显示,借款利率下降了约0.59个百分点,同时借款金额提高了15.24%,且均在1%的水平上显著,表明在开通征信认证后显著降低了借款利率并显著增加了借款金额。

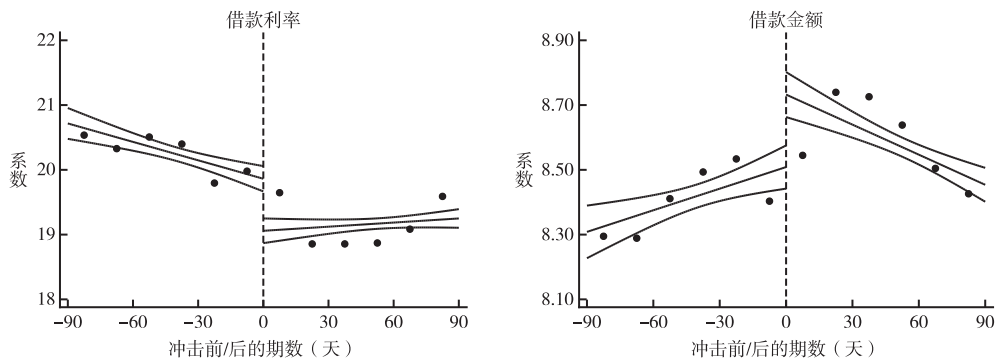


图9 借款利率和借款金额的影响

表5 借款利率和借款金额的参数回归结果

	(1)	(2)
	借款利率	借款金额
指示变量	-0.5901*** (0.2273)	0.1524*** (0.0525)
借款期限	控制	控制
借款金额	控制	未控制
借款利率	未控制	控制
样本量	5731	5731

说明:表中添加了贷款层面的控制变量,作者感谢审稿专家的这一建议。

上述结果说明,在开通征信认证后,A平台能通过获得的征信数据要素筛选出风险水平更低的借款人群以更低的利率向其发放贷款,且每笔贷款的金额得到提升即扩大了贷款供给。这实现了借贷双方的共赢,一方面优质借款人以更低的利率贷到了更多的资金,另一方面由于贷方有效区分了高风险和低风险借款人因此其承担的风险和收益也会更为匹配,从而促进整个信贷市场上资金的配置效率。由此可见缓解事前信息不对称带来的效果,即征信大数据的可得性帮助了贷方识别和评估借款人风险缓解了事前的信息不对称,从而能够收取更符合借款人真实风险水平的利率,同时增加了信贷供给,使信贷配置效率得到提高以及社会福利得到改善,这体现了解决逆向选择问题的重要性。

六 福利分析

基于前文分析可知,A平台开通征信认证功能提高了征信数据要素的可得性,进而在一定程度上解决了事前信息不对称的问题。下面,我们将估计这一事件冲击所产生的逆向选择缓解作用带来了多大的福利改善。

(一)福利改进的思路阐述

在建立关于福利分析的理论模型之前,我们使用一个简单的信贷供求框架阐述信息不对称造成的社会福利损失。在信贷市场的需求方存在高风险和低风险两类借款人,高风险借款人的违约风险更高,因此其借贷成本也更高($r_{H,fair}$)。反之,低风险借款人违约风险更低,借贷成本也更低($r_{L,fair}$)。但由于信息不对称的存在,贷方不能准确区分高风险借款人和低风险借款人,因此贷方只能采取统一定价(r_{pool}),即满足如下不等式关系 $r_{H,fair} > r_{pool} > r_{L,fair}$ 。

