Practical Application of the Models for Rectifying Selection Bias

Xiaohong Yu, Maonan Chen*
yuxiaohong@hit.edu.cn, *Corresponding author: tai70743060@163.com

Harbin Institute of Technology (Shenzhen) School of Economics and Management Shenzhen, China

Abstract—A sample of the annual data for A-share listed companies in China between 2010 and 2018 were used in this paper to explore the relationship between the native place of the chairman and company innovation with using multiple models for the solution of selection bias. The results of multiple models all led to the following consistent conclusion: When the chairman's native place was the same as the place of incorporation of the company, the company innovation capabilities as measured with indicators, such as patent numbers, R&D personnel numbers and R&D investment proportions, were all impaired. The results were consistent with multiple models that included the treatment effects model and the endogenous switching regression model. The same conclusions hold after further analyses of the subsequent treatment effects on the experimental group and the control group.

Keywords-native place of chairman; company innovation; selection bias; endogenous switching regression model

1 INTRODUCTION

Many problems can emerge in economic empirical research, including the problem of selection bias which is one of the types that are more difficult to be dealt with. Generally speaking, selection bias (or selection endogeneity) is usually caused by the following two factors: First, the data involved in the research cannot fully reflect study samples due to external (incomplete data because of secret information, administrative orders and the lack of the data in the database) and internal (the sampling methods or the selection of targeted industries or years) reasons. This situation is referred to as “sample selection bias” [1]. Second, the situation that is characterized by independent variables used in the research results from conscious behaviors of individuals in the study. The selection bias caused by this latter situation is called the "self-selection problem” [2]. The problem of self-selection leads not only to deviations of the coefficients obtained through OLS regression, but also erroneous judgements of causality between independent variables and dependent variables in the research.

Selection endogeneity is also a common problem in the financial research of listed companies. Since the selection bias has a significant impact on research results, many researchers have carried out in-depth studies from multiple perspectives. Although propensity score matching

ICEMME 2022, November 18-20, Nanjing, People's Republic of China

DOI 10.4108/eai.18-11-2022.2326849

(PSM) has been widely applied in order to narrow selection bias, Dehejia et al.[3] pointed out a shortcoming of this method. Since PSM could only narrow the bias resulting from observable factors, the influence brought by unobservable factors in the research was ignored. Heckman [1] proposed the Heckman two-step method to correct the sample selection bias. Generally speaking, the first step of this method is to forecast the tendency or probability of individuals in the study to follow a certain behavior pattern, and the second step is to calculate Imr(inverse Mills ratio) then put it into the regression for the estimation of the parameters. The treatment effects model is targeted at the self-selection problem. Its biggest difference from the Heckman two-step method is that the modeling of binary independent variables can be directly constructed and all data can be observed simultaneously [4]. The endogenous switching regression model proposed by Maddala [5] argued that the hypothesis of including all potentially influential factors in the equation during regression should be relaxed. This approach was essentially an extension of the Heckman two-step method. The advantage of the endogenous switching regression model is that it is possible to observe whether the same variable will perform differently in different groups, and it is possible to observe the net treatment effects involved in different groups. This advantage makes it possible to better determine the presence of causal effects.

The models noted above were adopted, therefore, to solve both sample selection bias and self-selection bias while analyzing the relationship between the chairman's native place and company innovation performance in this paper. The major contributions of this paper include the focus on the models and expands on their application for solving issues of selection bias in the financial research of listed companies. Further, it provides the reference for corporate governance.

2 THEORETICAL ANALYSIS AND HYPOTHESIS

Personal characteristics of managing executives have been researched for a long time. Prior research has indicated that overconfidence of corporate managing executives has an impact on the tendency of companies to avoid taxes. Overconfident executives are prone to cooperate with all related parties and to attempt to find measures to avoid taxes [6]. Gender has been an additional factor that is important for explaining the decisions of executive. Francis argued in his research that companies managed by female CEOs have less tolerance for risk-taking than firms with male CEOs [7]. Other research studies have indicated that in addition to the evidence that gender and the degree of self-confidence of executives affects the formulation of corporate policies, the age of the company's executives and whether they have overseas experiences also can have an impact on their management practices [8,9]. The results from in-depth studies of the individual characteristics of executives have led researchers to shift their focus to factors that are not individual characteristics but shared characteristics. For example, whether executives belong to the same society or the same ethnic group or whether they have studied in the same school can be important. This type of characteristic is not considered to be a “personal characteristic” but “identity acknowledgement” [10,11]. On the other hand, some scholars consider these characteristics to be an indication of the existence of an informal system. Existing studies have attached great importance to the role played by this kind of identity acknowledgement or informal systems. In addition to informal systems, Boiral et al. [12] found that the ideology of managing executives significantly affects the extent of the company's environment-friendly practices.

According to the research of Vaske et al. [13], this kind of feeling towards hometown nurtures and fosters an emotional bond among people. Further research has confirmed that this type of informal system may be to some degree as binding as any formal system. The studies carried out by Scannell et al. [14] argued that after controlling the influences of other key factors personal emotions toward hometown were significantly correlated to environment-friendly behavior of individuals. Specifically, the stronger the individuals' feelings towards their hometowns, their attitudes are more positive towards the environment in that town. It is worth noting that this kind of emotional attachment to hometown does not only play a role in views of regional pollution treatment and control but also profoundly affects the behavior of individuals in every respect [15,16].

After local chairmen taking control of the company, due to their better knowledge of the local business environment and excellent social resources they have, they may shift their focus from strengthening corporate business and innovation capabilities to maintaining their social relations. They may give a higher priority to the generation of excessive returns for the company through their own social capital and political connections. When a company has to rely on its chairman's social capital and political connections for profits, its innovation investment and innovation performance will inevitably be harmed. On the other hand, since chairmen from other places lack local political connections and social resources, the companies managed by they will pay more attention in promoting business capabilities and increasing company's R&D investment and efficiency than those firms managed by local chairmen. In order to compete at the same level as companies run by local chairmen, companies with chairmen from other places must provide products with higher quality and must have a better innovation performance. Companies controlled by chairmen from other places, therefore, will attach more importance to innovation than those managed by local chairmen. Given these possibilities, the following hypothesis was proposed:

When the chairman's native place is the same as the company's place of incorporation, the company innovation capability will be impaired resulting in a worse innovation performance.

