Applied Economics Letters

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/rael20

Big data and regional digital innovation: evidence from China

Dongfa Feng, Yao Zhang, Yupeng Lu & Litao Duan

To cite this article: Dongfa Feng, Yao Zhang, Yupeng Lu & Litao Duan (05 Apr 2024): Big data and regional digital innovation: evidence from China, Applied Economics Letters, DOI: 10.1080/13504851.2024.2339376

To link to this article: https://doi.org/10.1080/13504851.2024.2339376

Published online: 05 Apr 2024.

APPLIED ECONOMICS LETTERS
https://doi.org/10.1080/13504851.2024.2339376

Routledge
Taylor & Francis Group

Big data and regional digital innovation: evidence from China

Dongfa Feng [a], Yao Zhang [a], Yupeng Lu [b] and Litao Duan [c]

[a]Institute of Digital Finance, National School of Development, Peking University, Beijing, China; [b]Fu Zhou Branch, China Mobile Communication Group Jiangxi Co. Ltd, China; [c]Institute of Quantitative and Technological Economics, Chinese Academy of Social Sciences, Dongcheng, China; The CASS Laboratory for Economic Big Data and Policy, Chinese Academy of Social Sciences, Dongcheng, China Evaluation

ABSTRACT

We investigate the causal effect of big data on regional digital innovation by using the staggered difference-in-differences method based on panel data from 281 cities in China between 2011 and 2019. We find that big data positively affects regional digital innovation, especially in eastern China, regions with low economic policy uncertainty, high online attention, and strong legal environment. This innovation-promoting effect is achieved through enriching market opportunities and agglomerating production factors. This paper provides new empirical evidence that big data promotes digital innovation.

KEYWORDS

Big data; digital innovation; market opportunities; production factors

JEL CLASSIFICATION

O30; O38; R58

I. Introduction

In current digital era, digital innovation serves as a key driver for regional economic growth (Pedota 2023). Stimulating local digital innovation vitality is a crucial objective for governments. At the same time, big data is regarded as a vital raw material driving the development of the digital economy (Chen and Zhang 2014). A natural question arises: Can government encouragement of big data development promote regional digital innovation? In theory, the abundance of big data at the regional level can generate more digital entrepreneurial opportunities, enabling greater participation of market entities in digital innovation activities (Sahut, Iandoli, and Teulon 2021). Moreover, enhancements in enterprise big data analytics capabilities can also elevate innovation prowess (Mikalef and Krogstie 2020; Mikalef et al. 2019, 2020). However, the question of whether and how big data can promote regional digital innovation has not been fully explored in empirical research, particularly in the context of developing countries.

The main difficulties in exploring the causal relationship between big data development and regional digital innovation lie in two main aspects: Firstly, in the data source, scholars cannot accurately measure the unit's big data stocks. Secondly, in the identification strategy, there is a reverse causality between big data and regional digital innovation, wherein an increase in digital innovation helps regions gather more big data. Furthermore, both big data and regional digital innovation might be influenced by some common unobservable factors. These difficulties make the result obtained from ordinary least square estimation unreliable.

China, a typical centralized country, launched a pilot policy named National Comprehensive Big Data Pilot Zones (NCBDPZ) in 10 provinces and 57 cities in two rounds in 2015 and 2016, providing an ideal quasi-natural experiment. This policy urges pilot areas to open public data, build big data centres, and reform big data systems, resulting in an exogenous increase in local big data stocks. Recent studies utilize this exogenous shock to handle the difficulties mentioned above (Wang et al. 2023; Wei, Jiang, and Yang 2023). Following this identification strategy, we analyse the overall impact, mechanism, and heterogeneity of big data on regional digital innovation based on a staggered difference-in-differences (DID) setting and validate the robustness of regression results with a battery of tests. We find that the development of big data significantly promotes regional digital innovation. This effect is stronger in eastern China, areas with lower economic policy

CONTACT Litao Duan duanlt@cass.org.cn Institute of Quantitative and Technological Economics, Chinese Academy of Social Sciences, Dongcheng, China

uncertainty, higher online attention, and strong legal environments. Moreover, this promoting effect is achieved through enriching market opportunities and aggregating production factors.

