



OPEN ACCESS

EDITED BY

Xiaodong Luo,
Norwegian Research Institute (NORCE),
Norway

REVIEWED BY

Yuhao Deng,
University of Michigan, United States
Luis Castro Martin,
University of Granada, Spain

*CORRESPONDENCE

Yuejuan Zhao
✉ 20210980144@fudan.edu.cn

RECEIVED 08 February 2024

ACCEPTED 20 March 2024

PUBLISHED 15 April 2024

CITATION

Zhao Y (2024) The application of propensity score methods in observational studies. *Front. Appl. Math. Stat.* 10:1384217. doi: 10.3389/fams.2024.1384217

COPYRIGHT

© 2024 Zhao. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

The application of propensity score methods in observational studies

Yuejuan Zhao*

School of Data Science, Fudan University, Shanghai, China

Introduction: In research, it is crucial to accurately estimate treatment effects and analyze experimental results. Common methods include comparing outcome differences between different groups and using linear regression models for analysis. However, observational studies may have significantly different distributions of confounding variables between control and treatment groups, leading to errors in estimating treatment effects.

Methods: The propensity score methods can address this issue by weighting or matching samples to approximate the scenario of a randomized experiment and allow for more accurate estimation of treatment this paper.

Results: We use propensity score methods to analyze three datasets from observational studies and draw conclusions different from those in the original text. Furthermore, we simulate three scenarios, and the results demonstrate the superiority of propensity score methods over methods such as linear regression in addressing selection bias.

Discussion: Therefore, it is essential to thoroughly consider the characteristics of the data and select appropriate methods to ensure reliable conclusions in practical data analysis.

KEYWORDS

propensity score, observational study, confounding variable, treatment effect, linear regression

1 Introduction

In observational studies, such as in social public affairs, the formulation or assertion of various decisions requires robust evidence to support them. Analyzing the effects of decisions involves examining the relationships between variables, commonly using methods such as linear regression and logistic regression. However, real-world data often exhibit complex relationships, with variables potentially showing quadratic or even more intricate correlations. Additionally, due to various reasons, measurement errors might exist in the data, and linear models are heavily influenced by outliers. In such cases, linear models may not accurately describe the relationships between variables and may struggle to effectively address selection bias, thus impacting the accuracy of causal inference. The adoption of an inappropriate model could lead to decision errors and significant losses. Therefore, it is necessary to explore different methods, such as Propensity Score (PS) based methods—Propensity Score Matching (PSM), Inverse Probability Weighting (IPW) and Augmented IPW (AIPW), and conduct comprehensive comparative analyses to ensure reliable results. In this article, we apply PS methods to analyze datasets from three different fields: scientific research, natural biology, and demography. We discuss the advantages of PS methods over methods such as linear regression and introduce interaction terms in the linear regression model to provide partial corroboration for the results obtained from the analysis using PS methods.

In the study on the influence of research papers, Uzzi et al. [1] conclude that papers exhibiting both high conventionality and novelty are more likely to become hit papers. We replicate their experiment on the DBLP dataset and grouped the papers based on their novelty and conventionality to calculate the probability of becoming hit papers. Papers scoring high in both novelty and conventionality showed a 2.9 times higher probability of becoming hit papers compared to those with low scores in both aspects. However, when incorporating other confounding variables and using PS methods to calculate the probability ratio, the results were all smaller compared to the original conclusion. So relying solely on the ratio of probabilities may lead to a biased estimation of the quality of the papers. Similarly, when reviewing funding proposals, experts who assess projects solely based on their novelty and conventionality may make erroneous judgments and this will lead to the misallocation of resources toward relatively low-output and low-efficiency directions.

The study conducted by Vollaard and van Ours [2] on the fairness of herring ratings reveals that in the regression analysis of herring scores concerning whether they are supplied by Atlantic, the coefficient for Atlantic is significant at the 5% confidence level. Therefore, they concluded there was a bias in favor of Atlantic in the ratings. Vollaard's findings have had a significant impact on the reputation of Atlantic, leading to substantial economic losses and raising public concerns regarding the fairness of the ratings. However, according to all three PS analyses, after controlling for other variables related to herring, it can not be concluded that herring supplied by Atlantic receives significantly higher scores compared to others. Thus, we can not assert bias in the ratings. Through a more appropriate linear regression analysis, the coefficient for Atlantic is also found to be non-significant, providing further confirmation of this result.