3 RESEARCH DESIGN AND MODEL DESCRIPTION

3.1 Source and processing of the data

The samples for this research were drawn from the annual data between 2010 and 2018 for Chinese A-share listed companies. All the data came from the CSMAR and WIND databases. In addition, the data were processed as follows: (1) All companies in the financial sector were excluded while selecting the samples due to the special characteristics of the financial industry. (2) Those observations without financial data were excluded; (3) Companies with ST and PT status were excluded. (4) In order to avoid the adverse effects of possible extreme observations in the sample on the results, all continuous variables were subjected to a 1% and 99% winsorization. Since information for the native place of chairmen was incomplete, the data was supplemented by collecting the data manually. STATA was adopted in this paper as the statistical software.

3.2 Variable definitions

When the chairman's native place is at the same place as the company's registration place, the variable Native_if was taken as 1, otherwise 0. Because of the characteristics of the data available in the database, most of the information about the native places of the chairmen could only be traced at the level of provincial administrative region; therefore, the chairmen's native places and the companies' registration information were chosen at this level. However, in order to avoid major inconsistencies between the chairmen's native places and their birthplaces (native place refers to the place where an individual's grandfathers and older generations permanently lived or the place where they were born), Birth_if, a variable indicating the chairmen's birthplaces, was also used.

Based on the prior studies [17,18], the natural logarithm of the number of patent application of the company in a corresponding year plus 1 was used as the measure of Innovation_patent, the company's innovation capability. In addition, in order to fully reflect the innovation capability and innovation investment of a company, Innovation_spend (the natural logarithm of the company's total R&D investment in a corresponding year) and Innovation_ratio (the proportion of R&D investment to operating incomes) were also used to measure the company's innovation investment. An effort was made to judge how important companies considered their R&D personnel to be; therefore, Innovation_person, the natural logarithm of the value obtained by adding the number of R&D personnel with 1, and Innovation_pratio, the proportion of R&D personnel to total employees, were used to measure the company innovation capability. Obviously, the smaller the values of the above-mentioned indicators, the lower the companies' investment in innovation, and the less importance the companies had attached to innovation.

In addition, in keeping with previous research the following control variables (Cntrolvar) [19-23] were selected : (1) Da, corporate asset-liability ratio, measured as corporate total liabilities/corporate total assets; (2) Age, age of the company, measured by from time the company went public; (3) Share, the proportion of the shares held by the company's largest shareholder; (4) Multiple_if, valued as 1 if the company has multiple major shareholders (defined under current Chinese laws and regulations in China when the company has two or more shareholders who possess 10% or more shares of the company); (5) Multiple_n, the number of major shareholders who possess 10% or more shares of the company; (6) Growth, corporate growth rate, measured by the growth rate of total profits for the company; (7) Lev, corporate financial leverage measured by the formula (net profits + income tax expenses + financial expenses) / (net profits + income tax expenses); (8) TQ, the value of the company measured by Tobin Q; (9) Capital, corporate capital intensity measured by corporate total assets / corporate operating incomes; (10) Size, the size of the company measured by the natural logarithm of total assets of the firm.

3.3 Benchmark model regression

In order to test the hypothesis proposed above, the following OLS model was constructed as the benchmark model for the analysis:

$$Innovation_{i,t} = \alpha_1 + \beta_1 \times Native\_if_{i,t} + \sum \beta_i \times Cntrolvar_{i,t} + \sum YEAR + \sum INDUSTRY + \varepsilon_{i,t} \tag{1}$$

Innovation represents the various methods that measuring corporate innovation performance as described above. $\sum YEAR$ and $\sum INDUSTRY$ stand for the annual and industrial fixed effects. The OLS regression results were served as the starting point for the analysis of the results of the subsequent models. As noted above, if selection bias is present in the research, there will be the deviations of the estimated coefficients in OLS regression.

3.4 Propensity score matching (PSM)

In the research, PSM was first used to narrow the deviations of regression results caused by measurable factors. The companies whose places of incorporation and their chairmen's native places are at the same place were labelled as the experimental group (treatment group), while the companies whose places of incorporation are different from their chairmen's native places were considered to be the control group (untreated group). This division matched the logit model that was shown below that was designed to determine that there was no significant difference between the experimental group and the control group. It should be noted that in the process of propensity score matching, nearly no dependent variable was involved.

$$Native\_if_{i,t} = \alpha_1 + \sum \beta_i \times Cntrolvar_{i,t} + \sum YEAR + \sum INDUSTRY + \varepsilon_{i,t} \tag{2}$$

After the matching was completed, it was necessary to examine the matching results. Generally, the values for each variable between the experimental group and the untreated group deviated dramatically before the matching. After the matching was done, the deviations of all or most of the variables should be reduced to below the appropriate level of significance that would indicate an effective treatment during the matching. The kernel density function diagrams before and after matching will be reported in the research if necessary. This step was done for the following reasons. First, it is valuable to directly show the changes in deviations before and after matching and to judge the direction of the "policy treatment" effect. Of course, the matching must rely on the bootstrap method for accurate calculation of various parameters involved in the policy treatment effect, because the automatically estimated parameters in STATA might not be accurate (the assumption conditions of homoscedasticity were too strict) [24,25]. Specifically, treatment effects can be divided into the following categories: (1) average treatment effect on treatment group (ATT), representing the expected changes for the treatment group after the treatment; (2) average treatment effect on the untreated group (ATU), representing the expected changes for the control group after the treatment; and (3) average treatment effect on the population (ATE), representing the changes for all the samples after the treatment. Second, after the PSM, the samples that were not successfully matched could be excluded and main model (1) could subsequently be regressed again to determine whether the main model results were still valid.

3.5 Heckman two-step method

With the Heckman two-step method, in order to calculate the Imr, the first step is to use the probit model to predict the tendency or possibility of the individuals in the study making a certain choice. The second step is to calculate the parameters to be estimated in the regression with the Imr that is obtained. The first step of the Heckman two-step method resembles the PSM method since both rely on forecasts of the probability that an individual will undertake a certain action. Since the Heckman two-step method aims to overcome the influence of unobservable factors on the research results, however, it is necessary to incorporate “exclusive constraints” [26] in the prediction in the first stage. In other words, the deviations between the companies that do not disclose innovation data (or companies that are weak in innovation) and the companies that do (or companies that are strong in innovation) need to be corrected in the analysis.