This study enriches existing literature in two aspects. First, it provides new empirical evidence that big data promotes regional digital innovation based on data from China. Previous studies have discovered that companies in developed countries can stimulate innovation by enhancing their big data analytics capabilities, such as Greece (Mikalef et al. 2019, 2020) and Norway (Mikalef and Krogstie 2020). However, literature discussing regional digital innovation in developing countries remains relatively scarce. Second, our study enriches the literature on the economic and social effects of big data development. Previous studies have discussed the impact of big data development on the sustainable development of enterprises (Wang et al. 2023) and regional low-carbon development (Wei, Jiang, and Yang 2023). However, to the best of our knowledge, due to difficulties in data acquisition, there has been no discussion of its impact on regional digital innovation.

II. Research design

Empirical model

To investigate the causal relationship between big data and digital innovation, we view the establishment of the NCBDPZ as an exogenous shock on local big data stocks and implement a staggered DID estimate:

$$DigitInnov_{it} = \alpha + \beta \times BigData_{it} + \boldsymbol{Z}_{it}\delta + \mu_i + \lambda_t + \varepsilon_{it} \quad (1)$$

Where $DigitInnov_{it}$ is digital innovation for city i in year t, which measured by the digital invention patents granted per 10,000 people. A patent will be identified as a digital patent when its international patent classification belongs to the *Reference Relationship Table between the Digital Economy Core Industry Classification and the International Patent Classification (2023)*. $BigData_{it}$ is a dummy that equals one if city i is affected by the NCBDPZ at year t and zero otherwise. Following Feng et al. (2023), We add a series of control variables $\boldsymbol{Z}_{it}$ that may affect regional digital innovation vitality. For example, we include economic development (*EcoDev*), measured using real GDP per capita, as higher economic development levels generally enhance market returns for innovation outcomes; population density (*PopDen*), measured by the ratio of population to area, as higher population density leads to scale effects that promote innovation vitality; financial development (*Finance*), measured by the ratio of the loan balance to GDP, as innovation activities rely on financial support; research and development investment (*R&D*), measured by the percentage of science expenditure in fiscal expenditure, as increased investment in R&D generally leads to more innovation output; openness (*Open*), measured by the percentage of foreign direct investment in GDP, as higher degrees of openness leads to easier absorb foreign technologies; marketization level (*Market*), measured by the ratio of GDP to fiscal expenditure, as regions with higher levels of marketization are more likely to generate innovation. More details on these variables are reported in the left half of Table 1.

Data

Patent data was from the PatSnap database. The core dependent variable is manually curated. Control variables were from the *China City Statistical Yearbook*. We first collected all available

Table 1. Definitions, constructions, and descriptive statistics of key variables.

Variable	Description	Construction	Obs.	Mean	Std.	Min	Median	Max
DigitInnov	Regional digital innovation	Digital invention patents granted per 10,000 people	2529	2.215	8.533	0	0.288	150.127
BigData	Treatment status	A dummy variable equals 1 if the city is affected	2529	0.093	0.290	0	0	1
EcoDev	Economic development	Real GDP per capita	2529	0.553	0.498	0.080	0.397	4.810
PopDen	Population density	The ratio of population to area	2529	0.443	0.344	0.005	0.365	2.759
Finance	Financial development	The ratio of the loan balance to GDP	2529	0.985	0.614	0.118	0.811	9.622
R&D	R&D investment	The percentage of science expenditure in fiscal expenditure	2529	1.654	1.668	0.060	1.154	20.683
Open	Openness	The percentage of foreign direct investment in GDP	2529	1.718	1.752	0.000	1.220	19.880
Market	Marketization level	The ratio of GDP to fiscal expenditure	2529	6.043	2.466	1.092	5.730	22.789

sample data, then excluded 19 cities with severe data missingness, such as Chaohu, Laiwu, and Sansha. The final sample is a balanced panel data from 281 cities between 2011 and 2019. We report the descriptive statistics of key variables in the right half of Table 1.

III. Empirical results

Baseline estimate

We estimate the empirical model and report the results in Table 2. It shows that *BigData* is consistently significantly positive at a 5% level with or without control variables, indicating that the big data, made richer by NCBDPZ impacts, can boost regional digital innovation. According to column (2), big data can increase digital invention patents granted per 10,000 people by 1.472 compared to the untreated cities.