In the study on how education influences health, Cheng et al. [3] verified that education can have a positive impact on the physical health of the elderly through influencing health behaviors and socioeconomic factors. In their analysis, the former has a greater effect. However, according to other three PS analyses, when only controlling for leisure index, one variable in the health behavior category, the coefficient for education is no longer significant, indicating that we cannot definitively claim that education has a significant impact on health. Furthermore, when controlling for variables related to diet, the coefficient for education decreases compared to the baseline model, suggesting that education may not have an impact on health through this specific channel. Therefore, in health education, it may be more effective to focus on leisure activities rather than promoting healthy eating, or increasing financial subsidies for the elderly.

In this paper, we employ PS methods to estimate the treatment effects and provide supplementary analysis to the original conclusions. PS methods achieve an approximation of a randomized experiment by weighting or matching samples based on their propensity scores. The paper is structured as follows, in Section 2 we introduce the PS methods, while in Sections 3–5 we present the analyses of the DBLP dataset, the herring dataset, and the CLHLS dataset, respectively. In Section 6 we conduct simulations to compare the estimation of treatment effects using

various methods under different scenarios. In the final section we summarize the conclusion.

2 PS methods

Let Z represents different experimental exposure, where $Z = 1$ if sample is assigned to treatment group and $Z = 0$ otherwise. The outcomes when samples are assigned to control and treatment groups are denoted as $(Y(0), Y(1))$. But for a specified individual only one of $(Y(0), Y(1))$ can be observed; hence, they are usually referred to as potential outcomes. In observational studies, there is often an issue of selection bias— $(Y(0), Y(1))$ are not independent of the group assignment variable Z . In this situation, the distribution of confounding variables differs between the two groups, and these confounding variables may also influence potential outcomes. Consequently, the difference between the average observed responses of the two groups contains both the effect of the treatment and confounding variables. Comparing the difference in average observed responses directly to estimate the treatment effect may lead to errors.

For example, if we want to calculate the average treatment effect $ATE = E(Y(1) - Y(0))$. This represents the difference in the expected outcomes of the sample assigned to treatment and control group. The average response of the treatment group, i.e., $E(Y(1)|Z = 1)$, is not equal to $E(Y(1))$ due to confounding variables. In order to estimate ATE more accurately, various methods such as linear regression, PSM, IPW and AIPW can be used for analysis.

Let X denote confounding variables. When calculating treatment effects, VanderWeele [4] suggest selecting all variables that occur before the experiment. If the conditional independence assumption holds, i.e.,

$$(Y(0), Y(1)) \perp\!\!\!\perp Z|X,$$

then we can conclude that $E(Y(0)|Z = 0, X) = E(Y(0)|Z = 1, X)$. Hence we can select samples with the same values of the confounding variable and use the average outcomes of the control group as a substitute for $E(Y(0)|Z = 1, X)$. However, as the dimension of the confounding variables increases, it becomes infeasible to select samples with the same values of the confounding variable. The PS methods [5–7] can address this issue. Propensity score refers to the probability of a sample receiving treatment given the confounding variables, denoted as $e(X) = p(Z = 1|X) < 0 < e(X) < 1$. Under the assumption of conditional independence, it can be proven that $(Y(0), Y(1)) \perp\!\!\!\perp Z|e(X)$ holds, which means that we can describe the sample's propensity for the treatment solely based on the propensity score without considering all the confounding variables.

Intuitively, if two samples have similar propensity scores but receive different experimental treatments, it can be assumed that the potential outcomes for these samples are identified. This provides the possibility to estimate the treatment effect. To better utilize the samples, they can be divided into K strata Q_1, \dots, Q_k , and

the PSM estimator is

$$\Delta_{psm} = \sum_{j=1}^K \frac{n_j}{n} \{n_j^{-1} \sum_{i=1}^n Z_i Y_i \mathbf{1}(e(X_i) \in Q_j) - n_{0j}^{-1} \sum_{i=1}^n (1 - Z_i) Y_i \mathbf{1}(e(X_i) \in Q_j)\},$$

where n_{kj} is the number of samples belonging to both group $Z = k$ and strata Q_j , moreover, $n_j = \sum_k n_{kj}$ and $n = \sum_j n_j$.