Patent_if, a dummy variable, was constructed in this paper to measure the companies' patent situation. The variable was 1 when there was at least one patent application for the company in a corresponding year, otherwise it was 0. Additionally, Iv_ratio, a variable representing the average level of a company's innovation investment in the same region and year, was included in the prediction of the first stage to forecast whether the number of patent application of the company in that year was 0 or not. This variable was designed to determine whether there was a strong atmosphere favoring innovation. It was more likely that highly innovative companies would be successful in applying for patents. The reason for choosing Innovation_patent, representing the number of patent application in a corresponding year, as the dependent variable in the second step of the Heckman two-step method is that it was more representative for distinguishing its observation status (0 or otherwise). If other indicators (such as R&D investment) were used, however, their roles in distinguishing the observation status of dependent variables would be very limited (most of listed companies had more or less invested in R&D). The Heckman two-step model is as follows: $Patent\_if_{i,t}\begin{cases}1, if\ \ Innovation\_patent_{i,t} > 0\\ 0, if\ \ Innovation\_patent_{i,t} = 0\end{cases}$

$$Patent\_if_{i,t} = \alpha_1 + \beta_1 \times Iv\_ratio_{i,t} + \sum \beta_i \times Cntrolvar_{i,t} + \sum YEAR + \sum INDUSTRY + \varepsilon_{i,t}$$

$$Innovation\_patent_{i,t} = \alpha_1 + \beta_1 \times Native\_if_{i,t} + \beta_2 \times Imr_{i,t} + \sum \beta_i \times Cntrolvar_{i,t} + \sum YEAR + \sum INDUSTRY + \varepsilon_{i,t} \tag{3}$$

MLE is a default estimation method in STATA, but it is time-consuming. As a consequence, researchers can choose the two-step estimation method. This technique has the advantage of faster operational speed, but it is unable to adjust for heteroscedasticity or carry out clustering adjustments at the level of individual companies. In addition, the error in the first step will be brought into the estimation in the second step. While interpreting the results of the model, it is necessary to pay special attention to the Wald test results and whether the Imr coefficients are significant. If the Wald test results exceed the critical value, it indicates that the sample selection model is needed and that the estimation of OLS is biased. Likewise, if the Imr coefficients are significant, it indicates that sample selection bias has occurred.

3.6 Deformation of the Heckman two-step method

It can be seen from the above description that in terms of selection bias, the Heckman two-step method targeted the bias caused by the data selection problem (sample selection bias), instead of the selection bias that results from the conscious selection of the individuals (self-selection problem). In the case when the explanatory variable was an endogenous dummy variable, some researchers would use the Heckman two-step method to correct the self-selection problem [27,28]. Generally, the probit model would be used to predict the probability for the independent variable X to be 1, and then the Imr or Lambda coefficient (referred to as the Hazard coefficient in some portions of the literature but essentially the same) would be calculated. If the coefficient is found to be insignificant but there are significant independent variables in the final regression, there is no self-selection problem. If this coefficient is significant at the same time there are significant independent variables, then the conclusions from the main model remain valid after the self-selection problem has been controlled. As the Imr or Lambda coefficient is calculated manually (without following the existing commands in STATA), inaccurate standard errors might be obtained, thus this is affecting the coefficient significance. This manual estimation for the two stages has also been strongly criticized [29].

It is worth noting that some scholars view the deformed Heckman two-step method as equal to the original one. As mentioned above, however, the Heckman two-step method was used to deal with the failure of the sample to represent the population. The deformed Heckman two-step method was used to solve endogeneity issues with the independent variable X. It is not the recommendation of this paper that researchers use this method in practical research since it is still controversial.

In order to make comparisons convenient for analysis, the deformed Heckman two-step method was used for testing. Iv_area, the average employment of chairmen in their native place within the same region and year, and Iv_inds, the average employment of chairmen in their native place within the same industry and year, which were considered to be available to be used as instrumental variables. The model is as follows:

$$Native\_if_{i,t} = \alpha_1 + \beta_1 \times Iv\_area_{i,t} + \beta_2 \times Iv\_inds_{i,t} + \sum \beta_i \times Cntrolvar_{i,t} +$$
$$\sum YEAR + \sum INDUSTRY + \varepsilon_{i,t}$$
$$Innovation_{i,t} = \alpha_1 + \beta_1 \times Native\_if_{i,t} + \beta_2 \times Imr_{i,t} + \sum \beta_i \times Cntrolvar_{i,t} +$$
$$\sum YEAR + \sum INDUSTRY + \varepsilon_{i,t} \tag{4}$$

3.7 Treatment effects model

The condition for application of the treatment effects model is that endogeneity exists in the binary independent variables involved in the study. These independent variables in the study, therefore, were directly used for modeling in the first stage of the model, unlike the Heckman two-step method, which involved the observation state of the dependent variables. This method also demonstrated that the two models were aimed to solve different problems.

$$Native\_if_{i,t} = \alpha_1 + \beta_1 \times Iv\_area_{i,t} + \beta_2 \times Iv\_inds_{i,t} + \sum \beta_i \times Cntrolvar_{i,t} +$$
$$\sum YEAR + \sum INDUSTRY + \varepsilon_{i,t}$$
$$Innovation_{i,t} = \alpha_1 + \beta_1 \times Native\_if_{i,t} + \beta_2 \times Hazard_{i,t} + \sum \beta_i \times Cntrolvar_{i,t} +$$
$$\sum YEAR + \sum INDUSTRY + \varepsilon_{i,t} \tag{5}$$

It can be seen that models (4) and (5) are basically the same. It is hard to deny that the deformation of Heckman two-step method looks similar to the treatment effects model, but the calculation methods of the coefficients are completely different. Like the Heckman two-step method, the default MLE estimation method or the two-step estimation method that requires less time can be chosen with reference to the data characteristics of the research being undertaken. The interpretation of the results of the treatment effects model resembles that of the Heckman two-step method.