Mechanism tests

We propose that big data can promote regional digital innovation through enriching market opportunities and agglomerating production factors. To verify the market opportunities channel, we use three different indicators: the number of digital companies per capita (*DigComCount*), the scale of digital companies per capita (*DigComScale*), and the spirit of innovation (*InnovSpirit*) to measure regional market opportunities. To verify the factor agglomeration channel, we use the proportion of ICT employees to measure the degree of talent agglomeration (*Talent*) and use the 'outside investment' sub-index of IRIEDEC, complied by Dai et al. (2022) to measure the degree of capital agglomeration (*Capital*). We use these variables as dependent variables and report the results in Table 3. It shows that the coefficients of *BigData* are significantly positive, suggesting that the market opportunities channel and the factor agglomeration channel are verified.

Table 2. Baseline estimate results.

Variables	(1) *DigitInnov*	(2) *DigitInnov*
BigData	2.845**	1.472**
	(1.286)	(0.684)
EcoDev		15.686***
		(2.904)
PopDen		45.819***
		(8.195)
Finance		0.461**
		(0.231)
R&D		0.311**
		(0.152)
Open		−0.302***
		(0.107)
Market		−0.544***
		(0.123)
Obs.	2529	2529
Adj. R^2	0.873	0.938

Standard errors in parenthesis are clustered at city level. *, **, and *** denote significant at the 10%, 5%, and 1% levels respectively. All columns include individual fixed effects and time-fixed effects.

Heterogeneous effects

To explore the heterogeneous effects of big data on regional digital innovation, we divide all cities into two groups based on four different classification methods and implement subsample regression. Considering the significant regional disparities in the Chinese economy, with the eastern coastal regions being more developed compared to non-eastern regions, our first heterogeneity analysis divides cities into eastern and non-eastern groups based on geographical location. The corresponding regression results are reported in columns (1) and (2) of Table 4. The test for differences in coefficients between the groups indicates a significant disparity, with big data only promoting digital innovation in the eastern regions. This suggests a strong correlation between digital innovation and the local level of economic development.

Considering that one of the mechanisms through which big data promotes regional digital innovation is by enriching market opportunities, and economic policy uncertainty affects the realization of market opportunities, we categorize all cities into high uncertainty and low uncertainty groups based on the median of the China provincial economic policy uncertainty index compiled by Yu et al. (2021). The corresponding regression results are reported in columns (3) and (4) of Table 4. The test for differences in coefficients between the groups indicates a significant disparity, with big data only promoting digital innovation in the regions with lower economic policy uncertainty. This suggests that lower economic policy uncertainty facilitates the realization of market opportunities and leads to more digital innovation.

Considering that big data promotes digital innovation by attracting external digital talent,

Table 3. Mechanism test results.

	(1)	(2)	(3)	(4)	(5)
	Market opportunities channel			Production factors channel	
Variables	*DigComCount*	*DigComScale*	*InnovSpirit*	*Talent*	*Capital*
BigData	9.178**	16.276**	1.801**	0.214***	2.849*
	(4.076)	(7.602)	(0.910)	(0.074)	(1.633)
Obs.	2529	2529	2520	2529	2529
Adj. R^2	0.931	0.993	0.894	0.750	0.829

Table 4. Heterogeneous test results.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Geographical location		Economic policy uncertainty		Online attention		Legal environment	
	Non-eastern	Eastern	Low	High	Low	High	Weak	Strong
BigData	−0.066	2.931**	1.972**	0.361	−0.125	2.560***	0.092	6.167***
	(0.229)	(1.426)	(0.873)	(0.838)	(0.116)	(0.975)	(0.174)	(2.357)
Empirical p-value	0.000***		0.000***		0.000***		0.000***	
Obs.	1629	900	1413	1116	1260	1269	1053	1476
Adj. R^2	0.763	0.956	0.959	0.880	0.784	0.949	0.726	0.953

The dependent variables of all columns are *DigitInnov*. The empirical p-values are obtained by bootstrap 1000 times.

and regions with higher attention are more likely to attract talent, we categorize all cities into high online attention and low online attention groups based on the median of the Baidu Index corresponding to cities' name. The regression results, reported in columns (5) and (6) of Table 4, show that the difference between different group is statistically significant, and big data only promoting digital innovation in the regions with higher online attention. This suggests that regions with higher attention are more prone to digital innovation.

Considering that innovation outcomes require protection by a favourable legal environment, we categorize all cities into groups with strong and weak legal environments based on the strength of intellectual property protection in each city. The regression results, reported in columns (7) and (8) of Table 4, show that the difference between different group is statistically significant, and big data only promoting digital innovation in the regions with strong legal environment. This suggests that regions with strong legal environment are more prone to digital innovation.