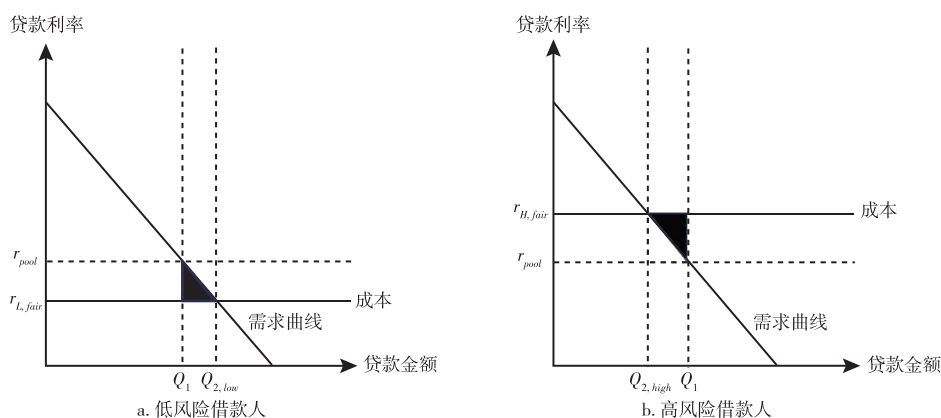


图 10 信息不对称与社会福利损失

当放贷者获得了额外的信息(即征信报告)使其能区分高风险和低风险借款人时,放贷者可以采用三级价格歧视的方法对高风险借款组和低风险借款组进行分别定价,并由此实现社会福利的改进。对于低风险借款人,其定价曲线将从 r_{pool} 降低到 $r_{L,fair}$,均衡的贷款额将从 Q_1 增加到 $Q_{2,low}$ (见图 10a),整个社会福利将增加黑色三角形部分。类似的,对于高风险借款人而言,其面临的借贷成本将会从 r_{pool} 上升到 $r_{H,fair}$,

均衡的贷款额将从 Q_1 减少到 $Q_{2,high}$ (见图 10b), 此时社会福利增加的部分同样为黑色三角形。因此, 当贷方获得了额外信息能准确区分高风险借款人和低风险借款人, 并进行差别定价时, 整个社会福利增加的部分为上述两个组别的黑色三角形面积之和。

(二)理论模型

本文主要基于 Jansen *et al.* (2025) 的模型计算 A 平台开通征信认证功能后带来的社会福利改进。借款人先在 $t = 0$ 时选择借贷额 L 并在随后的还款期内选择每期偿还相同的本金和利息额, 即 $\pi(L, r)$, 其效用函数为^①:

$$U(c) = \underbrace{u_0(c_0)}_{\text{消费}} + \underbrace{\sum_{t=1}^T \beta^t (1 - \delta)^t u(c_t)}_{\text{偿还}} + \underbrace{\sum_{t=1}^T (1 - \delta)^{t-1} \delta \sum_{i=t}^T \beta^i u(c_D)}_{\text{违约}} \quad (6)$$

$$c_0 = w_0 + L, c_t = w_t - \pi(L, r), \pi(L, r) = L\phi(r) \quad (7)$$

其中, δ 为借款人的违约概率, w 表示借款人的财富水平。借款人获得贷款后选择在 0 期就消费完, 即获得效用水平“消费”部分。当借款人在未来的每一期都选择还款时, 其未来效用水平的折现部分则为“偿还”部分。如果借款人在未来的某一期选择违约, 那么其获得的效用水平为“违约”部分。一般来说 c_D 小于 c_t , 因为借款人违约后会影响以后的借贷可得性, 并降低可用消费额。对 $u(c_t)$ 线性展开可得:

$$u(c_t) \approx u(w_t) + u'(w_t)(c_t - w_t) = u(w_t) - u'(w_t)\pi(L, r) \quad (8)$$

结合(6)-(8)式, 可将借款人的最优化问题表示为:

$$V(r) = \max_L u_0(w_0 + L) + \sum_{t=1}^T \beta^t (1 - \delta)^t [u(w_t) - u'(w_t)\pi(L, r)] + \sum_{t=1}^T (1 - \delta)^{t-1} \delta \sum_{i=t}^T \beta^i u(c_D) \quad (9)$$

当利率水平为 r 时, 借款人选择最优借贷额, 即为:

$$L^*(r) = \arg \max_L u_0(w_0 + L) + \sum_{t=1}^T \beta^t (1 - \delta)^t [u(w_t) - u'(w_t)\pi(L, r)] + \sum_{t=1}^T (1 - \delta)^{t-1} \delta \sum_{i=t}^T \beta^i u(c_D) \quad (10)$$