3.8 Endogenous switching regression model

The endogenous switching regression model has been most frequently used to analyze agricultural economic policies and family welfare circumstances [30]. More recently, however, it has been widely used in financial empirical research. As noted above, the biggest difference between the endogenous switching regression model and treatment effects model is the three equations, including a selection equation and two result equations in the experimental group and the control group:

$$Native\_if_{i,t} = \alpha_1 + \beta_1 \times Iv\_area_{i,t} + \beta_2 \times Iv\_inds_{i,t} + \sum \beta_i \times Cntrolvar_{i,t} + \sum YEAR + \sum INDUSTRY + \varepsilon_{i,t} \tag{6}$$

Equation (6) is a selection equation for determining whether a company would employ a person as chairman whose native place is the same as its place of incorporation. As mentioned above, Iv_area, the average employment of chairmen in their native place within the same region and year, and Iv_inds, the average employment of chairmen in their native place within the same industry and year, were used as the instrumental variables for prediction. When the variable Native_if was 1, the company innovation level was determined by the following formula (7), in which the T indicated that the company was in the treatment group:

$$Innovation_{Ti,t} = \alpha_{T1} + \sum \beta_{Ti} \times Cntrolvar_{Ti,t} + \sum YEAR + \sum INDUSTRY + \varepsilon_{Ti,t} \tag{7}$$

When the variable Native_if was 0, the company innovation level was determined by the following formula (8), where the U indicated that a company was in the control group:

$$Innovation_{Ui,t} = \alpha_{U1} + \sum \beta_{Ui} \times Cntrolvar_{Ui,t} + \sum YEAR + \sum INDUSTRY + \varepsilon_{Ui,t} \tag{8}$$

It can be seen that the variable Native_if is not in the result equation. Since it is impossible to observe the company innovation level in different situations at the same time, missing data was then determined by the endogenous switching regression model, the Imr was calculated and put into the equation, turning formula (7) and formula (8) into:

$$Innovation_{Ti,t} = \alpha_{T1} + \sum \beta_{Ti} \times Cntrolvar_{Ti,t} + \sum YEAR + \sum INDUSTRY + \sigma_T \times Imr_{Ti,t} + \varepsilon_{Ti,t} \tag{9}$$

$$Innovation_{Ui,t} = \alpha_{U1} + \sum \beta_{Ui} \times Cntrolvar_{Ui,t} + \sum YEAR + \sum INDUSTRY + \sigma_U \times Imr_{Ui,t} + \varepsilon_{Ui,t} \tag{10}$$

σ_T and σ_U are the covariance of the error term of the resulting equation and the selection equation. If they are found to be significant in the regression, there has been a problem of "simultaneous decision-making" in the research and the estimated coefficient of OLS was biased.

In the endogenous switching regression model, there were two counterfactual conditional expectations for the outcome variables besides the two factual conditional expectations of the outcome variables. Hence, the treatment effects can be expressed as the formula below:

$$\begin{gathered} ATT = E(Innovation_{Ti,t} \mid Native\_if_{i,t} = 1) - E(Innovation_{Ui,t} \mid Native\_if_{i,t} = 1) \\ ATU = E(Innovation_{Ti,t} \mid Native\_if_{i,t} = 0) - E(Innovation_{Ui,t} \mid Native\_if_{i,t} = 0) \end{gathered} \tag{11}$$

The ATE (the average treatment effects of the population) is the weighted average value of ATT and ATU. When the endogenous switching regression model was used, normally it was necessary for the significance level and specific values of ATT, ATU and ATE to be reported in addition to the results reported for the three equations. Further, the kernel density function diagram should be drawn according to the reported results.

4 EMPIRICAL RESULTS

4.1 Descriptive statistics

The descriptive statistics of this paper are shown in Table 1 below. It can be seen from the indicators that measure the company innovation level that they are all relatively low for the companies that the native places of the chairmen are the same as their places of incorporation. This result indicates that the employment of such chairmen would damage the company innovation capability. Hence hypothesis was preliminarily verified.

Table 1. Descriptive Statistics

	Native_if=0					Native_if=1				
variable	obs	mean	sd	min	max	obs	mean	sd	min	max
Innovation_person	815	5.9453	1.3641	2.7081	9.5373	1439	5.6873	1.1911	2.7081	9.5373
Innovation_pratio	815	0.1556	0.1332	0.0032	0.6504	1439	0.1413	0.1195	0.0032	0.6504
Innovation_patent	669	3.2463	2.0425	0.0000	8.3139	1018	2.6350	1.8689	0.0000	8.3139
Innovation_ratio	1718	0.0428	0.0406	0.0003	0.2324	3303	0.0395	0.0363	0.0003	0.2324
Innovation_spend	1718	18.1131	1.5614	14.3421	21.9443	3303	17.8150	1.3637	14.3421	21.9443

It can be seen from the Table 2 that the T-test of each explained variable strongly rejected the null hypothesis, indicating that the innovation levels for the two groups of companies were significantly different.

Table 2. Difference Testing Between Groups

variable	T value
Innovation_person	4.6840
Innovation_pratio	2.6032
Innovation_patent	6.3325
Innovation_ratio	2.9251
Innovation_spend	6.9858

4.2 OLS regression results

The regression results of Model 1 are as shown in Table 3: All the indicators that measure the innovation level of the company are negatively correlated with the variable Native_if. In other words, the company innovation capability would be damaged when the native place of its chairman was the same as its place of incorporation. Hence, hypothesis was completely verified.

Table 3. OLS Regression Results

VARIABLES	OLS	OLS	OLS	OLS	OLS
	Innovation_person	Innovation_pratio	Innovation_patent	Innovation_spend	Innovation_ratio
Native_if	-0.0839**	-0.0140***	-0.3410***	-0.1023***	-0.0032***
	(-2.1392)	(-2.8962)	(-3.8706)	(-3.6099)	(-3.3842)
Constant	-11.0099***	0.1118**	-13.3497***	-2.2255***	-0.0072
	(-21.1578)	(1.9792)	(-10.6637)	(-5.5943)	(-0.6723)
CONTROLS	YES	YES	YES	YES	YES
YEAR	YES	YES	YES	YES	YES
INDUSTRY	YES	YES	YES	YES	YES
Observations	2,254	2,254	1,687	5,021	5,021
R-squared	0.529	0.341	0.265	0.586	0.391

Note: The robust T values are in brackets; ***, ** and * represent significance at the levels of 1%, 5% and 10%, respectively.

The following robustness test has been conducted to verify the validity of the conclusions. First, the independent variable Native_if, the native place of the chairman, was replaced with Birth_if, the birthplace of the chairman before running the regression. Second, the independent variables were lagged for one period. Third, controls for the annual trend and fixed province effects were used in the regression to prevent the annual trend for the change of the company innovation level, as well as the differences in the feelings toward hometown at the provincial administrative level, having a confounding effect. The conclusion proved to be consistent with the previous conclusion from the analysis.