In applications, the number of strata in PSM method needs to be finely adjusted. Insufficient samples in certain strata can limit the effectiveness of matching. The idea behind IPW is to assign weights to the samples based on their propensity scores, aiming to make the distribution of confounding variables more similar between two groups [8]. This decouples the influence of treatment and confounding variables on the outcomes, so the difference in outcomes is only due to different experimental exposures, making it analogous to the scenario of randomized experiment. Taking the calculation of ATE as an example. When estimating $E(Y(0))$, we assign weights $w(X)$ to samples in control group, ensuring that the distribution of confounding variables between the two groups is similar, i.e.

$$w(X) \propto \frac{f(X)}{f(Z = 0|X)} = \frac{1}{1 - e(X)},$$

It means that if the propensity score of sample in control group is higher, the probability of it receiving treatment is also higher, and the confounding variable distribution of it is more similar to the treatment group. By increasing the weight of samples in the control group with higher propensity scores and decreasing the weight of samples with lower propensity scores, the distributions of confounding variables for the two groups of samples can be made more similar. The estimated value of $E(Y(0))$ is $\hat{E}(Y_0) = \frac{1}{n} \sum_{i=1}^n (1 - Z_i) Y_i / (1 - e(X_i))$, where C means control group and $EY(1)$ can be obtained in a similar manner. Finally, the IPW estimator of ATE is

$$\Delta_{ipw} = \frac{1}{n} \sum_{i=1}^n \frac{Z_i Y_i}{e(X_i)} - \frac{1}{n} \sum_{i=1}^n \frac{(1 - Z_i) Y_i}{1 - e(X_i)}$$

The IPW estimator highly depends on the accuracy of the propensity score. To address this issue, Robins et al. [9] proposed AIPW method, which includes the outcome model $m_k(X) = E(Y|Z = k, X)$. When at least one of the outcome model and the propensity score model is accurate, the estimation of the treatment effect is accurate. Taking ATE as an example, the AIPW estimator is

$$\Delta_{aipw} = \frac{1}{n} \sum_{i=1}^n \left(\frac{Z_i Y_i}{e(X_i)} - \frac{Z_i - e(X_i)}{e(X_i)} m_1(X_i) \right) - \frac{1}{n} \sum_{i=1}^n \left(\frac{(1 - Z_i) Y_i}{1 - e(X_i)} + \frac{Z_i - e(X_i)}{1 - e(X_i)} m_0(X_i) \right).$$

Furthermore, the asymptotic variance of the AIPW estimator is smaller than that of the IPW estimator [9].

PS method uses tree-based models to estimate the propensity score. It splits the data into different subtrees by selecting a splitting

point each time. The complexity of the model can be controlled by limiting the tree's depth and the number of samples in each leaf node. This method bypasses the linear assumption and uses non-parametric functions for classification. Moreover, tree-based models automatically select variables and are less sensitive to outliers.

However, when using tree models for analysis, we typically can only use the Bootstrap method to compute the variance or standard error (SE) of the estimates. In contrast, the estimation of variance in linear regression is relatively straightforward, but the assumption of linear relationship does not always hold true. Introducing a large number of cross-term features may cause overfitting issues, particularly in situations with limited data.

The propensity score method divides the estimation of treatment effects into two steps: the first step involves using the estimated propensity scores for weighting or matching, and the second step entails the estimation of effects based on the simulated random experiment scenarios. In practical applications, we assume that the estimation of the propensity scores is accurate and evaluate the treatment effects using t-test based solely on the results of the first step, thereby avoiding the time-consuming Bootstrap method.

PS methods are widely applied in various fields such as healthcare and social sciences. Li et al. [10] utilized IPW method to adjust the weights of the samples, making the distribution of women's confounding variables—age, BMI, and others similar between the treatment and control groups. They conclude that there is no significant reduction in the probability of developing pregnancy-induced hypertension by taking a certain nutrient during early pregnancy. Thomas et al. [11] employed IPW method to analyze the relationship between food security and children health. They find that children from households with food security issues are more likely to delay treatment due to financial constraints. These children are also more likely to have difficulty affording healthcare, dental care, and mental health services. Additionally, they are more prone to experiencing chronic symptoms such as asthma, skin allergies, and depression. Besides, while their acute illness symptoms are similar to those of other children, they are more susceptible to colds and gastrointestinal problems in the past two weeks. PS methods have also been widely applied in other studies [12, 13].

3 DBLP dataset

Uzzi et al. [1] conducted an in-depth study on the importance of novelty and conventionality in determining the impact of scientific papers. Their research revealed a significant phenomenon in scientific research and provided insights and profound understanding of scientific development. They analyzed the Web of Science (WOS) dataset and measured the novelty and conventionality of paper citations by examining the citation counts in reference combinations. The findings show that papers with high level of conventionality and novelty has a probability of 9.11% of becoming hit papers, while papers with low levels of both have a probability of 2.05% of becoming hit papers. Thus, they conclude that papers with high novelty and conventionality are more likely to become hit papers.

