CLC _____ Number _____

UDC _____ Available for reference ☐Yes ☐No

SUSTech Southern University of Science and Technology

Undergraduate Thesis

Thesis Title: **Confounded Logistic Model with Instrumental Variables: Bias, Mediation and Identifiability**

Student Name: **Langtian Ma**

Student ID: **12012917**

Department: **Department of Statistics and Data Science**

Program: **Statistics**

Thesis Advisor: **Assistant Professor Yifang Ma**

Date: April 15, 2024

COMMITMENT OF HONESTY

1. I solemnly promise that the paper presented comes from my independent research work under my supervisor's supervision. All statistics and images are real and reliable.

2. Except for the annotated reference, the paper contents no other published work or achievement by person or group. All people making important contributions to the study of the paper have been indicated clearly in the paper.

3. I promise that I did not plagiarize other people's research achievement or forge related data in the process of designing topic and research content.

4. If there is violation of any intellectual property right, I will take legal responsibility myself.

Signature:

Date:

Confounded Logistic Model with Instrumental Variables:
——Bias, Mediation and Identifiability

[**ABSTRACT**]:

A significant issue in determining causal parameters from observational data is the presence of unmeasured confounders. Potentially consistent estimators can be obtained if instrumental variables (IVs) are available. An IV must satisfy conditions constraining it to be associated with the treatment but not with the outcome except through the treatment. In linear models, consistent estimators for causal parameters can be obtained by a simple two-stage least squares (TSLS) algorithm. However, challenges arise when researchers encounter categorical outcomes, necessitating generalized linear models. In these cases, the TSLS estimator tends to be biased. Although several methods have been proposed to address this bias, their reliability is often contingent on unrealistic conditions. This paper analyzes the bias inherent in two commonly used methods and introduces a novel procedure that achieves more reliable estimates. The existing techniques suffer from model misspecification problems, and our simulations reveal that they yield highly biased estimates, even under moderate conditions. By utilizing the residuals from the first stage of the TSLS algorithm, we propose a new method that can mediate the bias caused by the unmeasured confounder. We present simulation studies and real data experiments that demonstrate the performance of our proposed method. Finally, we further discuss the identifiability problem in the confounded logistic model and provide a theoretical guarantee for the identifiability of causal parameters when the confounder follows a uniform distribution. Code to reproduce our simulation is publicly available at https://github.com/LangtianM/IV_Logistic.

[**Key words**]: Causal Inference, Instrumental Variable, Logistic Model

[摘要]：从经验数据中确定因果参数的一个重要阻碍是可能存在未测量的混淆因子。如果有一个或多个工具变量（IV）可用，那么我们可能借此得出对因果参数的一致估计。工具变量必须满足某些特定条件，即它与处理变量（treatment）有关联，但除了通过处理变量之外，不与结果变量（outcome）相关。在线性模型中，可以通过两阶段最小二乘（TSLS）算法获得因果参数的一致估计。然而，结果变量为分类型变量（categorical variable）时，就需要使用广义线性模型对其进行刻画。在这些情况下，TSLS 估计往往是有偏的。尽管现在已有几种方法来解决这种偏差，但它们往往依赖于一些不现实的条件。在本文中，我们分析了两种常用方法固有的偏差，并引入了一种能够获得更准确估计的新方法。我们发现现有方法存在模型误设（Model-Misspecification）的问题，我们的模拟表明，即使在温和的条件下，它们得出的估计量也存在很大的偏差。利用 TSLS 算法第一阶段的残差，我们提出了一种可以缓解由未测量混淆因子引起的偏差的新方法。我们通过模拟实验和真实数据展示了新方法的优越性。最后，我们进一步讨论了混杂的逻辑模型中的可识别性问题，并在混杂因子遵循某些特定分布时提供了因果参数可识别性的理论保证。我们的模拟代码可在https://github.com/LangtianM/IV_Logistic上公开获取。

[关键词]：因果推断, 工具变量, 逻辑斯蒂模型

Table of Content

1. Introduction

A variety of methods can obtain the correlation within observational data. However, to understand the effect of our interventions, called *treatment*, on the variable of interest, called *outcome*, we need to assess the causal mechanism underlying the data. Learning causal structure from observational data is a challenging task if confounders exist that influence both the treatment and the outcome. For illustration, one may observe that education years positively correlate with income. However, this is not enough to conclude that education *causes* the income to increase. One may argue that factors such as innate ability or family background might cause both education years and income to increase. This does not necessarily mean there is a direct causal relationship between years of education and income, although it can still result in a positive correlation between the two.

One way to deal with this problem is *instrumental variable* (IV) regression[1]. With available instrumental variables, which are associated with the treatment but not with the outcome except through the treatment, we may obtain unbiased estimates of the causal effect. In the above example, we can exploit the birth month and the related school start age policy, which affects when a child can start school and subsequently when they are allowed to drop out, based on the compulsory schooling laws. The birth month is a valid instrumental variable because it affects the education years but is not directly related to income[2].

The two-stage least squares method is typically used for linear regression models to identify the causal parameters. However, researchers may also encounter data with categorical outcomes that should be modeled by a generalized linear model, under which the two-stage least squares method is generally biased. Several methods have been proposed to deal with the nonlinearity, such as generalized method of moments (GMM)[3], structural mean models[4], and Deep IV[5]. Nevertheless, data with categorical outcomes hardly satisfies the conditions required by these methods, and they still suffer from high bias caused by model misspecification[4].

In this paper, by utilizing the residuals from the first stage of the TSLS algorithm, we

propose a novel method to mediate the effect of confounders and produce more reliable estimations for the causal parameters. Section 2 introduces the problem setting and the existing methods in causal inference with confounders. Section 3 describes our mediation method and shows why it works under certain conditions. Section 4 illustrates the bias of existing methods and the effectiveness of our method by simulation experiments. Theoretical explanations of the simulation results are also provided. In Section 6, we discuss the identifiability issue in the confounded logistic model and provide a theoretical guarantee for the identifiability of causal parameters when the confounder follows a uniform distribution. We complete the paper with a brief summary of our work and future directions in Section 7.