IV. Conclusion

This paper utilizes the NCBDPZ to explore the causal relationship between big data and regional digital innovation. We find that big data can promote regional digital innovation through market opportunities channel and factor agglomeration channel. We also document that in the areas located in eastern China, areas with low economic policy uncertainty, high online attention, and strong legal environment, big data has a more substantial innovation promotion effect.

This paper theoretically explores the mechanism by which big data promotes regional digital innovation, while also providing practical implications. On one hand, we argue that local governments can foster digital innovation by enriching local big data reserves. On the other hand, to fully leverage the innovation-promoting role of big data, local governments also need to focus on reducing economic policy uncertainty, enhancing online attention, and strengthening the legal environment.

The paper has two main limitations. Firstly, in this paper, pilot policy shocks are used to proxy the level of regional big data development. In future research, more suitable methods could be explored to directly measure local big data reserves or development levels, thereby more precisely assessing the role of big data. Secondly, this paper utilizes regional-level data for analysis. In the future, using firm-level data could provide a more detailed examination of the key role played by big data in the production and operations of enterprises.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the National Natural Science Foundation of China under Grant [72273005, 72241421], and the CASS Laboratory for Economic Big Data and Policy Evaluation (Project No. 2024SYZH004)

ORCID

Dongfa Feng http://orcid.org/0000-0003-0246-7643
Yao Zhang http://orcid.org/0000-0003-3367-8697
Litao Duan http://orcid.org/0000-0003-4850-4980

References

Chen, C. P., and C. Y. Zhang. 2014. "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data." *Information Sciences* 275:314–347. https://doi.org/10.1016/j.ins.2014.01.015.

Dai, R., A. Wang, and B. Chen. 2022. "Innovation and Entrepreneurship in Core Industries of Digital Economy in China: Stylized Facts and Index Compilation." *Economic Perspectives* 4:29–48. in Chinese.

Feng, D., Y. Shen, X. Xie, and Y. Huang. 2023. "Digital Economy and Carbon Emission Reduction: Evidence from China." *China Economic Journal* 16 (3): 272–301. https://doi.org/10.1080/17538963.2023.2244276.

Kong, D., and N. Qin. 2021. "China's Anticorruption Campaign and Entrepreneurship." *The Journal of Law and Economics* 64 (1): 153–180. https://doi.org/10.1086/711313.

La Ferrara, E., A. Chong, and S. Duryea. 2012. "Soap Operas and Fertility: Evidence from Brazil." *American Economic Journal: Applied Economics* 4 (4): 1–31. https://doi.org/10.1257/app.4.4.1.

Mikalef, P., M. Boura, G. Lekakos, and J. Krogstie. 2019. "Big Data Analytics Capabilities and Innovation: The Mediating Role of Dynamic Capabilities and Moderating Effect of the Environment." *British Journal of Management* 30 (2): 272–298. https://doi.org/10.1111/1467-8551.12343.

Mikalef, P., M. Boura, G. Lekakos, and J. Krogstie. 2020. "The Role of Information Governance in Big Data Analytics Driven Innovation." *Information & Management* 57 (7): 103361. https://doi.org/10.1016/j.im.2020.103361.

Mikalef, P., and J. Krogstie. 2020. "Examining the Interplay Between Big Data Analytics and Contextual Factors in Driving Process Innovation Capabilities." *European Journal of Information Systems* 29 (3): 260–287. https://doi.org/10.1080/0960085X.2020.1740618.

Pedota, M. 2023. "Big Data and Dynamic Capabilities in the Digital Revolution: The Hidden Role of Source Variety." *Research Policy* 52 (7): 104812. https://doi.org/10.1016/j.respol.2023.104812.

Sahut, J. M., L. Iandoli, and F. Teulon. 2021. "The Age of Digital Entrepreneurship." *Small Business Economics* 56 (3): 1159–1169. https://doi.org/10.1007/s11187-019-00260-8.

Sun, L., and S. Abraham. 2021. "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects." *Journal of Econometrics* 225 (2): 175–199. https://doi.org/10.1016/j.jeconom.2020.09.006.