进一步地, 我们要测算由于利率变化导致的借款人效用水平的变化, 这可以依赖

① 借款人按照等额本息的方式进行每期还款, 每期还款额为总借贷额乘以还款比例, 即复利率 $\phi(r)$, 还款比例计算方式为 $\phi(r) = \frac{r(1+r)^T}{(1+r)^T - 1}$, 其中 T 为总的还款期数。

包络定理,通过对(10)式求导可得:

$$\frac{dV(r)}{dr} = -L^*(r) \frac{d\phi(r)}{dr} \sum_{i=1}^T \beta^i (1-\delta)^i u'(w_i) \quad (11)$$

因此,可以得到当借款人面临的利率水平从 \tilde{r} 降低到 r 时,借款人的效用变化为:

$$V(r) - V(\tilde{r}) = \underbrace{\left[\sum_{i=1}^T \beta^i (1-\delta)^i u'(w_i) \right]}_{\text{效用权重}} \underbrace{\left[\int_{\tilde{r}}^r L^*(\hat{r}) \frac{d\phi(\hat{r})}{d\hat{r}} d\hat{r} \right]}_{\text{借款额变动}} \quad (12)$$

使用(12)式中的权重对上述效用变化进行标准化,并对借贷市场上的借款人进行加总,可以得到消费者剩余的变化为:

$$\Delta CS = CS(r) - CS(\tilde{r}) = \int_{\tilde{r}}^r \Lambda(\hat{r}) \frac{d\phi(\hat{r})}{d\hat{r}} d\hat{r} \quad (13)$$

$$\Lambda(\hat{r}) = \sum_i L_i(r) \quad (14)$$

生产者剩余的变化为:

$$\Delta PS = PS(r) - PS(\tilde{r}) = \Lambda(r) [\phi(r) - \phi(r_c)] - \Lambda(\tilde{r}) [\phi(\tilde{r}) - \phi(r_c)] \quad (15)$$

那么当利率水平从 \tilde{r} 降低到 r 时,总的社会福利变化为:

$$TS(r) - TS(\tilde{r}) = \Delta CS + \Delta PS \quad (16)$$

具体到本文的应用场景,当贷方通过获得额外信息(即征信报告)可以对不同风险的借款人进行差别定价时,整个借款人从原有的统一利率水平 r_{pool} 转变为:低风险借款人面临的利率下降到 $r_{L,fair}$,高风险借款人面临的利率上升到 $r_{H,fair}$ 。在这种情况下,贷方的利润净增加额为:

$$\sum_j \Lambda(\phi(r_{pool})) [\phi(r_{j,fair}) - \phi(r_{pool})], j = H, L \quad (17)$$

借款人的消费者剩余净增加额为:

$$\sum_j [\phi(r_{pool}) - \phi(r_{j,fair})] \left\{ \frac{\Lambda(\phi(r_{j,fair})) + \Lambda(\phi(r_{pool}))}{2} \right\}, j = H, L \quad (18)$$

总社会福利净增加额为:

$$\sum_j [\phi(r_{j,fair}) - \phi(r_{pool})] \left\{ \frac{\Lambda(\phi(r_{pool})) - \Lambda(\phi(r_{j,fair}))}{2} \right\}, j = H, L \quad (19)$$

其中:

$$\Lambda(\phi(r)) = a - b\phi(r) \quad (20)$$

因此,从(19)和(20)式可知总的社会福利净增加额可以表达为需求价格弹性的

充分统计量。即在知道相关的利率水平或者贷款需求后,可以通过简单代数运算估计出社会福利的变化,由此也避免了采用结构估计进行福利分析时的复杂计算,这也是充分统计量方法的优势(Chetty, 2009)。

(三)估计结果

我们利用原始数据中的初始信用等级划分高风险借款组和低风险借款组,其中初始信用等级为1和2的是高风险借款组,初始信用等级为3-6的是低风险借款组^①。为了与表5中的估计结果保持一致,在计算福利水平时我们使用的样本量限制在征信认证冲击的前90天之内。并且为了计算反事实状态,我们仅用事前的数据(即不受征信认证的影响)推算福利变化,以此避免计算时的内生性问题。