2. Background

2.1 Structural Casual Model

One way of describing the causal mechanisms underlying observational data is the structural causal model (SCM), which depicts how nature assign values to the variables of interest[6]. Formally, A structural causal model consists of two sets of random variables $\mathcal{U}$ and $\mathcal{V}$, and a set of functions $\mathcal{F}$ that assigns each variable in $\mathcal{V}$ a value based on the values of other variables in the model. A variable X is a direct cause of Y if X appears in the function that assigns Y's value. The variables in $\mathcal{U}$ are mutually independent and called exogenous variables, meaning that they are external to the model. The variables in $\mathcal{V}$ are called endogenous variables, meaning that they are changed or determined by other variables in the model.

Every SCM is associated with a causal graph, which consists a set of nodes representing the variables in $\mathcal{U}$ and $\mathcal{V}$ and a set of directed edges representing the direct causal relationships between the variables. We give an example of a structural casual model and its corresponding causal graph for illustration:

Example 1. *Let U represent the temperature, X represent the ice cream sales, and Y represent the criminal rate. We have a structural equation model: $\mathcal{U} = \{U\}$, $\mathcal{V} = \{X, Y\}$, and*

$\mathcal{F} = \{f_Y, f_X\}$, *where* f_Y *is defined as:*

$$X = f_X(Z) = 0.5U$$
$$Y = f_Y(X) = 2U$$

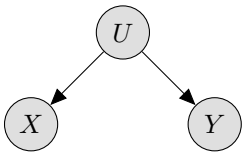


Figure 1 Causal Graph for Example 1

2.2 Confounding

In the context of causal inference, confounding is a phenomenon that occurs when a variable influences both the treatment and the outcome. In other words, a confounder is a common cause of the treatment and the outcome. Confounders can significantly impact causal inference in various scenarios, often leading to incorrect conclusions.

In example 1, the temperature U is a confounder because it influences both the ice cream sales X and the criminal rate Y. We may observe a positive correlation between ice cream sales and the criminal rate, but this correlation is not causal. Instead, the correlation is due to the common cause U. In analyzing the effect of maternal smoking during pregnancy on severe mental illness in offspring[7], as shown in Figure 2, confounding variables like maternal mental health and familial socioeconomic status can create a false association between maternal smoking and mental illness in offspring.

Various methods exist to identify the actual causal effect in the presence of confounding, including stratification[8], matching[9] and propensity score weighting[10] . These methods aim to balance the confounders by adjusting for the observed covariates to estimate the

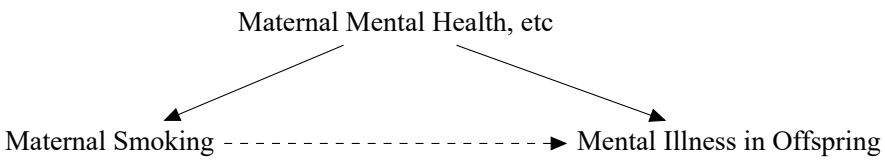


Figure 2 Causal Graph for Maternal Smoking and Mental Illness in Offspring. The dashed line indicates the causal relationship we are interested in but not certain about.

treatment effect without bias. However, they rely on the *ignorability assumption*:

Assumption 1. *(Ignorability Assumption, from an SCM Perspective) Let X be the treatment of interest, Y be the outcome, V be all other observed covariates, and U be unmeasured variables in the SCM. $f_Y(x, v, u)$ is the structural function that generates Y. The unconfoundedness assumption states that:*

$$f_Y(x, V, U) \perp\!\!\!\perp X|V. \tag{1}$$

Intuitively, it states that conditioning on observed covariates can rule out all unmeasured confounders that affect both the treatment and outcome simultaneously.

Unfortunately, this assumption is untestable from data and often violated in practice[8]. For example, there have been observational studies[11-12] show that vaccination against influenza remarkably reduces the elderly's risk of hospitalization and death after adjusting for observed covariates. Jackson et al. [13] doubt the results of these studies. The effect of vaccination should be specific to influenza season, but they found a greater effect before the influenza season, suggesting that the ignorability assumption fails and the observed effect is due to unobserved confounding.

While we are always uncertain whether the ignorability assumption holds, we can use *sensitivity analysis* to examine how fragile a result is against the possibility of unobserved confounders. It explores the impact of varying the assumptions under the estimated causal effect. By systematically modifying assumptions about the data and observing how these changes affect the results, researchers can assess the degree to which their findings are dependent on specific assumptions[14]. It helps identify the conditions under which the causal conclusions remain valid even if the original assumptions do not hold. However, sensitivity analysis does not directly control for the unmeasured confounders and cannot provide a way to identify the causal effect.

Nevertheless, when some additional variables satisfying certain conditions are available, we might be able to identify the causal effect without the ignorability assumption. Now,

we introduce the *instrumental variable method*.

2.3 Linear Instrumental Variable Model

The *instrumental variable method* has been widely used in econometrics to estimate causal effects in the absence of ignorability assumption. It relies on an additional variable, called the instrumental variable (IV), that satisfies two conditions: associated with the treatment and independent of the outcome conditional on the treatment. Formally, let X be the treatment, Y be the outcome, U be the confounder, Z be the IV and ϵ be random noise independent of all other variables. All variables are centered. If we assume the structural functions in $\mathcal{F}$ are linear, we have the following linear instrumental variable model:

Definition 1. *(Linear IV Model)* $\mathcal{U} = \{U, Z, \epsilon\}$, $\mathcal{V} = \{X, Y\}$ *and* $\mathcal{F}$ *is specified as:*

$$Y = \beta X + \eta U \tag{2}$$
$$X = \alpha Z + \gamma U + \epsilon. \tag{3}$$

where β, η, α, *and* γ *are the model parameters.*

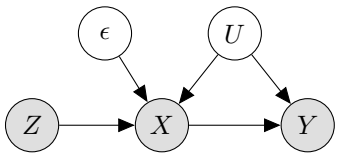


Figure 3 Causal Graph for Instrumental Variable Model

Note that the random varaible Z satisfies $Z \perp\!\!\!\perp Y|X$ and $Z \perp\!\!\!\perp U$. The causal graph for this model is shown in Figure 3. We want to identify the causal effect between X and Y, that is, to estimate the parameter β in the model.