4.3 Results of the propensity score matching

The dependent variable Innovation_ratio was taken as an example for matching in this paper. The 1:2 nearest neighbor matching was used to prevent the difference of propensity scores between the two groups from exceeding 0.01. The table of differences between variables and the kernel density function diagram before and after matching are shown as follows. It can be seen that there were significant differences between most variables before matching, but that no significant differences were found among most variables after matching occurred. The two kernel density function diagrams indicate that the difference between the experimental group and the control group was greatly reduced after matching.

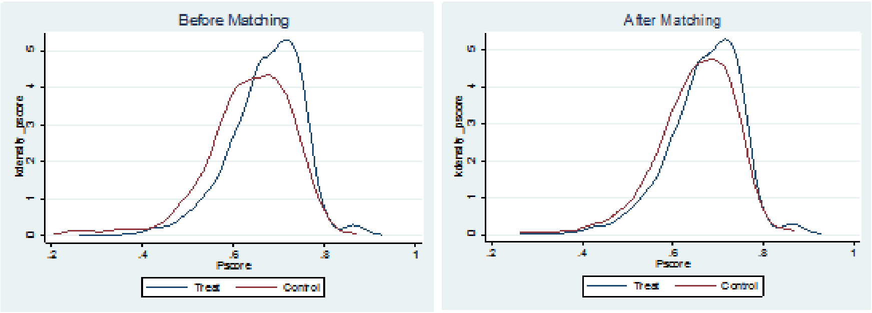

Figure 1. Kernel Density Function Diagram Before and After Matching

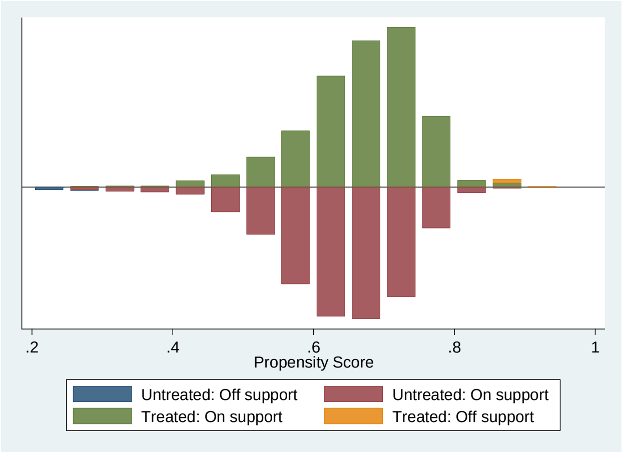

Figure 2. Mapping of Matching Results

The reported value of ATT was -0.002495329, which meant that if a company in treatment group employs another chairman from a different place, its R&D investment would be increased by approximately 0.2495%. The reported value of ATU was -0.00248879, which indicated that when a company with a chairman from a place other than its place of incorporation employs a local chairman, the R&D investment would be reduced by approximately 0.2489%. The reported value of ATE was -0.002493091 indicating that if all the companies employ chairmen from their places of incorporation, R&D investment would be reduced by 0.2493%. The standard error of ATT reported by STATA was 0.0014277. It is known, however, from the above analysis that the standard error obtained by using the bootstrap method was more reliable (even though not absolutely accurate). After 500 iterations of calculations by using the bootstrap method, the standard error of ATT was 0.001398. This result demonstrated that the standard error obtained by automatic calculation is to some extent different from the one obtained by using the bootstrap method. In addition, the bootstrap method also showed significant scores for ATT and ATE with an insignificant level for ATU. The successfully matched samples were used for the regression formula (1). It proved that the above conclusion was still valid.

4.4 Results of the Heckman two-step method

After the bias caused by observable variables was resolved, Innovation_patent was taken as the dependent variable to control for the sample selection bias by using the Heckman two-step method. The results of model (3) are as follows: whether either the MLE method or two-step method was used for estimation, sample selection bias was determined to be present in this study. The value of the Wald test reported by MLE method reached 54.05, providing for a strong rejection of the null hypothesis that there was no sample selection bias.

In addition, in the first stage of the process, the average value of R&D investment ratio of companies within the same year and region was discovered to have a significant positive correlation with the binary variables used to measure whether patent applications in that year was 0 or not. This result demonstrates that the regional innovation atmosphere affected the observed patent applications for companies (0 or not). In the second stage, Native_if was still found to have a significant negative correlation with the company innovation level. This finding provides evidence that the above conclusion was still valid after there were controls for the sample selection bias. As can be seen in the table, the results obtained by using the two methods are only slightly different. The OLS regression results with the same samples were also included for comparison. It is obvious that the estimated coefficients of the OLS are different from the estimated coefficients of the Heckman two-step method, and that the Heckman two-step method should be used to correct for the sample selection problem.

4.5 Results of the deformed Heckman two-step method

Based on the above-mentioned application of the Heckman two-step method, Iv_area and Iv_inds were added in the first stage to predict whether a company would employ a chairman whose native place is the same as its place of incorporation. The regression results of model (4) indicate that in the first stage, whether or not the company would employ a chairman whose native place is the same as the place of incorporation was affected by the industry tendency and regional tendency, both of which were significant at the 1% level. The Imr calculated according to the prediction results of the first stage was found to be insignificant in the second stage, and the Native_if still proved to have a negative correlation with the level of company innovation. These regression results indicate that there were no self-selection problems in this study. The corresponding defects of such a method have been studied in detail above.