要估算出(19)式的大小,首先要得到不同组别的利率差值和贷款量差值,即 $\phi(r_{j, fair}) - \phi(r_{pool})$ 和 $\Lambda(\phi(r_{pool})) - \Lambda(\phi(r_{j, fair}))$,而这两者之间的关系则依赖于(20)式中的系数 b ,系数 b 同贷款的需求价格弹性紧密相关。因此,我们利用表5第(1)和(2)列的估计结果并结合冲击前的利率平均水平可以计算得到贷款的需求价格弹性为-5.2239(如表6所示),这一结果同Jansen *et al.* (2025)估算的弹性大小(-2.32)相比略大,但都在同一个数量级。这两者之间存在差异的一个可能原因是数据结构的不同,他们利用的是个体层面的月度贷款面板数据进行估算,而本文则是基于个体单笔借贷合同的总贷款额估算的,其中每笔借贷合同的平均借款期限为10个月左右。所以相比他们估计的每月层面的贷款额对利息价格的敏感程度,10个月的贷款额更多,对于价格的敏感程度也越高,因此本文估计出的需求价格弹性更大。

需求价格弹性 e 与 b 的关系如下:

$$b_L \frac{\phi(r_{pool})}{\Lambda_L(\phi(r_{pool}))} = b_H \frac{\phi(r_{pool})}{\Lambda_H(\phi(r_{pool}))} = e \quad (21)$$

进一步地,假设在面临相同的利率水平时,高风险借款人和低风险借款人会借贷相同的贷款额,即有:

^① 由图10可知,在征信认证开通的冲击下,高风险借款人的借款利率会上升,而低风险借款人的借款利率则会下降,因此我们通过分别比较各初始信用等级组别的借款人在征信认证开通的冲击前后其借款利率水平的变化划分高风险借款人组和低风险借款人组。其中,初始信用等级为1和2的借款人其借款利率在冲击后上升,则划分为高风险借款人组;初始信用等级为3-6的借款人的借款利率水平在冲击后下降则为低风险借款人组。

$$\frac{\Lambda_L(\phi(r_{pool}))}{N_L} = \frac{\Lambda_H(\phi(r_{pool}))}{N_H} \quad (22)$$

其中 N_L 和 N_H 分别表示低风险借款人和高风险借款人的数量。此外,我们假设均衡时贷方在低风险人群中获得的利润 $(\phi(r_{pool}) - \phi(r_{L,fair}))$ 会用来补贴高风险人群中的亏损 $(\phi(r_{H,fair}) - \phi(r_{pool}))$,但由于市场势力的存在,其仍能获得垄断利润。即有:

$$(1 - \psi_L)\Lambda_L(\phi(r_{L,fair}))\phi(r_{L,fair}) + (1 - \psi_H)\Lambda_H(\phi(r_{H,fair}))\phi(r_{H,fair}) = \psi_L\Lambda_L(\phi(r_{L,fair})) + \psi_H\Lambda_H(\phi(r_{H,fair})) \quad (23)$$

其中, ψ_L 和 ψ_H 分别表示低风险借款人和高风险借款人的违约率。结合(20)和(21)式则有:

$$b_L = \frac{\Lambda_L(\phi(r_{L,fair})) - \Lambda_L(\phi(r_{pool}))}{\phi(r_{pool}) - \phi(r_{L,fair})}, b_H = \frac{\Lambda_H(\phi(r_{pool})) - \Lambda_H(\phi(r_{H,fair}))}{\phi(r_{H,fair}) - \phi(r_{pool})} \quad (24)$$

另外,从估计系数 $\hat{\beta}_{loan}$ 的定义出发则存在如下等式,即总贷款额在事件冲击后的值等于潜在高低风险借贷组的贷款总量的加权平均:

$$\Lambda(\phi(r_{pool}))(1 + \hat{\beta}_{loan}) = p_L\Lambda_L(\phi(r_{L,fair})) + p_H\Lambda_H(\phi(r_{H,fair})) \quad (25)$$