The ordinary least square (OLS) regression of Y on X would fail to produce an unbiased estimate of β. This is because regressing Y on X directly is actually estimating the conditional expectation of Y:

$$\mathbb{E}[Y|X] = \beta X + \eta \mathbb{E}[U|X].$$

The estimated coefficient would not converge to β because $\mathbb{E}[U|X]$ is a nonzero function of

X, and it would introduce bias.

The two-stage least squares (TSLS) method can provide an unbiased estimate of β in this model. The first stage of TSLS is to regress the treatment variable X on the instrumental variable Z, then obtain the predicted value $\hat{X} = \hat{\mathbb{E}}[X|Z]$. The second stage is to fit a linear regression model of Y on $\hat{X}$, and the coefficient of $\hat{X}$ is the causal parameters of X on Y.

To see why this procedure works, note that conditioning both sides of (2) on Z gives:

$$\mathbb{E}[Y|Z] = \beta\mathbb{E}[X|Z], \tag{4}$$

where $\mathbb{E}[U|Z]$ degenerates since U and Z are independent. Then we can write

$$Y = \beta\mathbb{E}[X|Z] + w \tag{5}$$

for some random variable w independent of Z. We can see that (5) is a standard linear regression model for Y and $\mathbb{E}[X|Z]$. Therefore, the second stage of TSLS, which regress Y on $\hat{\mathbb{E}}[X|Z]$, would produce an unbiased estimate of β.

Although the TSLS method is efficient in linear models, there are many scenarios in which the outcome of interest is binary. For example, Johnston [15] studied the effect of surgical clipping on in-hospital death, which is a binary outcome and is improper to be characterized by a linear model. We need to use generalized linear models to model the data in such cases. We introduce the confounded logistic model to represent the situations in which the outcome is binary and is affected by unmeasured confounders.

2.4 Confounded Logistic Model

A confounded logistic model shares the same casual graph as the linear IV model (Figure 3). The difference is that the outcome Y follows a Bernoulli distribution, and the conditional expectation of Y is determined by a logistic function of X and U. The model is defined as:

Definition 2. *(Confounded Logistic Model)* $\mathcal{U} = \{U, Z, \epsilon\}$, $\mathcal{V} = \{X, Y\}$, *and* $\mathcal{F}$ *is specified*

as:

$$\mathbb{E}[Y|X,U] = h(\beta X + \eta U) \tag{6}$$
$$X = \alpha Z + \gamma U + \epsilon, \tag{7}$$

where $h(x) = 1/(1+e^{-x})$ *is the logistic function.*

Note that, unlike the linear regression model, X and U are nonseparable in this generalized linear model, which causes the major difficulty in applying traditional methods. Next, we review the existing IV methods for binary outcomes and show their limitations.

2.4.1 Two-Stage Least Squares Method

Consider the confounded logistic model defined by (6) and (7). The first stage of the TSLS method is the same as that in the linear model, which is to regress X on Z and obtain the predicted value $\hat{X} = \mathbb{E}[X|Z]$. The second stage becomes fitting a logistic regression model of Y on $\hat{X}$. However, Conditioning both sides on Z would not eliminate the effect of U and we could not obtain a formula like (4) because X and U are nonseparable in the logistic model:

$$\mathbb{E}[Y|Z] = \mathbb{E}[h(\beta X + \eta U)|Z] \neq h(\beta \mathbb{E}[X|Z] + \eta \mathbb{E}[U|Z]).$$

The only case in which the process is consistent is when $\beta = 0$, under which Z and Y are independent, and the estimating result would also be 0. Therefore, the TSLS method in confounded logistic models can be used for testing purposes[3].

2.4.2 Generalized Method of Moments

It has been suggested that the generalized method of moments (GMM) should be used for nonlinear models[16-18]. It assumes that X and U are separable, in which case the structural equation for generating Y can be written as:

$$Y = h(\beta_0 + \beta_1 X) + U. \tag{8}$$

The GMM method leverages the moment conditions that Z is independent of U implies

$$\mathbb{E}[f(Z)U] = \mathbb{E}[f(Z)(Y - h(\beta_0 + \beta_1 X))] = 0 \tag{9}$$

for all measurable function f. Given functions f_1 and f_2, we can estimate β_0 and β_1 by solving equation (9) for all f_i. A common choice of f in the unit variable case is $f_1(z) = 1$ and $f_2(z) = \mathbb{E}[X|Z = z]$. Therefore, to apply the GMM method, we need to solve the following equation numerically:

$$\frac{1}{n}\sum \begin{pmatrix} 1 \\ \mathbb{E}[X|Z = z_i] \end{pmatrix} (y_i - h(\beta_0 + \beta_1 x_i)) = 0. \tag{10}$$

Sound theoretical guarantees have been provided for the GMM method[19]. However, in real scenarios with categorical outcomes, the GMM method is not practical because

1. It is improper to model data with categorical outcomes using (8) since the outcome is not continuous.

2. If we forcibly use (8) to model such data, it would be *impossible* to find a variable Z related to X and not to U[3].

3. The choice of f is not unique, which may lead to different results in practice.

4. Solving (15) might be numerically unstable.

The only case that the GMM method is consistent with model (6) is when $\beta = 0$. In this case, X and U are naturally separable since X does not appear in the model. In section 5, we will further show the bias and instability of the GMM method in the confounded logistic model through simulation experiments.