Uzzi et al. combined referenced papers in pairs and recorded their citations counts received in the year of publication. For computational feasibility, they used journal combinations to replace paper combinations. They then standardized these citation counts based on the observed citation counts of the same combinations in randomized citation networks. To create the randomized citation networks, they randomly selected two papers and two references published in the same year respectively, exchanged their citation relationships, and repeated this process multiple times. They repeated this procedure to generate 10 random citation networks and recorded the citation counts for each journal combination. The normalized citation count for each journal combination was calculated as the z -score, which describes the rarity or commonness of the frequency with which the references are cited together compared to the random networks. For a given paper, the z -score values of all journal combinations formed by its references were computed. The high novelty of a paper was defined as having 10th-percentile below zero, while the high conventionality of a paper was defined as having a median z -score value higher than the median z -score value of all journal combinations.

Because WOS dataset is not publicly available, we conduct our analysis on the DBLP-Citation-network V12 [14] dataset (DBLP), which yield similar results to the original study. We select papers published between 1965 and 2015 and define hit paper as one ranking in the top 5% of citations within a 5-year period. The DBLP dataset contains 4.89 million papers and 45.56 million citation records. Using the aforementioned method, we calculate the conventional and novel characteristics for each paper. The results show that the probability of papers with both high conventional and novel characteristics becoming hit papers is 8.6%, while the probability for papers with low scores in both characteristics is 2.9%. Papers with high scores in both characteristics are 2.94 times more likely to become hit papers compared to papers with low scores in both characteristics, which is roughly consistent with the results from the WOS dataset.

Assuming $Z = 1$ indicates both high novelty and conventionality of the paper, and $Z = 0$ indicates both are low, then the assessment of the impact of novelty and conventionality on the paper requires the estimation of

$$RR = \frac{EY(1)}{EY(0)}.$$

In the analysis by Uzzi et al., they did not consider other confounding variables that could influence citation counts, such as the number of citations the author has received in the past which represents the author's academic impact. The distribution of confounding variables is different between the two groups of papers, so the estimation of the probability ratio of becoming a hit paper $E(Y(1)|Z = 1)/E(Y(0)|Z = 0)$ includes the effects of confounding variables, novelty and conventionality, and this may lead to biased estimates.

We rigorously select variables occurring before the paper under analysis, including the number of hit papers published by the corresponding author in the past 5–10 years ($hit_cnt_corr_5yrs$), the number of papers published by the first author in the past 3 and 5 years ($pub_cnt_fir_3yrs, pub_cnt_fir_5yrs$), the citation counts

of corresponding author in the past 5 years ($cite_cnt_cor_5yrs$), and the citation counts of the first author in the past 3 years ($cite_cnt_fir_3yrs$) to estimate the propensity score. By sequentially adding these variables for analysis, the results, as shown in Table 1, indicate that after including the first five variables, the probability ratios of becoming hit paper are all much lower than 2.94.

If we include more variables, such as the ranking of the authors' respective schools, the average citations of the first author over the past 5 years, the number of citations within 3 years of publication for papers published by the first author in the last 5–10 years, publication date, and so on, using all 13 variables to estimate the propensity score, the results indicate that this ratio is revised to 1.8–1.9.

4 Herring dataset

Every year, the Algemeen Dagblad newspaper organized a competition focusing on the quality of herring. Vollaard and van Ours [2] found that Atlantic-supplied herring received exceptionally high ratings, with an average score of 8.2. In contrast, the herring supplied by other outlets had an average score of 5.5. Furthermore, despite Atlantic's market share being <10%, their herring occupied more than half of the top ten positions in 2016 and 2017. To examine the fairness of these ratings, a regression analysis was conducted, considering factors such as cleanliness, ripeness, freshness, verbal judgment and "Atlantic" (indicating whether herring is supplied by Atlantic). The results revealed that the coefficient for the variable indicating whether the herring was supplied by Atlantic was significant at the 5% level. Considering that one of the critics responsible for the ratings had a vested interest in Atlantic, the authors concluded that the ratings exhibited a bias toward Atlantic, suggesting an unfairness in the evaluation process.

The average score of herring supplied by Atlantic is higher than that of other outlets. There may be two possible reasons for this situation. On one hand, there might be bias in the scoring process favoring Atlantic, which means even if characteristics of the herring are similar, the herring supplied by Atlantic still receives higher score. Therefore whether supplied by Atlantic directly affects the final score. On the other hand, the quality of herring supplied by Atlantic is higher, and this will have an impact on final score and whether they are supplied by Atlantic. The analysis of the characteristics of the two groups of herring is shown in Table 2, indicating a significant difference in the distribution of herring characteristics between Atlantic and other companies. Therefore, the score difference may arise from confounding variables such as maturity, cleanliness, freshness, and linguistic evaluations by critics. Merely relying on the difference in average scores is insufficient to confirm bias toward Atlantic in the scoring process.