同理可得借款利率的表达式为:

$$\phi(r_{pool} + \hat{\beta}_{rate}) = p_L\phi(r_{L,fair}) + p_H\phi(r_{H,fair}) \quad (26)$$

其中, p_L 和 p_H 代表低、高风险借款人的比例, $\hat{\beta}_{loan}$ 和 $\hat{\beta}_{rate}$ 即表5第(1)和(2)列估计的借款金额和借款利率的回归系数。因此,结合(21)–(26)式可以计算出 $\Lambda_L(\phi(r_{L,fair}))$ 和 $\Lambda_H(\phi(r_{H,fair}))$, b_L 和 b_H , $\phi(r_{L,fair})$ 和 $\phi(r_{H,fair})$,最后按照(19)式可以计算得到总的福利水平变化。

表6列出了按照上述方法计算的福利情况。计算结果表明,在征信认证开通后降低了无谓损失,进而提升了社会福利,并且量化的结果显示福利改进了0.0927%。进一步地,我们利用表6的估计结果可以外推整个样本期间内福利改进的绝对值大小。具体计算方式为:首先,统计出在2015年4月27日至2016年12月31日之间高、低风险借款合同的数量,分别是16 031和266 932笔;其次,利用表6中估计的单个高、低风险借款人的福利改进值,计算出加总的福利改进水平为161.4万元;最后,由于本文使用的数据是原始平台数据10%的随机抽样,所以总的福利改进水平在此基础上还要乘以10,即1614万元。因此,通过本文建立的一个理论模型,我们估计出该金融科技平台在开通征信认证后,自2015年4月27日到2016年12月31日这一时间段累计提高了社会福利1614万元。

表 6

社会福利变化

	数值	数据来源
冲击前(<i>pool</i>):		
利率平均水平(%)	20.2272	原始数据
借款金额(元)	5700	原始数据
高风险组违约率	0.2852	原始数据
高风险组占比	0.3067	原始数据
低风险组违约率	0.1541	原始数据
低风险组占比	0.6933	原始数据
弹性(<i>e</i>):		
需求价格弹性	-5.2239	由表 5 第(1)和(2)列估计结果计算得到
福利水平:		
单个高风险借款人的需求	3420.5954	
单个低风险借款人的需求	7961.3203	
单个高风险借款人需求斜率	67 779	由(21)-(26)式计算得到
单个低风险借款人需求斜率	153 217	
单个高风险借款人的复利率	24.1021	
单个低风险借款人的复利率	23.9526	
单个高风险借款人的福利改进(元)	4.1404	由(19)式计算得到
单个低风险借款人的福利改进(元)	5.7979	由(19)式计算得到
福利改进占平均借款金额之比(%)	0.0927	
总体福利改进额(万元)	1614	

七 主要结论与政策启示

随着大数据、人工智能、云计算等数字技术的加速融合发展和金融科技的不断创新,数据要素已经成为一类重要的资产(Jones and Tonetti, 2020)并得到了理论和实务界的高度重视,但到底应当如何激活数据要素,以有效发挥其价值目前鲜有研究进行回答。本文利用金融科技平台开通征信认证功能以提高征信报告可得性这一冲击构造断点回归,研究发现征信数据要素能帮助平台在贷前审核阶段挑选出优质借款人从而解决逆向选择问题。但由于征信数据的可得性本身并没有改变借款人事后违约成本以及按时还款收益,每笔贷款的事后违约率没有显著变化,因而必须对借款人施加一定的违约惩罚机制或按时还款激励才能约束其事后行为以解决道德风险问题。此外,本文基于 Jansen *et al.* (2025) 的框架构建充分统计量估计了征信数据要素的经