3. Our Mediation Method: ResIV

We have seen the limitations of existing methods. Due to the nonseparability of X and U in the confounded logistic model, it is hard to give an unbiased point estimation of β. Instead, we propose Residual Logistic Regression with Instrumental Variables (ResIV)

to mediate the effect of the confounders and obtain a more reliable estimation of the causal parameters. It is based on a simple reparameterization trick. We define $\tilde{U}$ as the residual of X after regressing on Z, which can be easily estimated from data:

$$\tilde{U} = X - \alpha Z = \gamma U + \epsilon.$$

Then, we can rewrite the confounded logistic model (6) as:

$$\mathbb{E}[Y|X, U] = h(\beta X + \frac{\eta}{\gamma}\tilde{U} - \frac{\eta}{\gamma}\epsilon). \tag{11}$$

Note that in (11), X and $\tilde{U}$ are known, but the model is still confounded because $\frac{\eta}{\gamma}\epsilon$ becomes a new confounder term. However, we argue that the strength of random noise ϵ is always weaker than that of the unmeasured confounder U. Actually, when $\text{Var}(\frac{\eta}{\gamma}\epsilon) < \text{Var}(\eta U)$, i.e. under the condition

$$\text{Var}(\epsilon) < \gamma^2 \text{Var}(U) \tag{12}$$

the strength of the confounder in (11) is weaker than that in (6). In such scenarios, the effect of the confounder can be mitigated by fitting a logistic regression model using X, $\tilde{U}$ and Y. We summarize the procedure in the following algorithm.

Algorithm 1 ResIV

1: Regress X on Z to obtain the estimate of α, $\hat{\alpha}$.
2: Obtain the estimation of $\tilde{U}$: $\hat{U} = X - \hat{\alpha}Z$.
3: Fit a logistic regression model on Y with X and $\hat{U}$ as predictors.
4: The fitted coefficient of X is the estimation of β.

Interestingly, even if equation (12) is not satisfied, the ResIV method may still be effective. This is because a small γ, which leads to the failure of (12), also implies a minimal confounding effect on X, thus maintaining a low bias in estimating β. Furthermore, our simulation studies indicate that a high variance of ϵ does affect the effectiveness of ResIV, but it still outperforms the other methods. We plan to explore this phenomenon further in

section 4.4, although a comprehensive explanation for this behavior is not yet available.

In fact, we do not need to assume equation (7) is linear for ResIV to work. We only require U to be separable in the equation, i.e., we may replace (7) with

$$X = g(Z) + \gamma U + \epsilon, \tag{13}$$

for some nonlinear function g, in which case we may use more sophisticated machine learning models to estimate $\hat{\mathbb{E}}[X|Z] = \hat{g}(Z)$ and obtain $\hat{U} = X - \hat{g}(Z)$. However, we should be careful about overfitting when using complex models, which may produce a deviated estimate of $\tilde{U}$.

4. Simulation Studies

In this section, we demonstrate the performance of the three methods above by simulation studies and give explanations for the results. An ordinary logistic regression of Y on X (labeled as "Ordinary" in all figures) is included for comparison in all settings. We employ Powell's hybrid method to solve equation (15) for the GMM method. We generate data according to (6) and (7) with varying parameters or sample sizes and investigate how the bias of the estimators varies across different settings. Unless specified, parameters are set to 1 and the sample size is set to 10000. The random variables U and Z are generated from a standard normal distribution, and the random noise ϵ is generated from a normal distribution with a mean of zero and a standard deviation of 0.1 unless otherwise specified.

Before conducting simulation experiments, it is crucial to determine the appropriate range for varying parameters. Using improper parameters can cause the optimization algorithm to fail, even without confounding, which may result in misleading results. For example, we vary γ, which can be interpreted as the strength of confounder on X, across a wide range, from 0.1 to 10^9, and apply all previously described methods. Additionally, we fit a complete logistic regression model (labeled as "Complete"), which includes both X and U as explanatory variables. This model is not confounded and is expected to yield unbiased results.

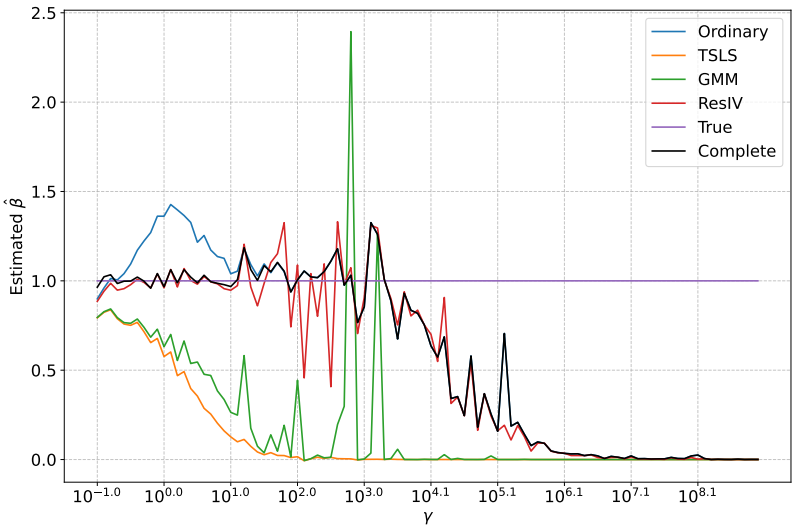
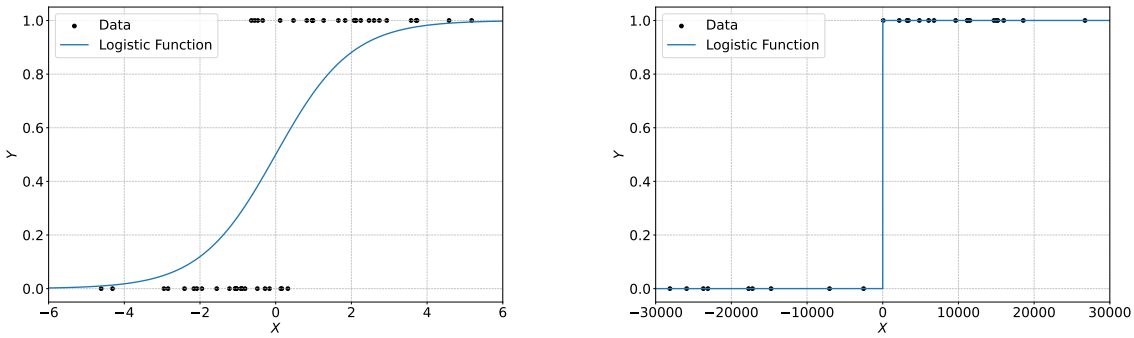


Figure 4 Estimation of various methods with varying γ.