By utilizing linear regression to analyze the relationship between herring scores and the supply from Atlantic, we can partially address the issue of selection bias. However, the model may suffer from endogeneity issues, such as the quality of herring not being fully captured by the features, which is a real concern. The temperature feature of the herring is omitted in the regression analysis, and the verbal comments about temperature are also ignored. Additionally, there could be a non-linear relationship

TABLE 1 Result of DBLP dataset.

Variables added	IPW-RR	IPW-SE	PSM-RR	PSM-SE	AIPW-RR	AIPW-SE
<i>hit_cnt_corr_5yrs,</i> <i>pub_cnt_fir_3yrs</i>	2.203	0.086	2.523	0.057	2.447	0.055
<i>pub_cnt_fir_5yrs</i>	2.203	0.086	2.510	0.057	2.435	0.054
<i>cite_cnt_cor_5yrs</i>	1.941	0.077	2.191	0.050	2.140	0.048
<i>cite_cnt_fir_3yrs</i>	1.901	0.075	2.102	0.048	2.063	0.046
All 13 variables	1.804	0.074	1.952	0.047	1.870	0.042

TABLE 2 Comparison of covariates for herring.

Confounding variables	χ^2 statistic	<i>p</i> -value
Verbal judgment	8.775	0.000
Degree of ripening	3.694	0.012
Quality of cleaning	4.353	0.005
Freshly prepared	4.196	0.041

between the herring features and final score and whether the supply from Atlantic is correlated with the residual term in linear regression. All of these could lead to significant coefficient test results.

Indeed, PS methods cannot fully address endogeneity issues. But potential correlations often exist between confounding variables, such as the temperature and freshness of herring, where higher temperatures typically lead to quicker spoilage and reduced freshness. These correlations provide an opportunity to approximate unobserved variables. Linear models are limited to capturing linear relationships. Tree-based models can more accurately calculate proxies for unobserved variables, therefore they can partially mitigating endogeneity issues. Christos Louizos et al. [15] employed variational autoencoders to approximate the joint distribution of observed and unobserved variables and got a more precise estimation.

To test the hypothesis of non-linearity, we attempt to include interaction terms of variables in the linear model. Due to the fact that all variables are categorical factors in linear regression, the number of variables becomes significantly large compared to the limited sample size. In order to reduce the number of variables, we first select splitting points for the variables and then discretize them into factor types. For instance, we choose 1 and 2 as splitting points for the linguistic evaluations of critics, resulting in three variables: (0, 1][1, 2] and (2, 5]. Similarly, we select 1 and 3 as splitting points for cleanliness and maturity variables. After including quadratic and partial cubic interaction terms of these variables in the linear regression, the coefficient for Atlantic is no longer significant. Therefore, we cannot conclude that there is bias in favor of the Atlantic in the scoring solely based on regression analysis.

After adding interaction terms of variables in the linear regression models of Volland et al., the coefficient test results of Atlantic changed. This may indicates a non-linear relationship between the score of herring and its features. However, continuously adding higher-order interaction terms of variables

TABLE 3 Result of herring dataset.

	ATE	Std. Err	<i>p</i> -value
Linear model	0.251	0.109	0.023
PSW	0.542	0.891	0.543
PSM	0.781	0.639	0.223
AIPW	0.319	0.247	0.198

to approximate the regression function in the linear regression model may lead to overfitting issues. In the subsequent analysis, we attempt to use PS methods to analyze the herring dataset. As shown in Table 3, the *p*-values of the coefficients of Atlantic are all not significant which indicates that we cannot conclude a significant difference in scores between herring provided by Atlantic and those provided by other companies if their quality is similar.

5 CLHLS dataset

Cheng et al. [3] analyzed the Chinese Longitudinal Healthy Longevity Survey (CLHLS) dataset from 2008 to examine how education affects people's health. They used Instrumental Activities of Daily Living (IADL) as a measure of health status among the elderly. They considered baseline characteristics such as education status, gender, age, ethnicity, and parental longevity etc. They progressively introduced health behavior variables and socioeconomic variables into the analysis. The health behavior variables included smoking, alcoholism, diet, exercise, and leisure index. The socioeconomic variables included financial independence, source of livelihood, control over household expenses, health insurance, and medical care. Incorporating these variables led to an increase in the coefficient for the education variable, particularly when including diet, exercise, and leisure index. In all regression analyses, the coefficient for education was found to be statistically significant. This indicated that education could influence physical health by affecting health behaviors and economic capabilities. Among these factors, the effects resulting from influencing diet, exercise, and the leisure index were found to be the most significant. Furthermore, even after controlling for these variables, education continued to have an impact on health.