Table 4. Difference Between Variables Before and After Matching

VARIABLES	Type	Mean		Bias (%)	T test	
		treated	control		T value	P value
Da	Unmatch	0.38009	0.40685	-13.2	-4.52	0.000
	Match	0.38064	0.37393	3.3	1.35	0.177
Age	Unmatch	8.5375	9.727	-18.1	-6.18	0.000
	Match	8.5173	8.551	-0.5	-0.21	0.834
Share	Unmatch	0.35225	0.36073	-5.6	-1.92	0.055

	Match	0.35249	0.35756	-3.4	-1.40	0.162
Multiple_if	Unmatch	0.45444	0.39988	11.0	3.70	0.000
	Match	0.4541	0.44373	2.1	0.84	0.399
Growth	Unmatch	0.71022	0.8631	-4.3	-1.46	0.145
	Match	0.71561	0.6533	1.8	0.76	0.448
Lev	Unmatch	1.3049	1.3141	-1.1	-0.38	0.701
	Match	1.3056	1.2831	2.8	1.11	0.266
TQ	Unmatch	2.0821	2.1517	-5.7	-1.93	0.054
	Match	2.0826	2.0927	-0.8	-0.35	0.726
Capital	Unmatch	2.107	2.1985	-7.1	-2.43	0.015
	Match	2.1026	2.1202	-1.4	-0.58	0.563
Multiple_n	Unmatch	1.5559	1.4738	11.5	3.85	0.000
	Match	1.555	1.5287	3.7	1.48	0.138
Size	Unmatch	22.139	22.437	-21.7	-7.63	0.000
	Match	22.141	22.127	1.0	0.41	0.679

4.6 Results of the treatment effects model

The regression results of model (5) are shown below in Table 7. Based on the results of the Wald test, the null hypothesis of “no self-selection problem” could not be rejected since there was an estimated P value reaching 0.99. According to the regression results from the first stage, there was a strong correlation between the probability of a company employing a local chairman and both the industry and regional tendencies. According to the regression results of the second stage, Native_if under the MLE estimation was still found to have a negative correlation with the company innovation level, while the estimation results when the two-step method was used remained insignificant. This finding may resulted from the loss of estimation efficiency present in the two-step method. The results in Table 6 indicate that the conclusion of the treatment effects model was consistent with that of the deformed Heckman two-step method. Even so, however, the regression coefficient of the deformed Heckman two-step method remained to some extent different from the coefficient obtained by the treatment effects model. This finding indicates that there is a defect with manual estimation.

Table 5. Results of the Heckman Two-Step Method

VARIABLES	Heckman twostep (MLE)		Heckman twostep (twostep)		OLS
	Patent_if	Innovation_patent	Patent_if	Inovation_patent	Innovation_patent
Iv_ratio	11.2211***		7.3863**		
	(4.1202)		(2.4790)		
Native_if		-0.1428**		-0.2023***	-0.3527***
		(-2.1309)		(-2.9677)	(-3.7733)

athrho		-1.2669***			
		(-7.3517)			
lnsigma		0.2281***			
		(6.9618)			
Imr				-1.1804*	
				(-1.8083)	
Constant	-3.9768***	-9.0672***	-4.6837***	-8.7891***	-13.1868***
	(-3.1832)	(-6.9889)	(-3.6411)	(-4.6485)	(-9.4254)
CONTROLS	YES	YES	YES	YES	YES
YEAR	YES	YES	YES	YES	YES
INDUSTRY	YES	YES	YES	YES	YES
Observations	1,448	1,448	1,448	1,448	1,448
R-squared					0.252

Note: The robust Z values (T values) are in brackets; ***, ** and * represent significance at the levels of 1%, 5% and 10%, respectively.

Table 6. Regression Results of the Deformed Heckman Two-Step Method

VARIABLES	Probit	OLS
	Native_if	Innovation_ratio
Iv_area	3.0259***	
	(29.2509)	
Iv_inds	2.6690***	
	(6.4097)	
Native_if		-0.0034***
		(-3.1145)
Imr		-0.0006
		(-0.4159)
Constant	-2.4035***	-0.0075
	(-4.0651)	(-0.6995)
CONTROLS	YES	YES
YEAR	YES	YES
INDUSTRY	YES	YES
Observations	5,021	5,021
R-squared		0.391
Pseudo R-squared	0.181	

Note: The robust T values are in brackets; ***, ** and * represent significance at the levels of 1%, 5% and 10%, respectively.

Table 7. Regression Results of the Treatment Effects Model

VARIABLES	Treatment effect model (MLE)		Treatment effect model(twostep)	
	Native_if	Innovation_ratio	Native_if	Innovation_ratio
Iv_area	3.0260***		3.0259***	
	(29.2757)		(28.6228)	
Iv_inds	2.6691***		2.6690***	
	(6.4095)		(6.1748)	
Native_if		-0.0032*		-0.0032
		(-1.7760)		(-1.6061)
athrho		0.0005		
		(0.0134)		
lnsigma		-3.5221***		
		(-174.3909)		
Hazard				0.0000
				(0.0155)
Constant	-2.4035***	-0.0071	-2.4035***	-0.0071
	(-4.0655)	(-0.6388)	(-3.9953)	(-0.6307)
CONTROLS	YES	YES	YES	YES
YEAR	YES	YES	YES	YES
INDUSTRY	YES	YES	YES	YES
Observations	5,021	5,021	5,021	5,021

Note: The robust Z values are in brackets; ***, ** and * represent significance at the levels of 1%, 5% and 10%, respectively.

4.7 Results of the endogenous switching regression model

The endogenous switching regression model was also used to solve the endogeneity of independent variables (self-selection problem), but its advantage over the treatment effects model is that it could find the functional difference for each variable in different groups at the same time while calculating the treatment effects. The treatment effects model, in contrast, has to predict before the calculation. The results of the endogenous switching regression model are shown in Table 8. The results of the selection equation (6) are shown in the first column, which is similar to the conclusion derived in the treatment effects model in the first stage. Whether or not the company employs a local chairman is greatly affected by the regional tendency and the industry tendency. The second column and the third column show the relationship between the control variables and company innovation in the control group and the experimental group respectively, that is, the results from regression equation (7) and equation (8). It can be seen that the speed of company development significantly promoted innovation in firms that had employed chairmen from places other than their places of incorporation. This relationship, however, could not be found in companies that had employed chairmen from their places of incorporation. There was a significant negative correlation between the financial leverage and its innovation investment in companies that were led by chairmen from other places. This correlation remained insignificant in companies controlled by local chairmen. As for the existence of self-selection problem, consistent conclusions have been reached from the endogenous switching regression model and the treatment effects model. The value of the Wald test would not permit the significant rejection of the null hypothesis of "no self-selection problem" with an estimated P value that reached 0.88; The estimated value and robust standard error of σ_T were 0.0589 and 0.5085. The estimated value and robust standard error of σ_U were -0.0092 and 0.0980 respectively, which were not significant. In other words, the test results of the endogenous switching regression model provide clear evidences that negate the existence of endogenous problem of "simultaneous decision-making" in the research about company innovation and the native place of its chairman.