As shown in Figure 4, the estimations of all methods collapse to zero as γ become significantly large. However, it is not caused by the increasing strength of the confounder because the complete logistic regression also fails. The sole reason lies in the nature of the logistic function h. When the scale of X is considerably large, $h(\beta X + \eta U)$ tends to be either very close to 0 or very close to 1 and the distribution of generated Y would be extreme. Figure 5a shows a logistic function and the generated data with $\gamma = 2$, and Figure 5b shows the same function and the data with $\gamma = 10000$. The latter is almost a step function because of the large scale of X, which is unsuitable and unrealistic for logistic regression and would lead the optimization algorithm to collapse.

Therefore, our simulation studies cannot vary parameters in a huge scope. Instead, we fit a complete logistic regression in advance to determine the appropriate range of parameters, which should ensure that the data is still suitable for logistic regression. A proper range would be in which a complete logistic regression is unbiased and stable. We set the range of β to $[0, 10]$, the range of η to $[-10, 10]$, the range of γ to $[0, 10]$, and the range of $\text{sd}(\epsilon)$ to $[0, 10]$.



a) $\gamma = 2$ **b)** $\gamma = 10000$

Figure 5 The distribution of X and Y with moderate and extreme γ.

4.1 Varying the causal parameter β

We first investigate how different values of β affect the estimations of various methods. As shown in figure 6, when β is close to 0, TSLS and GMM methods give accurate estimations while the ordinary logistic regression is biased, which aligns well with our discussion in section **??**. However, as β increases, the bias of TSLS and GMM proliferates and becomes much worse than the ordinary logistic regression. It suggests that the TSLS and GMM methods are only suitable for testing purposes and unreliable if we want to obtain estimations of true parameters. By contrast, our ResIV method gives accurate estimations in all scenarios.

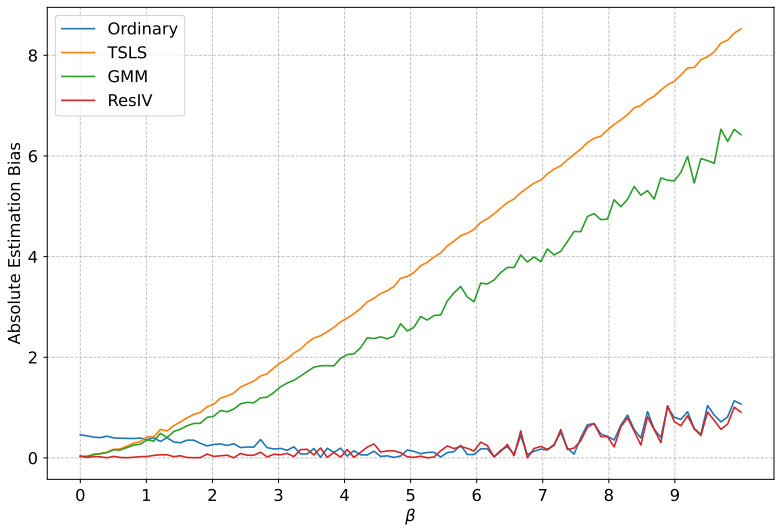


Figure 6 Estimation of various methods with different β.

4.2 Varying the strength of the confounder on Y

Next, we are interested in how η, which can be interpreted as the strength of the confounder on Y, affects the estimations. Figure 7 shows that the ordinary logistic regression deviates from the true value as η goes far from 0. The TSLS and GMM methods are also generally biased. Note that $TSLS$ gives accurate estimation when $\eta = -1$. This is because when $\eta = -1$, we have $\eta = -1 = -\beta = -\gamma = -\alpha$, equation (6) collapses to

$$\mathbb{E}[Y|X, U] = h(\beta X + \eta U) = h(\beta Z + \epsilon), \tag{14}$$

which is a standard logistic model for Z without any confounder, and the TSLS procedure is equivalent to an ordinary logistic regression of Y on Z.

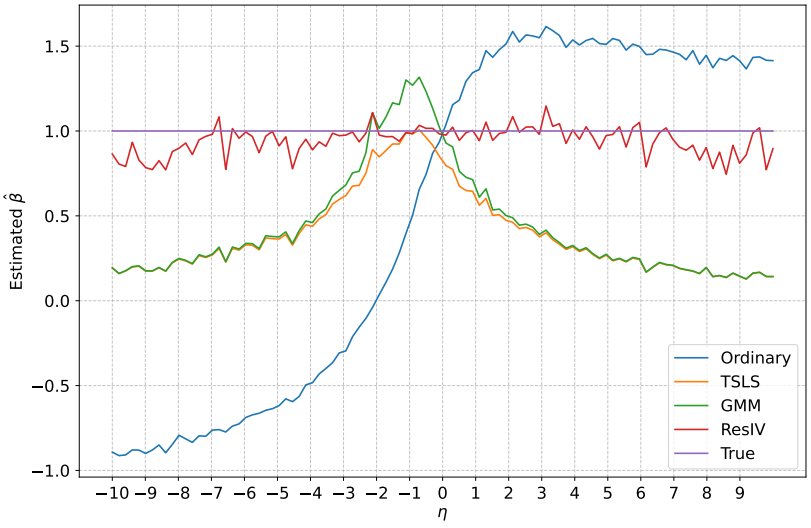


Figure 7 Estimation of various methods with different η.

4.3 Varying the strength of the confounder on X

As we are varying γ (Figure 8), TSLS and GMM methods are biased in most scenarios, even worse than the ordinary logistic regression. We can also observe that the estimations given by GMM method are unstable, which may be due to the instability of the numerical solution of equation (15). Note that when γ is small, the result of ResIV does not change much, which is consistent with our discussion in section 3.