We recalculate the regression results, as shown in Table 4. The coefficient for education is consistent with the original findings, indicating that our variable selection and handling align closely with the original study. By utilizing the PS methods on this

TABLE 4 Result of CLHLS dataset.

Variable added	Linear model			PSW			PSM			AIPW		
	ATE	SE	<i>p</i> -value	ATE	SE	<i>p</i> -value	ATE	SE	<i>p</i> -value	ATE	SE	<i>p</i> -value
Base variables	-0.052	0.0090	7.06E-09	-0.052	0.0198	0.0084	-0.037	0.0125	0.0027	-0.038	0.0100	0.0002
Alcoholism	-0.051	0.0090	1.06E-08	-0.057	0.0185	0.0021	-0.048333	0.0147	0.0010	-0.038	0.0102	0.0002
Smoking	-0.052	0.0089	5.72E-09	-0.058	0.0182	0.0013	-0.046376	0.0190	0.0148	-0.037	0.0100	0.0002
Diet	-0.049	0.0090	4.90E-08	-0.056	0.0199	0.0047	-0.041097	0.0161	0.0107	-0.038	0.0100	0.0002
Exercise	-0.047	0.0089	1.63E-07	-0.055	0.0180	0.0024	-0.040797	0.0144	0.0046	-0.033	0.0102	0.0011
Leisure index	-0.030	0.0088	0.00062	-0.028	0.0204	0.1650	-0.0057945	0.0112	0.6034	-0.013	0.0094	0.1695
Health behavior	-0.029	0.0088	0.00106	-0.027	0.0203	0.1820	-0.0010988	0.0114	0.9233	-0.011	0.0095	0.2348
Financial independence	-0.040	0.0089	7.52E-06	-0.049	0.0182	0.0069	-0.023408	0.0138	0.0907	-0.031	0.0100	0.0016
Source of livelihood	-0.047	0.0089	1.39E-07	-0.046	0.0197	0.0197	-0.038054	0.0135	0.0048	-0.035	0.0100	0.0005
Control over household expenses	-0.046	0.0088	2.09E-07	-0.049	0.0203	0.0155	-0.028334	0.0117	0.0154	-0.033	0.0097	0.0006
Health insurance	-0.052	0.0088	8.18E-09	-0.054	0.0196	0.0059	-0.042124	0.0144	0.0035	-0.039	0.0104	0.0002
Medical care	-0.052	0.0088	8.66E-09	-0.052	0.0199	0.0096	-0.037951	0.0124	0.0023	-0.036	0.0097	0.0002
Socioeconomic variables	-0.035	0.0089	8.59E-05	-0.037	0.0202	0.0659	-0.025	0.0132	0.0611	-0.028	0.0102	0.0059
All variables	-0.020	0.0088	0.02371	-0.024	0.0200	0.2260	-0.002	0.0130	0.8720	-0.008	0.0095	0.4166

TABLE 5 Result of simulation.

		Scenario 1	Scenario 2	Scenario 3
AIPW	ATE	0.0026	-0.0058	0.0025
	SE	0.0999	0.0027	0.0308
	$p > 0.1$	93	81	84
	$p > 0.05$	98	91	88
IPW	ATE	-0.0048	0.1356	0.0480
	SE	0.1210	0.0216	0.0232
	$p > 0.1$	97	29	53
	$p > 0.05$	99	60	62
PSM	ATE	0.0496	-0.0179	0.0021
	SE	0.0529	0.0220	0.0303
	$p > 0.1$	100	91	86
	$p > 0.05$	100	96	90
LR or CP	AME	0.1059	-0.0138	0.1979
	SE	0.0327	0.0041	0.0206
	$p > 0.1$	8	7	0
	$p > 0.05$	14	12	0

1. LR, linear regression; CP denotes conditional probability, in scenario 1 and scenario 2, we compare linear regression, PSM, AIPW and IPW methods. In condition 3 we compare conditional probability, PSM, AIPW and IPW methods.

2. In linear regression, Average of Margin Effect(AME) is the coefficient of Z, in condition 3, AME is the differenc in the probability of $Y = 1$ between the two groups.