Wang, W., H. Zhang, Z. Sun, L. Wang, J. Zhao, and F. Wu. 2023. "Can Digital Policy Improve Corporate Sustainability? Empirical Evidence from China's National Comprehensive Big Data Pilot Zones." *Telecommunications Policy* 102617–21. https://doi.org/10.1016/j.telpol.2023.102617. 47 9

Wei, X., F. Jiang, and L. Yang. 2023. "Does Digital Dividend Matter in China's Green Low-Carbon Development: Environmental Impact Assessment of the Big Data Comprehensive Pilot Zones Policy." *Environmental Impact Assessment Review* 101:107143. https://doi.org/10.1016/j.eiar.2023.107143.

Yu, J., X. Shi, D. Guo, and L. Yang. 2021. "Economic Policy Uncertainty (EPU) and Firm Carbon Emissions: Evidence Using a China Provincial EPU Index." *Energy Economics* 94:105071. https://doi.org/10.1016/j.eneco.2020.105071.

Appendix A. Robustness tests

A.1. *Pre-event parallel trends tests*

Following La Ferrara et al. (2012), we examine the pre-event parallel trends based on the event study method. Figure A1a shows that the coefficients are insignificant at a 5% level in pre-event periods, suggesting no significant difference between the treatment and control groups before the establishment of the NCBDPZ. However, Sun and Abraham (2021) proved that the coefficient of dynamic treatment term has no direct meaning while heterogeneous treatment effects exist. To eliminate the potential impact of this issue, we re-implemented this test using the method proposed by Sun and Abraham (2021). Figure A1b presents the new test results, which are very similar to Figure A1a and support the robustness of the baseline result.

A.2. *Policy exogeneity test*

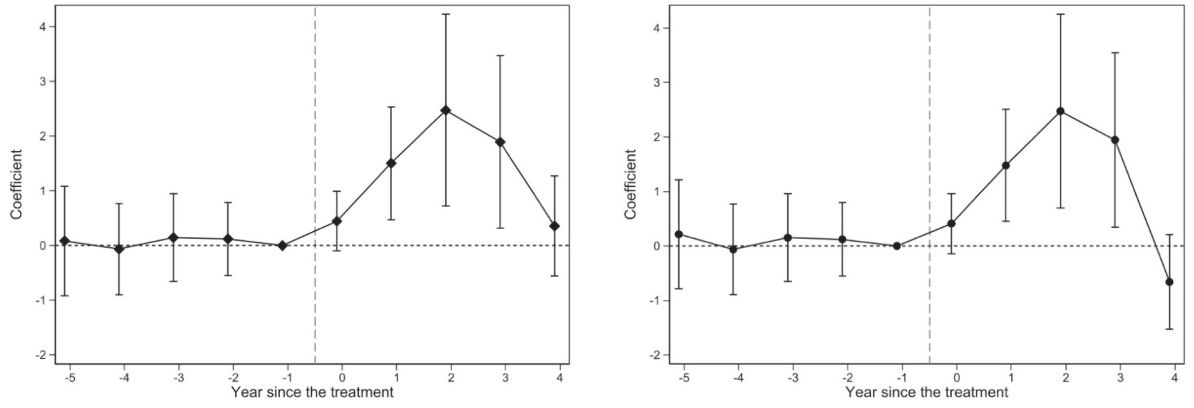
Our identification strategy assumes that the pilot policy is exogenous. If this hypothesis holds true, innovation entities should not be able to anticipate the occurrence of the policy or change their innovation levels before the introduction of the NCBDPZ. Conversely, if they do, it suggests that the hypothesis is not valid. To test this assumption, we introduce two dummy variables in the empirical model, representing the first period (*L1event*) and second period (*L2event*) before the event, respectively, and report the results in columns (1) and (2) of Table A1. It shows that these dummy variables are insignificant, indicating that cities cannot anticipate the implementation time of NCBDPZ before the event, suggesting that the policy has a certain exogeneity. Therefore, our baseline regression results should reflect the causal relationship between big data development and regional digital innovation.

A.3. *Considering unobservable factors*

Following Kong and Qin (2021), we implement two tests to mitigate the interference of unobservable factors. The first introduces the interaction between city-fixed effects and time trends in the empirical model. The result, reported in column (3) of Table A1, shows that *BigData* is significant at a 10% level. The second one permutes *BigData* with a random variable and repeats 500 times. Figure A2 shows that the kernel distribution of coefficients is around 0 and is like the normal distribution; most p-values exceed 0.05; and the baseline estimate falls outside the kernel curve. In summary, our baseline results are robust to considering unobservable factors.