An interesting observation is that the bias of the ordinary logistic regression slowly decreases as γ increases. We can give a qualitative explanation here: the confounding affects the result of estimation through the non-zero term $\mathbb{E}[U|X]$, and by equation 7, we can write it as:

$$\mathbb{E}[U|X] = \frac{1}{\gamma}(X - \alpha\mathbb{E}[Z|X] - \mathbb{E}[\epsilon|X]).$$

Thus, as γ grows, the scale of $\mathbb{E}[U|X]$ decreases, which may lead to a smaller bias in the ordinary logistic regression.

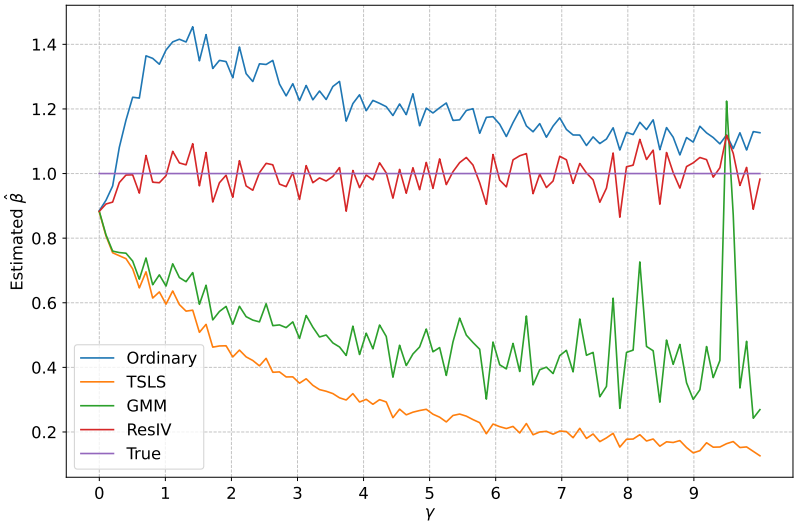


Figure 8 Estimation of various methods with different γ.

4.4 Varying the standard deviation of ϵ

As discussed in section 3, ResIV solves the reparameterized model (11). A high variance of ϵ would lead to a strong confounder effect in the model, which may cause ResIV to be more biased. However, our simulation studies show that ResIV still outperforms other methods when the variance of ϵ is large. Figure 9 shows that ResIV gives estimates better than all other methods when sd(ϵ) is small and behaves similarly to an ordinary logistic regression when sdϵ becomes large. The reason why ResIV maintains a moderate bias when sd(ϵ) grows large still requires further exploration.

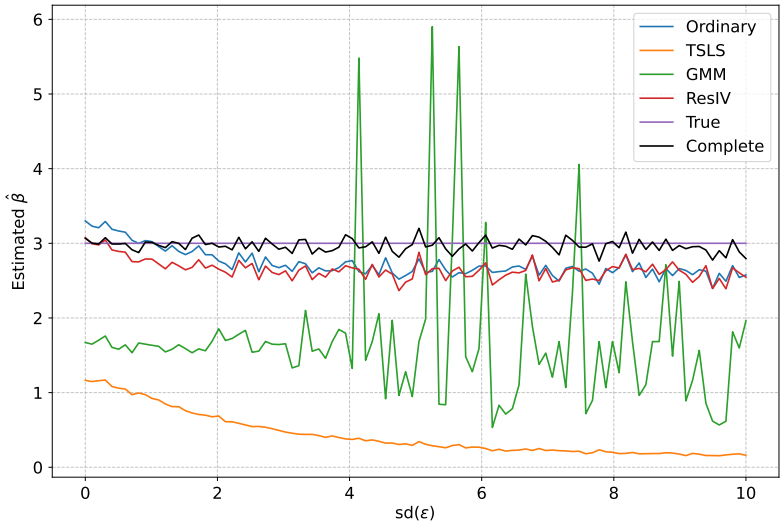


Figure 9 Estimation of various methods with different sd(ϵ).

5. Real Data Experiments

In this section, we use the cross-sectional labor force participation data from Mroz [20] to illustrate the performance of the proposed method. Suppose we are interested in the causal effect of education years on whether a woman participates in the labor force. We use the education years of her mother and father as instrumental variables. Other variables are also available in the data set. We select age, family income, education years of the husband, the unemployment rate in the country of residence, and an indicator of whether she lives in a city as observed confounders, which is supposed to affect both education years and labor force participation. We also select covariates, including the number of children under six years old, the number of hours worked by the husband, the hourly wage of the husband, and the federal marginal tax rate faced by the woman, which is supposed to affect the labor force participation but not the education years. There are 753 observations for each variable in the data set.

To fit TSLS and ResIV to data with multiple covariates, we can simply add the covariates to the final stages of each method. For GMM, we use the moment conditions as used by Henneman et al. [3]. Let $X \in \mathbb{R}$ be the education years, $Y \in \{0, 1\}$ be the indicator of

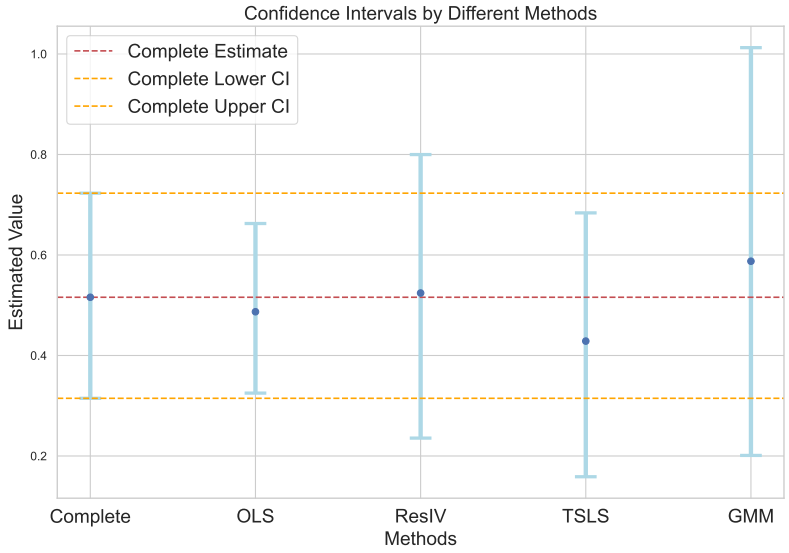


Figure 10 Experimental results. The blue dot corresponds to parameter estimates obtained through a specified method along the x-axis, while the light blue line signifies the 90% confidence interval around these estimates. The dotted red line represents parameter estimates derived from the complete model, accompanied by orange lines representing the upper and lower bounds of the 90% confidence interval.