济价值,结果显示征信大数据可得性的提高创造了价值1614万元的福利改进。

本研究结论为如何激活数据要素潜能以有效解决信息不对称问题并推进数字中国建设提供重要政策启示。一方面,大数据要素是解决逆向选择的有效工具,所以应该加大建设征信体系等重要金融基础设施的深度和广度,同时积极鼓励对其余各大市场中海量数据资源进行挖掘和共享并建立相应的数据基础设施,以优化整个市场的运行效率。这一启示同2023年3月中共中央、国务院组建国家数据局的重大举措相呼应,该部门的核心职责就是负责协调推进数据基础制度建设,统筹数据资源整合共享和开发利用,推进数字中国、数字经济的建设。另一方面,大数据本身仅能起到描绘客户信用风险画像的基础要素作用,必须配套制定事后违约惩罚机制才能有效解决道德风险问题,从而促进金融市场高效发展。因此,大数据作为重要的信息要素媒介需依赖于相应的市场制度完善才能实现更多丰富的应用场景落地以更好地发挥我国海量数据的规模优势。

参考文献:

- 曹光宇、刘畅、周黎安(2022):《大数据征信与平台流量:基于共享单车免押骑行的经验研究》,《世界经济》第9期。
- 方锦程、刘颖、高昊宇、董纪昌、吕本富(2023):《公共数据开放能否促进区域协调发展?——来自政府数据平台上线的准自然实验》,《管理世界》第9期。
- 黄炜、任昶宇、周羿(2023):《退休制度、劳动供给与收入消费动态》,《经济研究》第1期。
- 黄益平、邱晗(2021):《大科技信贷:一个新的信用风险管理框架》,《管理世界》第2期。
- 李稻葵、刘淳、庞家任(2016):《金融基础设施对经济发展的推动作用研究——以我国征信系统为例》,《金融研究》第2期。
- 马述忠、贺歌、郭继文(2024):《如何缓解跨境交易中的信息不对称?——来自跨境电商质量认证的经验证据》,《数量经济技术经济研究》第6期。
- 田利辉、李政、施炳展(2025):《企业上云对其创新的影响:数据要素化的视角》,《世界经济》第1期。
- 王正位、周从意、廖理、张伟强(2020):《消费行为在个人信用风险识别中的信息含量研究》,《经济研究》第1期。
- 杨俊、李小明、黄守军(2022):《大数据、技术进步与经济增长——大数据作为生产要素的一个内生增长理论》,《经济研究》第4期。
- 尹志超、甘犁(2011):《信息不对称、企业异质性与信贷风险》,《经济研究》第9期。
- 周沂、冯皓月、陈晓兰(2021):《中央环保督察的震慑效应与我国环境治理机制的完善》,《经济学动态》第8期。
- Akerlof, G. A. "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism." *The Quarterly Journal of Economics*, 1970, 84(3), pp. 488–500.

Arrow, K. J. "Uncertainty and the Welfare Economics of Medical Care." *The American Economic Review*, 1963, 53(5), pp. 941–973.

Begenau, J.; Farboodi, M. and Veldkamp, L. "Big Data in Finance and the Growth of Large Firms." *Journal of Monetary Economics*, 2018, 97, pp. 71–87.

Beraja, M.; Yang, D. Y. and Yuchtman, N. "Data-Intensive Innovation and the State: Evidence from AI Firms in China." *The Review of Economic Studies*, 2023, 90(4), pp. 1701–1723.

Bu, D. and Liao, Y. "Shaming Microloan Delinquents: Evidence from a Field Experiment in China." *Management Science*, 2022, 68(5), pp. 3768–3790.

Buchak, G.; Matvos, G.; Piskorski, T. and Seru, A. "Fintech, Regulatory Arbitrage, and the Rise of Shadow Banks." *Journal of Financial Economics*, 2018, 130(3), pp. 453–483.

Chan, Y. S. and Thakor, A. V. "Collateral and Competitive Equilibria with Moral Hazard and Private Information." *Journal of Finance*, 1987, 42(2), pp. 345–363.

Chetty, R. "Sufficient Statistics for Welfare Analysis: A Bridge Between Structural and Reduced-Form Methods." *Annual Review of Economics*, 2009, 1(1), pp. 451–488.

