Causal Inference Analysis of the Influence of Family Factors on Student Achievement——Based on CEPS Data

Qiuyu Liao *

Department of Materials Chemistry, Tianjin University of Science and Technology, Tianjin, China

* Corresponding Author Email: 18194266814@163.com

Abstract. The purpose of this paper is to explore the influence of family factors on student achievement by means of causal inference. In this paper, data of the China Education Tracking Survey is used to analyze the factors such as students' family status, parents' education level, family income and family relationship. To obtain the causal effect of family factors and student achievement, a structured causal model was established. Our Research shows that family factors have a positive impact on student achievement.

Keywords: ceps, causality theory, structural causality model.

1. Introduction

As we all know, grades are the most intuitive judgment basis for testing students' learning results, but how to analyses the influencing factors of students' grades are a complex issue. Education includes family education, school education and social education. Family education is both enlightening education and lifelong education, which is the foundation and supplement of school education and its role cannot be replaced [1]. Family factors include family economic status, parental education level, family relationship and family atmosphere, etc. However, the degree of influence of these factors on students' achievement has been controversial. Most of the current analysis of its association is based on traditional statistical methods, but existing statistics-based machine learning methods lack counterfactual level information, and therefore cannot conduct counterfactual level reasoning, so it is not a summary of past experience, or predict the outcome of current actions, there is a lack of evidence. Therefore, the method of causal reasoning [2] is used to analyze whether there is a causal relationship between things.

Causal inference refers to the observation and analysis of the relationship between events to infer the extent of the impact of one event on another event rather than simply judging the correlation between the two based on data. But without a clear semantics for causal claims, and because we do not have effective mathematical tools to pose causal questions or derive causal answers, causal reasoning has rarely been used to solve problems. But now, thanks to advances in graphical models, causality has undergone a major transformation: from a concept shrouded in mystery to a mathematical object with clearly defined semantics and well-grounded logic. Contradictions and disputes are resolved, vague concepts are explained, and practical problems that depend on causal information, long considered metaphysical or unmanageable, can now be solved using elementary mathematics. In short, causality has been mathematized [3]. Therefore, Judea Pearl proposed a structured causal model, through which the causal relationship between factors can be clarified and intervention factors can be removed, and the expected results can be achieved through intervention. Therefore, we can build a structured causal model to analyze the influence of family factors on students' achievement, so as to create better learning environment for students and achieve effective improvement of family education.

Causal theory is increasingly popular because it can help researchers determine the effectiveness and impact of interventions, and as data continues to grow and technology advances, researchers will be better able to collect and analyze data and draw more accurate and meaningful conclusions from it. Because of this characteristic, he is widely used in medicine, social science, economics and other fields.

2. Related Works

There are many methods of causality analysis, and there are many differences between them in the basic logic of causal analysis and the setting form of causal relationship [4]. In simple terms, there are two common methods of multiple regression analysis and structured causal model.

2.1. Multiple regression analysis

Multiple regression analysis is a popular statistical method that establishes a mathematical model to explore the impact of multiple independent variables on a dependent variable. It is widely applied in social science, medicine, and education to determine if the effect of independent variables on dependent variables is statistically significant and practical.

Multiple regression analysis was proposed by Galton and Quetelet in the early 19th century, but its modern form was developed by statisticians like Fisher and Pearson in the early 20th century. Multiple regression analysis is also widely used in China. As early as 1986, Zheng Zhongguang [5] applied multiple regression analysis to the determination of ore weight, and proved that this method has great superiority in practice. In addition, this method has also been widely used to forecast various meteorological parameters. Niu Guiping and Huang Zuying [6] used multiple regression analysis to make long-term forecast of heavy rain.

However, this approach has limitations. For example, the sample size was relatively small because only two classes were selected and the research was conducted in a specific context. Therefore, the generalizability of the findings may be limited. Therefore, this article will not introduce this method in detail, but introduce another method, structured causal model method.

2.2. Structural Causal Model

Structured causal modeling (SCM) is a method of causal inference used to study causal relationships between variables. It is a probabilistic inference method based on causal graphs that aims to infer causal relationships between variables and assess the direction and strength of these relationships [7], and has been widely used in genetics, economics, engineering, and social sciences [3]. One of the advantages of scm over other causal modeling tools is their ability to deal with circular causality [8,9].