whether a woman participates in the labor force, $Z \in \mathbb{R}^2$ be a vector of instrumental variables, and $V \in \mathbb{R}^4$ be a vector of covariates. The GMM method, in this case, solves the following equation:

$$\frac{1}{n}\sum \begin{pmatrix} 1 \\ \mathbb{E}[X|Z=z_i] \\ V \end{pmatrix} (y_i - h(\beta_0 + \beta_1 x_i + \beta^T v_i)) = 0, \tag{15}$$

where $\beta \in \mathbb{R}^4$ and $\mathbb{E}[X|Z=z_i]$ is estimated by a linear regression model in practice. The solved β_1 is the estimated causal parameter.

To evaluate various methods, we first fit a logistic regression model, referred to as the "Complete Model", using all variables described above except the instrumental variables. We assume that this regression reveals reliable parameters in the model since the confounders are observed. We then remove the confounders from the data to obtain incomplete data with unmeasured confounders and fit ordinary logistic regression, ResIV, TSLS, and GMM on it to obtain the estimated parameters. We obtain confidence intervals for the estimated parameters using the bootstrap method with 1000 replications. The results are shown in Figure 10.

We can observe that the estimation given by ResIV is the closest to the estimation given by the complete model, and the confidence interval of ResIV, though wider, behaves similarly to the complete model. Both TSLS and GMM give biased estimates, and the large confidence interval of GMM indicates that the method suffers from high variance.

6. Identifiability Analysis

Although our ResIV approach enhances the reliability of estimations, it does not eliminate the issues related to confounders. Consequently, it fails to achieve theoretical consistency. Deriving consistent estimators for the causal parameters in the confounded logistic model is still challenging.

Then, we may wonder if obtaining consistent estimators under this setting is possible. The identifiability of the confounded logistic model becomes another crucial issue to address. There have been some studies on the identifiability of the nonseparable confounded model with instrumental variables, which is defined as:

$$Y = h(X, U) \tag{16}$$

for some nonlinear function h, and h is restricted to be weakly monotonic on its final argument U. With the availability of instrumental variables, Chernozhukov et al. [21] shows that h is point-identifiable when it is continuous. Chesher [22] studied the case when h is discrete and showed that it is only partial-identifiable. However, though similar, model (16) is different from (6), since (16) gives a formula for Y while (6) models the conditional expectation $\mathbb{E}[Y|X, U]$. Further analysis is needed to determine whether the confounded logistic model is point-identifiable.

We provide a guarantee that the confounded logistic model is identifiable when the confounder follows a uniform distribution. The result is summarized in Theorem 1.

Theorem 1. *Under the model specified by (6) and (7), the causal parameter β is identifiable when the confounder U follows a uniform distribution.*

The proof of Theorem 1 is provided in the appendix. Please note that the result is still

limited due to our limited understanding of the unmeasured variable U in practical applications. Additionally, the general identifiability of this model remains an unresolved issue.

7. Discussion and Outlook

In this study, we examine the shortcomings of current methodologies for identifying causal parameters within the confounded logistic model. To address these limitations, we introduce the ResIV approach, designed to enhance the reliability of estimations. Nevertheless, while ResIV effectively mitigates issues related to confounders, it does not eliminate them and fails to achieve theoretical consistency. Moreover, ResIV relies on a strong model assumption that the function for generating X is linear (or, at least, separable in terms of U). This assumption may be violated in practice. Deriving consistent estimators for the causal parameters in the confounded logistic model remains an open challenge. We also have other ideas that were shown to be unsatisfactory during this project. We include them in the appendix A for future reference. In addition, we can only provide the theoretical guarantee for the identifiability of the causal parameters when the confounder follows a uniform distribution. The general identifiability of this model remains an unresolved issue. Interesting future directions arise from this study. First, a more rigorous theoretical characterization of ResIV is needed to understand its limitations and potential improvements. Second, relaxing the assumptions for ResIV would make it more practical. Finally, exploring the overall identifiability of the confounded logistic model is still an open question.

References

[1] STOCK J H, TREBBI F. Retrospectives: Who Invented Instrumental Variable Regression?[J/OL]. Journal of Economic Perspectives, 2003, 17(3): 177-194(2003-08-01) [2024-04-16]. https://pubs.aeaweb.org/doi/10.1257/089533003769204416. DOI: 10.1257/089533003769204416.

[2] ANGRIST J D, KRUEGER A B. Does compulsory school attendance affect schooling and earnings?[J]. The Quarterly Journal of Economics, 1991, 106(4): 979-1014.

[3] HENNEMAN T, LAAN M van der, HUBBARD A. Estimating Causal Parameters in Marginal Structural Models with Unmeasured Confounders Using Instrumental Variables[J/OL]. U.C. Berkeley Division of Biostatistics Working Paper Series, 2002(2002-01-01). https://biostats.bepress.com/ucbbiostat/paper104.

[4] CLARKE P S, WINDMEIJER F. Instrumental Variable Estimators for Binary Outcomes[J/OL]. Journal of the American Statistical Association, 2012, 107(500): 1638-1652(2012-12-01) [2024-03-19]. https://doi.org/10.1080/01621459.2012.734171. DOI: 10.1080/01621459.2012.734171.

[5] HARTFORD J, LEWIS G, LEYTON-BROWN K, et al. Deep IV: A Flexible Approach for Counterfactual Prediction[C/OL]. in: Proceedings of the 34th International Conference on Machine Learning. PMLR, 2017: 1414-1423 [2023-07-17]. https://proceedings.mlr.press/v70/hartford17a.html.