DeFusco, A. A.; Tang, H. and Yannelis, C. "Measuring the Welfare Cost of Asymmetric Information in Consumer Credit Markets." *Journal of Financial Economics*, 2022, 146(3), pp. 821–840.

Dell’Ariccia, G. and Marquez, R. "Information and Bank Credit Allocation." *Journal of Financial Economics*, 2004, 72(1), pp. 185–214.

Di Maggio, M. and Yao, V. "Fintech Borrowers: Lax Screening or Cream-skimming?" *The Review of Financial Studies*, 2021, 34(10), pp. 4565–4618.

Djankov, S.; McLiesh, C. and Shleifer, A. "Private Credit in 129 Countries." *Journal of Financial Economics*, 2007, 84, pp. 299–329.

Doblas-Madrid, A. and Minetti, R. "Sharing Information in the Credit Market: Contract-level Evidence from U. S. Firms." *Journal of Financial Economics*, 2013, 109, pp. 198–223.

Einav, L. and Finkelstein, A. "Moral Hazard in Health Insurance: What We Know and How We Know It." *Journal of the European Economic Association*, 2018, 16(4), pp. 957–982.

Einav, L.; Finkelstein, A. and Cullen, M. R. "Estimating Welfare in Insurance Markets Using Variation in Prices." *The Quarterly Journal of Economics*, 2010, 125(3), pp. 877–921.

Farboodi, M. and Veldkamp, L. "Long-run Growth of Financial Data Technology." *The American Economic Review*, 2020, 110(8), pp. 2485–2523.

Farboodi, M.; Mihet, R.; Philippon, T. and Veldkamp, L. "Big Data and Firm Dynamics." *AEA papers and proceedings*, 2019, 109, pp. 38–42.

Finkelstein, A. and Poterba, J. "Adverse Selection in Insurance Markets: Policyholder Evidence from the UK Annuity Market." *Journal of Political Economy*, 2004, 112(1), pp. 183–208.

Fuster, A.; Plosser, M.; Schnabl, P. and Vickery, J. "The Role of Technology in Mortgage Lending." *The Review of Financial Studies*, 2019, 32(5), pp. 1854–1899.

- Greenstone, M.; He, G.; Jia, R. and Liu, T. “Can Technology Solve the Principal-Agent Problem? Evidence from China’s War on Air Pollution.” *American Economic Review: Insights*, 2022, 4(1), pp. 54–70.
- Gupta, A. and Hansman, C. “Selection, Leverage, and Default in the Mortgage Market.” *The Review of Financial Studies*, 2022, 35(2), pp. 720–770.
- He, Z.; Huang, J. and Zhou, J. “Open Banking: Credit Market Competition When Borrowers Own the Data.” *Journal of Financial Economics*, 2023, 147(2), pp. 449–474.
- Igawa, K. and Kanatas, G. “Asymmetric Information, Collateral and Moral Hazard.” *Journal of Financial and Quantitative Analysis*, 1990, 25(4), pp. 469–490.
- Imbens, G. W. and Lemieux, T. “Regression Discontinuity Designs: A Guide to Practice.” *Journal of Econometrics*, 2008, 142(2), pp. 615–635.
- Jansen, M.; Nagel, F.; Yannelis, C. and Zhang, A. L. “Data and Welfare in Credit Markets.” *Journal of Financial Economics*, 2025, 174, 104171.
- Jappelli, T. and Pagano, M. “Information Sharing, Lending and Defaults: Cross-country Evidence.” *Journal of Banking & Finance*, 2002, 26, pp. 2017–2045.
- Jia, J.; Liang, X. and Ma, G. “Political Hierarchy and Regional Economic Development: Evidence from a Spatial Discontinuity in China.” *Journal of Public Economics*, 2021, 194, 104352.
- Jones, C. I. and Tonetti, C. “Nonrivalry and the Economics of Data.” *The American Economic Review*, 2020, 110(9), pp. 2819–2858.
- Karlan, D. and Zinman, J. “Observing Unobservables: Identifying Information Asymmetries with a Consumer Credit Field Experiment.” *Econometrica*, 2009, 77, pp. 1993–2008.
- Liao, L.; Martin, X.; Wang, N.; Wang, Z. and Yang, J. “What If Borrowers Were Informed about Credit Reporting? Two Natural Field Experiments.” *The Accounting Review*, 2023, 98(3), pp. 397–425.
- McCrary, J. “Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test.” *Journal of Econometrics*, 2008, 142(2), pp. 698–714.
- Menkhoff, L.; Neuberger, D. and Suwanaporn, C. “Collateral-based Lending in Emerging Markets: Evidence from Thailand.” *Journal of Banking & Finance*, 2006, 30, pp. 1–21.
- Michael, S. “Job Market Signaling.” *The Quarterly Journal of Economics*, 1973, 87(3), pp. 355–374.
- Pagano, M. and Jappelli, T. “Information Sharing in Credit Markets.” *Journal of Finance*, 1993, 43, pp. 1693–1718.
- Rothschild, M. and Stiglitz, J. “Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information.” *Foundations of Insurance Economics, Springer Netherlands*, 1976, pp. 355–375.
- Stiglitz, J. and Weiss, A. “Credit Rationing in Markets With Imperfect Information.” *The American Economic Review*, 1981, 71, pp. 393–410.
- Stroebel, J. “Asymmetric Information About Collateral Values.” *Journal of Finance*, 2016, 71(3), pp. 1071–1112.