SCM assumes causal relationships between variables and represents them as nodes connected by directed edges. Each node has a structural equation that describes its relationship with its parent node. SCM helps researchers identify and control for confounding factors and covariables to accurately assess causality using methods such as instrumental variable and propensity score matching. In practice, SCM requires model selection and parameter estimation, including determining the causal graph structure and structural equation for each node, and determining parameter values. These problems can be solved using common methods in causal inference frameworks, such as maximum likelihood, Bayesian, and instrumental variable estimation.

Structural causality model is the core of Pearl's causality theory and an important achievement in the field of artificial intelligence. Probability graph models, structured equations, counterfactual reasoning and perturbation models, which are data on the one hand and transitive on the other hand, make up for the poor operability of traditional causality theory in the field of new technology causality: It provides solutions to relevant problems in medical, legal, economic and other important fields. Scientific methods for studying causality and mining causality are universal to a certain extent in various disciplines, and causal inference has obtained quite fruitful research results in econometrics, computer science and other fields [11]. The deep learning of continuous iteration of artificial intelligence provides the theory and framework.

In summary, SCM is a method used to study causality, which is able to describe causal relationships between variables through causal graphs and quantify these relationships using structural equations to more accurately evaluate causality.

3. Our methods

The idea adopted in this paper is to use Judea Pearl's structural causal model. We first determine the outcome variables, causal variables and confounding variables, and then construct the structural equation and solve it by using the relevant data of these variables provided by ceps, so as to build an accurate causal diagram and provide a complete framework for causal reasoning. Thus, the causal relationship between the observed variables is intuitively reflected.

3.1. Data and variables

The data in this paper are based on the China Education Panel Survey, (CEPS), The project is a nationally representative large-scale tracking survey project designed and implemented by the China Survey and Data Center of Renmin University of China. (CEPS) Take the 2013-2014 school year as the baseline, adopt the multi-stage probability to Scale (PPS) sampling method, take the first grade of junior middle school and the third grade of junior middle school as the starting point, and take the average education level of the population and the proportion of floating population as stratified variables to randomly select 28 county-level units (counties, districts and cities) as the investigation points. The implementation of the survey was based on schools. 112 schools and 438 classes were randomly selected from the selected county-level units for investigation. All the students in the selected classes were sampled, and a total of about 20,000 students were surveyed for baseline investigation. CEPS uses questionnaire survey as the main means to conduct questionnaire survey on all the students under investigation and their parents or guardians. The contents of the questionnaire include: basic information of students, household registration and mobility, growth experience, parent-child relationship, school performance, basic information of family members, basic information of parents, family education environment, family education investment, family income and other relevant data (Figure 1).

stcog	tr_chn	b09	stfdrunk	stprfigh	stprrel	steco_5c
11	84	4	1	1	2	4
17	82	3	1	1	2	3
12	81	3	1	1	1	3
10	80	3	1	1	2	3
10	84	4	1	1	2	4
12	83	3	1	2	1	3
14	87	3	1	1	2	3
13	89	4	1	1	2	3
10	83	3	1	1	2	3
16	75	3	1	1	2	3
14	76	2	1	2	2	3
8	77	3	1	1	2	3
12	85	3	1	1	2	3
14	72	3	1	1	2	3
11	86	3	1	1	2	4
17	84	3	1	1	1	3
13	74	3	1	1	2	3
13	84	4	1	1	2	4
14	78	3	1	1	1	3
11	69	3	2	2	1	3

Fig 1. The sample of CEPS data

3.1.1 Outcome variable

The Outcome variable of this paper is students' academic performance, which is directly provided by the surveyed schools in the autumn semester of 2013 and not filled in by students themselves. The variables tr_chn/tr_mat/tr_eng represent the original scores of Chinese, math and English respectively. In this paper, the total score of Chinese, mathematics and English tr_sum is used as the result variable to measure students' achievement.

3.1.2 Causal variable

This paper aims to study the influence of family factors on student achievement, so the causal variables include the following

(1) Education level of parents

The education level of students' parents is judged by the questionnaire of students and parents, and the education level of students' father and mother is compared and the relatively higher value is taken. It is divided into 9 grades, 1 no education, primary school, junior high school, secondary school/technical school, vocational high school, high school, college, undergraduate, postgraduate and above.