Table 8. Results of the Endogenous Switching Regression Model

VARIABLES	Endogenous switching regression model		
	Native_if	Innovation_ratio(control)	Innovation_ratio(treat)
Iv_area	3.0292***		
	(29.1987)		
Iv_inds	2.6765***		
	(6.3992)		
Da	-0.3919**	-0.0115**	-0.0112**
	(-2.5386)	(-2.1440)	(-2.2663)
Age	-0.0067*	-0.0009***	-0.0008***
	(-1.6642)	(-7.8450)	(-5.2494)
Share	0.0963	-0.0325***	-0.0193***
	(0.6626)	(-6.8181)	(-5.0233)
Multiple_if	-0.0336	-0.0127***	-0.0054**
	(-0.4041)	(-3.3882)	(-2.4718)
Growth	0.0065	0.0006*	0.0003
	(1.1303)	(1.7515)	(1.5818)
Lev	0.0100	-0.0049***	-0.0008
	(0.3549)	(-6.5265)	(-0.8944)
TQ	-0.0603***	0.0044***	0.0052***
	(-2.8245)	(4.1906)	(6.8183)
Capital	-0.0133	0.0080***	0.0074***
	(-0.6111)	(8.6182)	(11.4421)
Multiple_n	0.0919	0.0105***	0.0025
	(1.5865)	(3.4844)	(1.4028)
Size	-0.0478**	0.0003	0.0006
	(-2.1351)	(0.4068)	(0.9399)

Constant	-2.4226***	-0.0137	-0.0183
	(-3.9635)	(-0.6163)	(-1.3658)
YEAR	YES	YES	YES
INDUSTRY	YES	YES	YES
Observations	5,021	5,021	5,021

Note: The robust Z values are in brackets; ***, ** and * represent significance at the levels of 1%, 5% and 10%, respectively.

Next, ATT, ATU and ATE were estimated according to the estimation results of the endogenous switching regression model in equation (11). Compared with the PSM results (ATT: -0.002495329, ATU: -0.00248879 and ATE: -0.002493091) found above, there are obvious differences in specific values. The values estimated by the endogenous switching regression model are all higher, except for ATT. In addition, the conclusions on the significance of the treatment effects are also different. The three treatment effects estimated by the endogenous switching regression model are all significant. In order to demonstrate more clearly the influence on innovation of the decision of whether or not to employ a chairman whose native place is the same as the place of incorporation of the company, the difference of innovation investment between the two groups of companies under different circumstances are shown in Table 9.

Table 9 Treatment Effects

Variable	Obs	Mean	T value
$E(Innovation_{Ti,t} \mid Native\_if_{i,t} = 1)$	3303	0.0394791	-
$E(Innovation_{Ui,t} \mid Native\_if_{i,t} = 1)$	3303	0.0406552	-
ATT	3303	-0.0011762	-9.1011
$E(Innovation_{Ti,t} \mid Native\_if_{i,t} = 0)$	1718	0.0388671	-
$E(Innovation_{Ui,t} \mid Native\_if_{i,t} = 0)$	1718	0.0427703	-
ATU	1718	-0.0039032	-21.1991
ATE	5021	-0.0038769	-36.6482

Fig. 3 contains a graphical representation of the contents of Table 9. The left figure represents the change of innovation investment level of companies in the treatment group after treatment (whether to employ a local chairman or not). Treatture represents the current actual innovation investment level for companies in the treatment group ($E(Innovation_{Ti,t} \mid Native\_if_{i,t} = 1)$), and TreatCF represents the innovation level of companies in the treatment group if they do not employ chairmen whose native places are the same as their places of incorporation ($E(Innovation_{Ui,t} \mid Native\_if_{i,t} = 1)$). It can be seen that after companies in the treatment group refused the treatment, the level of innovation investment of them would be improved. The curve moves to the right, and the difference between the two is ATT. The right figure shows the change of the level of innovation investment in the control group after treatment. Controlture represents the actual innovation investment level of companies in the control group ($E(Innovation_{Ui,t} \mid Native\_if_{i,t} = 0)$), and ControlCF represents the innovation level of companies in the control group after treatment ($E(Innovation_{Ti,t} \mid Native\_if_{i,t} = 0)$). After companies in the control group accepted the treatment, their innovation investment levels would be reduced as the curve moves to the left. The difference between them is ATU.

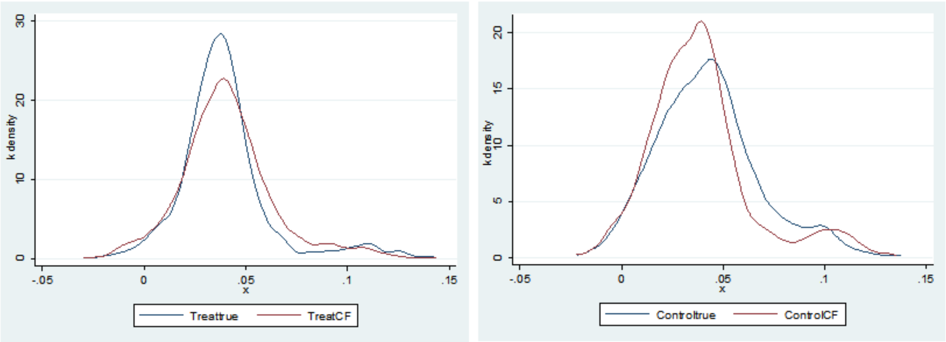

Figure 3. Kernel Density Function Diagram of the Treatment Effects

5 CONCLUSIONS AND SUGGESTIONS

After using OLS for testing and PSM to reduce errors caused by observable variables, it was concluded that when the native place of chairman is the same as the place of incorporation of a company, its innovation capability will be damaged. At the same time, after multiple models for dealing with selection bias were systematically explored, their practical application and results for financial research of listed companies were analyzed. Since the selection bias is comprised of both sample selection bias and self-selection bias, the Heckman two-step method was firstly used to test the samples. The conclusion from this analysis demonstrated the existence of sample selection bias. After the introduction of controls for sample selection bias, the clear conclusion that the chairman whose native place is the same as the place of incorporation of a company would inhibit the company innovation was still valid. In addition, as to the possible self-selection problem, the deformed Heckman two-step method, treatment effects model and the endogenous switching regression model were used to test. All the three models, moreover, proved that there was no self-selection problem in this study. In the end, the treatment effects were estimated by using the PSM method and the endogenous switching regression model. The analyses provided proof that there were both differences in the significance and levels of final estimated values.