3. In simulation, SE calculated as $\sqrt{\sum(\Delta_i - E(\Delta))^2}$ is different from the previous section, it contains the variance of propensity score estimating. Among them, Δ is the real value of ATE which is 0 in these scenarios and Δ_i is the i -th estimate.

dataset to analyze the impact pathway, we observe that the impact of education on physical health is no longer significant when including the leisure index as a confounding variable while controlling for baseline characteristics. This suggests that educated elderly individuals tend to engage in a more enriching leisure lifestyle, which contributes to better health outcomes. However, when incorporating other health behavior variables, contrary to the linear regression model used in the original study, the coefficient for education decreases. Therefore, we can not conclude that education promotes physical health by influencing other health behavior variables such as balanced diet.

Besides, socioeconomic has a certain influence on health status. After including all socioeconomic variables, the impact of education on health is higher than that of including all health behavior variables. This indicates that healthy lifestyle has a greater impact on physical health than economic capabilities, consistent with the findings of the original study.

After including both health and socioeconomic variables, the impacts of education on health all remain insignificant. It means controlling for these two categories of variables, we can not conclude that education still has an effect on health. In other words, when controlling for baseline characteristics, if an elderly individual does not receive education but possesses a healthy lifestyle and favorable economic conditions, their likelihood of being physically healthy would still be substantial. So we can not establish a relationship between education and health in such cases.

In the original linear regression model, after controlling for socioeconomic and health behavior variables, the impact of

education on health remained significant. This could be due to the presence of nonlinear relationships between some variables and health, which can not be captured by linear regression. Additionally, there might be multi-collinearity between education and other variables. To test this hypothesis, we introduce some nonlinear elements into the original regression model (including interaction terms of all variables with education, as well as squared terms for non-binary variables). We find that the coefficient for education was no longer significant. In the new regression model, certain interaction terms between health behavior and socioeconomic variables and baseline characteristics are found to have an impact on health. For example, the p -value for the interaction term between age and alcohol consumption is 0.0035, indicating that older individuals who consume alcohol tend to have poorer health conditions. The p -value for the interaction term between gender and control over household expenses was 0.0051, suggesting that if males have control over household expenses, their health status tends to be worse. However, in the original regression model the coefficient for gender is not significant which indicates that gender does not have a significant impact on health. This may suggests the presence of nonlinear relationships between health and other variables, and using a linear function to approximate is inappropriate. With the inclusion of interaction terms, after controlling for health behavior variables and socioeconomic variables, we can not establish a correlation between education and health. This finding aligns with the results obtained from the PSW method.

6 Simulation

Cepeda et al. [16] pointed out that in situations with a large number of features and a small number of samples, using propensity score method for estimation has advantages over logistic regression in terms of bias, empirical power, and robustness. We analyze three scenarios where the ATE of Z is examined with a true value of 0. We repeated the experiment 100 times and compared the mean, standard error, and p -values of different estimators. The experimental results are shown in Table 5. In most cases, PS methods especially AIPW method, exhibit smaller biases and standard errors compared to linear regression.

The first scenario is set as follows,

$$\begin{aligned} Z &\sim B(1, 0.3), \\ X_1 &\sim N(0.5Z + 0.5, 0.5), \\ X_2 &\sim N(0.2Z + X_1, 0.5), \\ Y &= X_1 + X_2 + 0.5 \exp(X_2). \end{aligned}$$

where $N(\mu, \sigma^2)$ refers to a normal distribution with mean μ and variance σ^2 , $B(1, p)$ refers to the Bernoulli distribution with parameter p . In this scenario, treatment variable Z only affects the observed values through the influence of confounding variables. The analysis results of the PS methods show that after controlling for confounding variables, the coefficients of the experimental treatment variable is no longer significant. This phenomenon is similar to what was observed in the analysis of the CLHLS dataset where linear assumption may not hold. In the results of the linear regression, the education has significant coefficients. However, when analyze using the PS methods and controlling for confounding variables, it appears to be insignificant.

The second scenario is set as follows,

$$\begin{aligned} X_1 &\sim N(0, 1), \\ X_2 &\sim U(0, 1), \\ X &\sim B(1, s(0.5X_1 + 0.6 \log(X_2))), \\ Y &= X_1 + X_2 + 1.5 \exp(X_2). \end{aligned}$$

where $s(x) = (1 + \exp(x))^{-1}$ is *sigmoid* function. In this situation, confounding variables can simultaneously affect both the outcomes and the treatment. However, the treatment has no causal effect on the observed values. Additionally, due to the presence of nonlinear terms, the results obtained from linear regression analysis still show that the coefficient of the treatment variable is significant. However, when analyzed using PS methods, this effect is no longer significant. This is similar to what we observed when analyzing the herring dataset.