(2) Financial status of students' families

Students and parents make subjective judgment of family economic status through parent questionnaires and student questionnaires. The data of parents questionnaire is used first. When the question of parents questionnaire is not answered, the data of students questionnaire is used. There are 5 values: very difficult, relatively difficult, medium, relatively rich, and very rich.

(3) Parental relationship

Parental relationship is a subjective judgment made by students on whether the relationship between parents is good, which can be divided into two situations.

3.1.3 Confounding variables

When studying the influence of family factors on student achievement, this paper considers that in addition to the above factors, the following factors may also affect the research results, so the following variables are selected as confounding variables.

(1) Whether the father is drunk

Whether the father is an alcoholic is the subjective judgment made by the middle school students on whether their father is often drunk, which can be divided into two situations.

(2) Parents quarrel

Parents quarrel is a questionnaire in which students give subjective answers to whether their parents often quarrel with each other, which can be divided into two situations.

3.2. The method adopted in this paper

The structural causality model was briefly introduced in the previous article. This study analyzed the possible causal relationship between family factors and student achievement by combining causality diagram and structural equation.

3.2.1 Cause-effect diagrams

The method adopted in this paper is based on Judea Pol's structural causal model. The structural causal model provides a complete framework for causal reasoning, while the causal graph can more intuitively reflect the causal relationship between the observed variables.

Since the 18th century, causality has been extensively discussed in philosophy. Then, based on the counterfactual framework, the causal relationship of modern social science is explored and various causal inference methods applicable to different situations are developed. However, the understanding of causal situations and the interpretation of relevant methods under the existing framework still rely on a large number of algebraic derivations, which are too complicated for scholars and readers of social sciences, and objectively limit the acceptance and influence of causal inference thinking in the discipline. Even systematically trained statisticians and econometricians are not immune to omissions in their presentation. However, it is the appearance of causality diagram that provides a set of intuitive and rigorous expression of causality problem, and directly illustrates causality in a form similar to flow charts. This is the causal graph first proposed by computer scientist and Turing Prize winner Judea Pearl, and developed and refined by the joint efforts of scholars in different fields, these advances are illustrated using the general theory of causality based on the Structural causal Model (SCM) described by Pearl. This model incorporates and unifies other causal approaches and provides a coherent mathematical basis for cause and counterfactual analysis [12].

Among them, the contributions of computer science and epidemiology scholars are the most prominent. (Greenland, Causal Diagrams for Epidemiologic Research, et al., 1999; Hernan and Robins, Causal inference: What if. Boca Raton: Chapman & Hall, 2020 [12-13]; Pearl, 2009). Sociologists Morgan, Winship et al. (Elwert and Winship, 2014; Morgan and Winship, 2014) realized early the important value of causal graphs in understanding causal problems, They use causal graphs (direct acyclic graphs, or DAGs) to highlight that endogenous selection bias stems from conditioning (e.g., controlling, stratifying, or selecting) on a so-called collider variable, i.e., a variable that is itself caused by two other variables, one that is (or is associated with) the treatment and another that is (or is associated with) the outcome. Endogenous selection bias can result from direct conditioning on the outcome variable, a post-outcome variable, a post-treatment variable, and even a pre-treatment variable. We highlight the difference between endogenous selection bias, common-cause confounding, and overcontrol bias and discuss numerous examples from social stratification, cultural sociology, social network analysis, political sociology, social demography, and the sociology of education[14].We believe that the framework of non-parametric causal inference based on causal graph can provide another useful perspective for humanities and social science researchers to examine the problem of causal inference. It helps those who are interested in the problem of causation to delve into the specific situation, clarify the source of bias, clarify the habitual misreading, and cultivate causal thinking. Therefore, this paper will use this method to explore the influence of family factors on students' achievement.

In the definition of discrete mathematics, a causal graph consists of a set of vertices and a set of edges connecting pairs of vertices. Vertices in a graph correspond to variables, and edges represent some relationship between pairs of variables. Two variables connected by an edge are called adjacent variables. Each edge in the graph can be directed (represented by a single arrow on the edge) or undirected (no arrow links). In some applications, "bidirectional" edges are also used to indicate unobserved common causes (sometimes called confounders). Each edge starts at the vertex after the previous edge, and it can be in the same direction as the arrow or in the opposite direction. In other words, a path is any complete, disjoint path drawn along the edges of the graph.