This study is supposed to provide a reference point for company governance. The attachment or the sense of belonging to the hometown is a double-edged sword. A local chairman may harm the company innovation after taking office, even though he knows the local business environment quite well and has more social resources. It is important, therefore, to focus on the heterogeneity of the management when choosing candidates for the chairman and other members of upper management.

It is important to be aware that in the financial research of listed companies, the selection bias is composed of the sample selection bias and the self-selection bias. The Heckman two-step method has proved to be effective for solving the sample selection bias. As for the self-selection problem or the endogeneity problem of binary independent variables, the treatment effects model and the endogenous switching regression model can be used as the effective solution. In addition, in some of the existing literature, the deformed Heckman two-step method was used to solve the self-selection problem. According to the manually calculated coefficient with this method, a qualitative judgment consistent with the above two methods can also be obtained. Due to the limitations of manual estimation, however, it is not recommended for using this method as the basis for a quantitative judgment.

References

[1] J. J. Heckman, "Sample selection bias as a specification error," Econometrica: Journal of the econometric society, pp. 153-161, 1979.

[2] J. M. Shaver, "Accounting for endogeneity when assessing strategy performance: Does entry mode choice affect FDI survival?," Management science, vol. 44, pp. 571-585, 1998.

[3] R. H. Dehejia and S. Wahba, "Propensity score-matching methods for nonexperimental causal studies," Rev. Econ. Stat., vol. 84, pp. 151-161, 2002.

[4] G. S. Maddala, Limited-dependent and qualitative variables in econometrics, Cambridge university press, 1983.

[5] G. Maddala, "Methods of estimation for models of markets with bounded price variation," International Economic Review, pp. 361-378, 1983.

[6] T.-S. Hsieh, Z. Wang and S. Demirkan, "Overconfidence and tax avoidance: The role of CEO and CFO interaction," Journal of Accounting and Public Policy, vol. 37, pp. 241-253, 2018.

[7] B. B. Francis, I. Hasan, Q. Wu and M. Yan, "Are female CFOs less tax aggressive? Evidence from tax aggressiveness," The Journal of the American Taxation Association, vol. 36, pp. 171-202, 2014.

[8] S. D. Dyreng, M. Hanlon and E. L. Maydew, "The effects of executives on corporate tax avoidance," The accounting review, vol. 85, pp. 1163-1189, 2010.

[9] M. Giannetti, G. Liao and X. Yu, "The brain gain of corporate boards: Evidence from China," The Journal of Finance, vol. 70, pp. 1629-1682, 2015.

[10] G. A. Akerlof and R. E. Kranton, "Economics and identity," The quarterly journal of economics, vol. 115, pp. 715-753, 2000.

[11] G. A. Akerlof and R. E. Kranton, "Identity and schooling: Some lessons for the economics of education," J. Econ. Lit., vol. 40, pp. 1167-1201, 2002.

[12] O. Boiral, N. Raineri and D. Talbot, "Managers’ citizenship behaviors for the environment: a developmental perspective," Journal of Business Ethics, vol. 149, pp. 395-409, 2018.

[13] J. J. Vaske and K. C. Kobrin, "Place attachment and environmentally responsible behavior," The Journal of Environmental Education, vol. 32, pp. 16-21, 2001.

[14] L. Scannell and R. Gifford, "The relations between natural and civic place attachment and pro-environmental behavior," Journal of environmental psychology, vol. 30, pp. 289-297, 2010.

[15] P. Bardhan and D. Mookherjee, "Decentralizing antipoverty program delivery in developing countries," Journal of public economics, vol. 89, pp. 675-704, 2005.

[16] D. C. Hambrick and P. A. Mason, "Upper echelons: The organization as a reflection of its top managers," Acad. Manage. Rev., vol. 9, pp. 193-206, 1984.

[17] V. V. Acharya and K. V. Subramanian, "Bankruptcy codes and innovation," The Review of Financial Studies, vol. 22, pp. 4949-4988, 2009.

[18] J. R. Brown and B. C. Petersen, "Cash holdings and R&D smoothing," Journal of Corporate Finance, vol. 17, pp. 694-709, 2011.

[19] D. Czarnitzki and K. Hussinger, "The link between R&D subsidies, R&D spending and technological performance," ZEW-Centre for European Economic Research Discussion Paper, pp., 2004.

[20] J. Cornaggia, Y. Mao, X. Tian and B. Wolfe, "Does banking competition affect innovation?," Journal of financial economics, vol. 115, pp. 189-209, 2015.

[21] V. W. Fang, X. Tian and S. Tice, "Does stock liquidity enhance or impede firm innovation?," The journal of Finance, vol. 69, pp. 2085-2125, 2014.

[22] J. J. He and X. Tian, "The dark side of analyst coverage: The case of innovation," Journal of Financial Economics, vol. 109, pp. 856-878, 2013.

[23] B. Balsmeier, L. Fleming and G. Manso, "Independent boards and innovation," Journal of Financial Economics, vol. 123, pp. 536-557, 2017.

[24] A. Abadie and G. W. Imbens, "On the failure of the bootstrap for matching estimators," Econometrica, vol. 76, pp. 1537-1557, 2008.

[25] A. Abadie and G. W. Imbens, "Bias-corrected matching estimators for average treatment effects," Journal of Business & Economic Statistics, vol. 29, pp. 1-11, 2011.

[26] C. S. Lennox, J. R. Francis and Z. Wang, "Selection models in accounting research," The accounting review, vol. 87, pp. 589-616, 2012.

[27] R. Lanis, G. Richardson and G. Taylor, "Board of director gender and corporate tax aggressiveness: An empirical analysis," Journal of Business Ethics, vol. 144, pp. 577-596, 2017.

[28] F. A. Gul, J.-B. Kim and A. A. Qiu, "Ownership concentration, foreign shareholding, audit quality, and stock price synchronicity: Evidence from China," Journal of financial economics, vol. 95, pp. 425-442, 2010.

[29] F. Rios-Avila and G. Canavire-Bacarreza, "Standard-error correction in two-stage optimization models: A quasi–maximum likelihood estimation approach," The Stata Journal, vol. 18, pp. 206-222, 2018.

[30] W. Ma and A. Abdulai, "Does cooperative membership improve household welfare? Evidence from apple farmers in China," Food Policy, vol. 58, pp. 94-102, 2016.