Unlocking Data Potential through Financial Infrastructure: Evidence from Central Bank Credit Reporting

Zhou Bo; Liang Yunjia

Abstract: As a core component of China's foundational data infrastructure, financial infrastructure holds significant practical importance for enhancing the circulation and allocation of data resources. This paper investigates the economic value of credit data by examining the credit reporting system, employing 285,603 daily loan transaction records from the fintech platform PPDAl between 2015 and 2016. To identify causal effects, an exogenous policy shock – namely, PPDAl's introduction of a credit authentication feature on 27 April 2015, which allowed borrowers to submit their central bank credit reports – is exploited to implement a regression discontinuity design. This approach facilitates an investigation into the economic value of credit data elements and their underlying mechanisms.

The findings reveal three principal results. First, the introduction of credit authentication significantly improved access to large-scale credit data, providing a critical input for the platform's pre-loan screening. On average, initial credit ratings increased by 0.31 units per loan, indicating that enhanced data availability helped to mitigate ex-ante information asymmetry. However, as this policy shock did not alter borrowers' ex-post incentives – such as default penalties or repayment rewards – default rates remained unchanged. This implies that, in this context, large-scale credit data can alleviate adverse selection but does not mitigate moral hazard. Second, following the introduction of credit authentication, the platform's average loan interest rates fell by 0.59 percentage points, while average loan amounts increased by 15.24%. This suggests that improved access to credit reports enabled lenders to better identify and assess borrower risk, adjust interest rates accordingly, and expand credit supply, thereby improving credit allocation efficiency. Third, a welfare analysis indicates that the improved availability of large-scale credit data increased social welfare by 0.0927% through the mitigation of adverse selection. In monetary terms, the policy generated a cumulative welfare gain of RMB 16.14 million between 27 April 2015 and 31 December 2016, underscoring the significant economic value of credit data as a productive factor.

These findings offer important policy implications. On the one hand, big data is an effective tool for addressing adverse selection. Thus, policymakers should prioritise the expansion of credit reporting systems and the broader development of data infrastructure, alongside efforts to integrate and share massive data resources across sectors. This aligns with China's broader institutional goals, exemplified by the establishment of the National Data Bureau in 2023, which is tasked with coordinating data infrastructure and promoting the digital economy. On the other hand, big data alone cannot resolve moral hazard. Without complementary mechanisms – such as robust ex-post default penalties – it serves only as a preliminary tool for profiling borrower risk. Consequently, concomitant institutional improvements are essential to fully realise the benefits of data elements and to leverage China's large-scale data advantage across diverse application scenarios.

Key words: information asymmetry, regression discontinuity, sufficient statistics, welfare analysis

JEL codes: D82, G14, D6

(截稿:2025年8月 责任编辑:王 徽)