A digraph may contain loops (e.g. X→Y, Y→X), indicating mutual causality or feedback processes, but not self-loops (e.g. X→X). A graph that does not contain a directed ring is called an acyclic graph. A graph that is both directed and acyclic is called a directed acyclic graph (DAG), and most of the graphs discussed in this article on causality belong to this class of graphs.

Causal models are mathematical models that represent causal relationships within a single system or population. It helps to infer causality from statistics, can teach us epistemology about causality, and shows the relationship between causality and probability. It has also been applied to topics of interest to philosophers, such as counterfactual logic, decision theory, and actual causality analysis. A causal graph is a directed graph that shows the causal relationship between variables in a causal model. A causal graph consists of a set of variables (or nodes), each node connected by an arrow to one or more other nodes that have a causal effect on it. Arrows depict the direction of cause and effect, for example, connecting variables A and B with an arrow pointing to B indicates that A change in A causes a change in B with some probability. A path is a traversal of the graph between two nodes along the causal arrow. As an intuitive and rigorous graphical expression system, causal diagram can help researchers clarify some inaccurate viewpoints that have long existed in empirical social science studies [16].

3.3. Causal reasoning analysis based on causal diagram

3.3.1 The construction of causal graph

Based on the above theory and research, the preliminary construction is carried out with common sense and empirical causal diagram. Figure 2 depicts the causal relationship between parental education level (x_1), parents fighting (x_2), parental relationship (x_3), whether the father is often drunk (x_4), family economic situation (x_5), and student achievement (x_6).

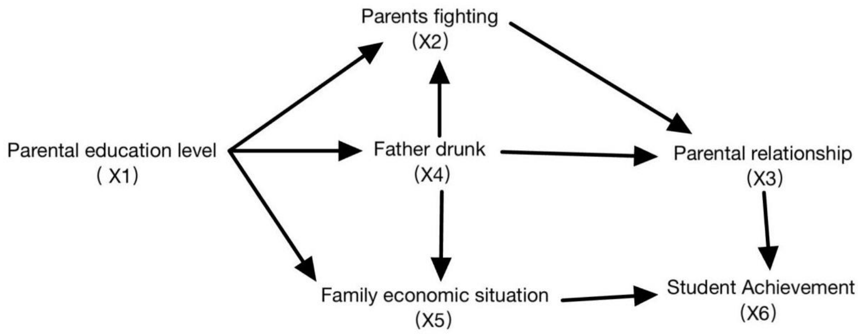

Fig 2. Cause-effect diagram based on empiricism

3.3.2 Construct structural equation

Structural equation modeling is a standard tool in economics and sociology. In a set of structural equations, if each variable has a unique equation, in which that variable appears on the left side of the equation (called the dependent variable), the model is called structural equation modeling or causal model for short.

The corresponding structural equation is constructed from the causal relationship shown in the causal diagram above and consists of six functions, each representing an autonomous mechanism that controls a variable. Where x(i.e., the parent variable) represents the set of variables with a directly determined value, and u represents the error resulting from the omitted factor. Its structural equation is as follows:

$$\begin{gathered} x_1 = u_1 \\ x_2 = f_1(x_1, x_4, u_2) \\ x_3 = f_2(x_2, x_4, u_3) \\ x_4 = f_3(x_1, u_4) \\ x_5 = f_4(x_1, x_4, u_5) \\ x_6 = f_5(x_3, x_5, u_6) \end{gathered} \tag{1}$$

3.3.3 Solution of the structural equation

(1) The result

According to the relevant variables and data in CEPS, the structural equation is solved, and the final causal diagram is shown in Figure 3. The figure is based on a large number of CEPS data and has high confidence. As can be seen from the figure, there is a causal relationship between students' family factors and students' grades, which specifically shows that there is a direct causal relationship between family economic conditions and students' grades. Parents' education level has a direct causal relationship with the three variables of whether the father is drunk, parents quarrel and family economic conditions, while the two variables of whether the father is drunk and parents quarrel have a direct causal relationship with parents.