The third scenario is similar to the second scenario in terms of setup, with the difference being that in the third scenario $Y \in \{0, 1\}$, and ,

$$\begin{aligned} Z &\sim B(1, s(1.5X_1 + 0.6 \log(X_2))), \\ Y &\sim B(1, s(X_1 + 0.3X_2 + 0.5 \exp(X_2))). \end{aligned}$$

In this situation, if we directly calculate the probability of $y = 1$ for the control and treatment group, there is an average difference of 0.20, and all the test results are significant. However, the treatment variable Z has no effect on the observed variable Y . The difference in the probability of $Y = 1$ between the two groups is entirely due to the influence of confounding variables. Therefore, the effect of the treatment should be close to zero in order to accurately reflect the actual situation. After eliminating the influence of confounding variables using PS methods, the treatment effects are estimated to be nearly zero. This phenomenon is similar to what was observed when analyzing the DBLP dataset, where the novelty and conventionality of a paper, along with confounding variables, can affect the number of citations. After removing the influence of confounding variables, the estimated impact of novelty and conventionality on citation counts decreases.

7 Conclusion

In this paper, we discuss the issue of selection bias in datasets from three different fields: scientific research, natural biology, and demography. Comparing the difference in responses or the conditional probability of $Y = 1$ between the treatment and control groups directly can lead to estimation bias. To mitigate this problem, some researchers choose to incorporate confounding variables into the analysis framework using linear regression, which can partially alleviate the issue. However, linear regression imposes rather strict assumptions on the model, so relying solely on linear model analysis may lead to erroneous conclusions, as discussed in Section 6 of this paper. In this study, we employ PS methods for analysis. PS methods addresses this problem by weighting or matching the samples to decouple the impact of treatment and confounding variables on the outcomes and estimate treatment effects more accurately. This suggests that it is necessary to compare different methods in decision-making and adjust them based on actual data to avoid making erroneous judgments.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

YZ: Formal analysis, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Uzzi B, Mukherjee S, Stringer M, Jones B. Atypical combinations and scientific impact. *Science*. (2013) 342:468–72. doi: 10.1126/science.1240474
2. Vollaard B, van Ours JC. Bias in expert product reviews. *J Econ Behav Organ*. (2022) 202:105–18. doi: 10.1016/j.jebo.2022.08.002
3. Cheng L, Zhang Y, Shen K. Understanding the pathways of the education-health gradient: evidence from the Chinese elderly. *China Econ Q*. (2014) 1:305–30. doi: 10.13821/j.cnki.ceq.2015.01.016
4. VanderWeele TJ. Principles of confounder selection. *Eur J Epidemiol*. (2019) 34:211–9. doi: 10.1007/s10654-019-00494-6
5. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. (1983) 70:41–55. doi: 10.1093/biomet/70.1.41
6. Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat Sci*. (2010) 25:1. doi: 10.1214/09-STS313
7. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc*. (1984) 79:516–24. doi: 10.1080/01621459.1984.10478078
8. Hirano K, Imbens GW, Ridder G. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*. (2003) 71:1161–89. doi: 10.1111/1468-0262.00442
9. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc*. (1994) 89:846–66. doi: 10.1080/01621459.1994.10476818
10. Li Zw, Liu Jm, Ren AG. Introduction to an individual-based standardization method-propensity score weighting. *Zhonghua Liu Xing Bing Xue Za Zhi*. (2010) 31:223–6. doi: 10.3760/cma.j.issn.0254-6450.2010.02.024
11. Thomas M, Miller DP, Morrissey TW. Food insecurity and child health. *Pediatrics*. (2019) 144:e20190397. doi: 10.1542/peds.2019-0397
12. Yang W, Tang J, Yi D, Li X, Wang X, Zhou X. GBM propensity score weighting for causal inference research. *World Sci Technol-Modernization Tradit Chin Med*. (2017) 1462–72.
13. McCaffrey DE, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol Methods*. (2004) 9:403. doi: 10.1037/1082-989X.9.4.403
14. Tang J, Zhang J, Yao L, Li J, Zhang L, Su Z. ArnetMiner: extraction and mining of academic social networks. In: *KDD'08*. San Jose, CA (2008). p. 990–8. doi: 10.1145/1401890.1402008
15. Louizos C, Shalit U, Mooij JM, Sontag D, Zemel R, Welling M. Causal effect inference with deep latent-variable models. In: *Advances in neural Information Processing Systems*. Long Beach, CA (2017). p. 30.
16. Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am J Epidemiol*. (2003) 158:280–7. doi: 10.1093/aje/kwg115