A.4. *More robustness tests*

Firstly, to alleviate the measurement error in the dependent variable, in columns (4) and (5) of Table A1, we replace the dependent variable with digital invention patents applied per 10,000 people (*DigitApply*) and digital patents granted per 10,000 people (*DigitGrant*), respectively. Secondly, considering a time lag between policy announcing and implementation, in column (6) of Table A1, we replace the key independent variable with its time lead term (*F.BigData*). Thirdly, to mitigate the self-selection bias, in column (7) of Table 3, we re-estimate the model with the PSM-DID method. Lastly, to exclude the impact of other innovation incentive policies, in columns (8) and (9) of Table A1, we introduce two dummy variables representing 'Broadband China' (*BroadbandChina*) and 'Social Credit System Construction Pilot' (*SocialCredit*), respectively. Overall, the coefficients of big data are significant, indicating that the baseline result is robust.



a. Test using La Ferrara et al. (2012) b. Test using Sun and Abraham (2021)

Figure A1. Pre-event parallel trends test results.

Table A1. Robustness test results.

Variables	(1) *DigitInnov*	(2) *DigitInnov*	(3) *DigitInnov*	(4) *DigitApply*	(5) *DigitGrant*	(6) *DigitInnov*	(7) *DigitInnov*	(8) *DigitInnov*	(9) *DigitInnov*
BigData	1.462**	1.488**	0.640*	1.891**	3.076**		0.928*	1.471**	1.362**
	(0.736)	(0.685)	(0.334)	(0.837)	(1.381)		(0.518)	(0.680)	(0.676)
L1event	−0.045								
	(0.362)								
L2event		0.078							
		(0.289)							
F.BigData						1.775**			
						(0.741)			
BroadbandChina								0.769**	
								(0.332)	
SocialCredit									2.274***
									(0.785)
city × *trend*	No	No	Yes	No	No	No	No	No	No
Obs.	2529	2529	2529	2529	2529	2529	2451	2529	2529
Adj. R^2	0.938	0.938	0.986	0.940	0.937	0.938	0.869	0.938	0.940

Standard errors in parenthesis are clustered at city level *, **, and *** denote significant at the 10%, 5%, and 1% levels, respectively. All columns include individual fixed effects, time-fixed effects, and control variables. The same is below.

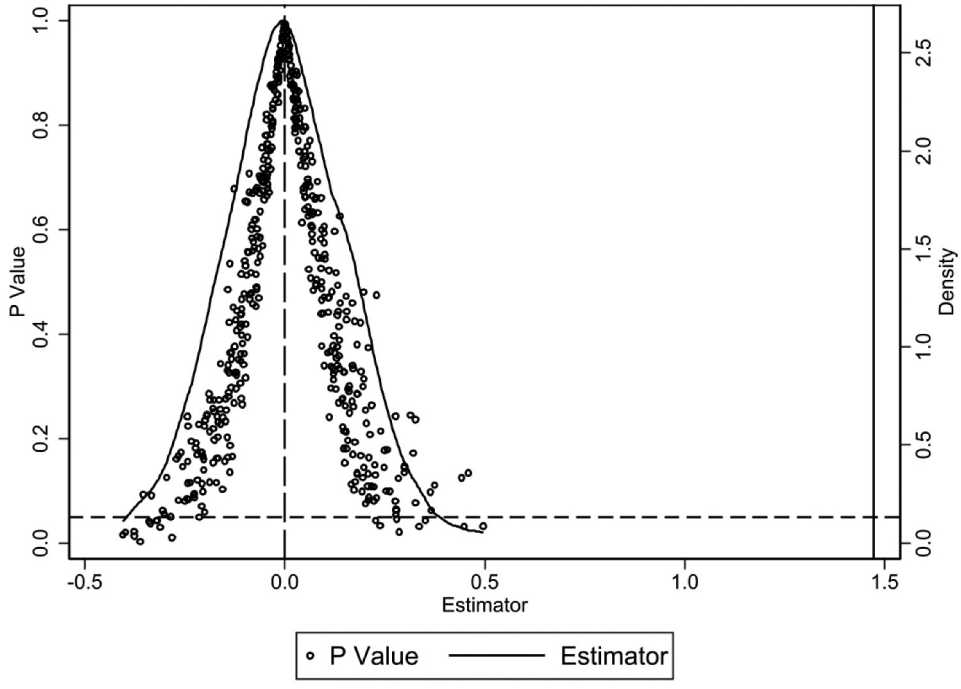


Figure A2. Permutation test result.