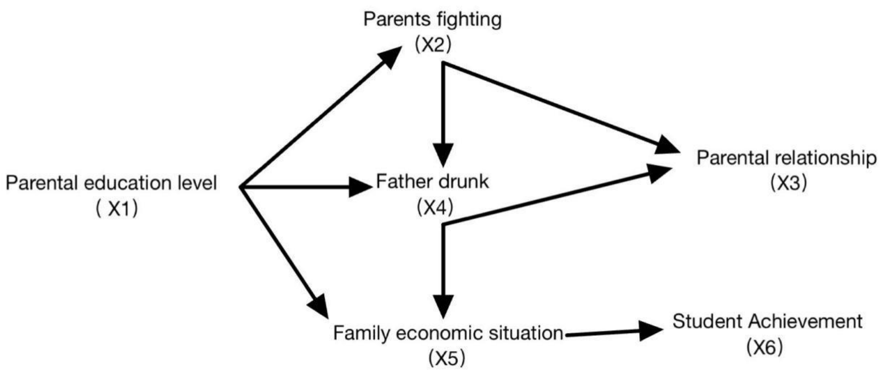

Fig 3. Cause-effect diagram based on CEPS data

4. Conclusion

This paper analyzes the possible causal relationship between family factors and students' achievement by using the causal reasoning method combined with causal diagram and structural equation. The results show that family factors have a significant impact on student achievement. Family economic level is the most important factor affecting student achievement, family education background and good relationship between parents also have a certain degree of influence on student achievement. Therefore, improving family education input, improving family background and maintaining good relationship between parents, as well as parents not drinking too much are all effective ways to promote the improvement of student achievement.

In addition, this conclusion has important reference value for the formulation of education policy and the practice of family education. In the formulation of educational policies, we should pay attention to the influence of family factors on students' achievements, pay attention to improving the level of family education, strengthen the support and guidance of family education, and provide students with a better family education environment. In the practice of family education, parents should pay attention to cultivate a good family atmosphere, improve their own education level, correct improper behavior, and provide better learning conditions and support for children.

Overall, the paper provides insights into the impact of family factors on student achievement and demonstrates the excellence of causal reasoning methods in examining causal relationships between variables.

References

[1] Du Xingjiao. Analysis of Influencing Factors on Middle School Students' Academic Performances in Math, Chinese and English Based on the CEPS Data[D].Tianjin University of Finance and Economics,2019.DOI:10.27354/d.cnki.gtcjy.2019.000951.

[2] Imbens G W, Rubin D B. Causal inference in statistics, social, and biomedical sciences[M]. Cambridge University Press, 2015.

[3] Pearl J. Causality[M]. Cambridge university press, 2009.

[4] Hao J. Comparison and Application of causal analysis methods in Social science research [J]. Statistics and Decision,2009,No.300(24):34-36.

[5] Zheng Zhongguang. Application of multiple regression analysis in determination of ore weight [J]. Geology and Exploration,1986(08):44-46.

[6] Niu Guiping, Huang Zuying. Long-term forecast of heavy rain using multiple regression analysis [J]. Shaanxi Meteorology,1996(01):8-9.

[7] Zhang Xin. Judea pearl causality theory research [D]. Xiangtan university, 2021.2021.DOI:10.27426/d.cnki.gxtdu.2021.001438.

[8] Hyttinen A, Eberhardt F, Hoyer P O. Learning linear cyclic causal models with latent variables[J]. The Journal of Machine Learning Research, 2012, 13(1): 3387-3439.

[9] Forré P, Mooij J M. Markov properties for graphical models with cycles and latent variables[J]. arXiv preprint arXiv:1710.08775, 2017.

[10] Ma Zhonggui, Xu Xiaohan, Liu Xuer. Causal inference three analysis framework and its application in [J]. Journal of engineering science, 2022, 44 (7) : 1231-1243. The DOI: 10.13374 / j.i ssn2095-9389.2021.07.04.002.

[11] Pearl J. Causal inference in statistics: An overview[J]. 2009.

[12] Greenland,Sander,Judea Pearl,and James M. Robins. 1999.“Causal Diagrams for Epidemiologic Research.”Epidemiology 10(1):37-48.

[13] Hernán,Miguel and James M. Robins. 2020. Causal Inference:What If. Boca Raton: Chapman & Hall.

[14] Elwert,Felix and Christopher Winship. 2014. “Endogenous Selection Bias:The Problem of Conditioning on a Collider Variable.”Annual Review of Sociology 40:31-53.

[15] Yu Guodong, Chen Yunsong. Graphic logic power: the concept of cause and effect diagram and its application [J]. Journal of society, and 2022 (03) : 195-221. The DOI: 10.15992 / j.carol carroll nki. 31-1123 / c. 2022.03.005.