[6] PEARL J, GLYMOUR M, JEWELL N P. Causal inference in statistics: A primer[M]. John Wiley & Sons, 2016.

[7] KENDLER K. Causal Inference in Psychiatric Epidemiology.[J]. JAMA psychiatry, 2017, 74 6: 561-562. DOI: 10.1001/jamapsychiatry.2017.0502.

[8] DING P. A first course in causal inference[J]. ArXiv preprint arXiv:2305.18793, 2023.

[9] RUBIN D B. Matched sampling for causal effects[M]. Cambridge University Press, 2006.

[10] ROSENBAUM P R, RUBIN D B. The central role of the propensity score in observational studies for causal effects[J]. Biometrika, 1983, 70(1): 41-55.

[11] PUIG-BARBERÀ J, DIEZ-DOMINGO J, HOYOS S P, et al. Effectiveness of the MF59-adjuvanted influenza vaccine in preventing emergency admissions for pneumonia in the elderly over 64 years of age[J]. Vaccine, 2004, 23(3): 283-289.

[12] VOORDOUW A, STURKENBOOM M, DIELEMAN J, et al. Annual revaccination against influenza and mortality risk in community-dwelling elderly persons[J]. Jama, 2004, 292(17): 2089-2095.

[13] JACKSON L A, JACKSON M L, NELSON J C, et al. Evidence of bias in estimates of influenza vaccine effectiveness in seniors[J]. International journal of epidemiology, 2006, 35(2): 337-344.

[14] CINELLI C, HAZLETT C. Making Sense of Sensitivity: Extending Omitted Variable Bias[J/OL]. Journal of the Royal Statistical Society Series B: Statistical Methodology, 2020, 82(1): 39-67(2020-02-01) [2023-10-26]. https://academic.oup.com/jrsssb/article/82/1/39/7056023. DOI: 10.1111/rssb.12348.

[15] JOHNSTON S C. Effect of endovascular services and hospital volume on cerebral aneurysm treatment outcomes[J]. Stroke, 2000, 31(1): 111-117.

[16] NEWEY W K. Efficient Instrumental Variables Estimation of Nonlinear Models[J/OL]. Econometrica, 1990, 58(4): 809-837. JSTOR: 2938351 [2024-04-17]. https://www.jstor.org/stable/2938351. DOI: 10.2307/2938351.

[17] FOSTER E M. Instrumental variables for logistic regression: an illustration[J]. Social Science Research, 1997, 26(4): 487-504.

[18] AMEMIYA T. The nonlinear two-stage least-squares estimator[J]. Journal of econometrics, 1974, 2(2): 105-110.

[19] HANSEN L P. Large sample properties of generalized method of moments estimators[J]. Econometrica: Journal of the econometric society, 1982: 1029-1054.

[20] MROZ T A. The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions[M]. Stanford University, 1984.

[21] CHERNOZHUKOV V, IMBENS G W, NEWEY W K. Instrumental Variable Estimation of Nonseparable Models[J/OL]. Journal of Econometrics. Endogeneity, Instruments and Identification 2007, 139(1): 4-14(2007-07-01) [2024-04-07]. https://www.sciencedirect.com/science/article/pii/S030440760600100X. DOI: 10.1016/j.jeconom.2006.06.002.

[22] CHESHER A. Endogeneity and Discrete Outcomes[R/OL]. (2007-03-06) [2024-04-05]. https://www.ifs.org.uk/publications/3888.

[23] MISHKIN A. Instrumental Variables, DeepIV, and Forbidden Regressions[J].,

[24] KINGMA D, SALIMANS T, POOLE B, et al. Variational Diffusion Models[C/OL]. in: Advances in Neural Information Processing Systems: vol. 34. Curran Associates, Inc., 2021: 21696-21707 [2024-03-20]. https://proceedings.neurips.cc/paper/2021/hash/b578f2a52a0229873fefc2a4b06377fa-Abstract.html.

Appendix

A Discussion of Previous Ideas

In this section, we introduce and discuss our previous ideas that were shown to be unsatisfactory or infeasible during this project. We provide a brief introduction to each idea and discuss why it was not adopted in the main text.

A.1 Tayler Expansion of Logistic Function

The major difficulty in the confounded logistic model is that in equation (6), X and U are nonseparable, i.e. we cannot write the equation to be an addition of two terms, one of which is a function of X and the other is a function of U. A natural thought to separate X and U is to expand the logistic function into a Taylor series and neglect the higher-order terms. We have:

$$\begin{aligned} \mathbb{E}[Y|X,U] &= h(\beta X + \eta U) \\ &= \frac{1}{2} + \frac{1}{4}(\beta X + \eta U) + o(\beta X + \eta U)^2 \end{aligned} \tag{17}$$

However, assuming $(\beta X + \eta U)^2$ has an ignorable scale is unrealistic. The higher-order terms are not negligible in most cases, and the Taylor expansion idea is not applicable.

A.2 Learning a Conditional Expectation Function

Another idea to address the nonseparability problem is to rewrite (6) as:

$$h^{-1}(\mathbb{E}[Y|X,U]) = \beta X + \eta U. \tag{18}$$

Conditioning both sides on Z, we have:

$$\mathbb{E}[h^{-1}(\mathbb{E}[Y|X,U])|Z] = \beta \mathbb{E}[X|Z]. \tag{19}$$

Note that the right-hand side of (19) is the same as the right-hand side of (4) in the main text. The left-hand side is a function of Z, written as $\phi(Z)$, which has no closed-form expression but is specified given a distribution of the random variables. Inspired by Mishkin [23] and Kingma et al. [24], we may learn a neural network to approximate the function $\phi(Z)$ and then use it to estimate the causal parameters. However, training such a neural network is

nontrivial because we have no direct knowledge of $\mathbb{E}[Y|X,U]$.

A.3 Approximating the Conditional Expectation Function

Though we have no direct knowledge about $\mathbb{E}[Y|X,U]$, if we can find a satisfactory approximation, we can obtain the left-hand side directly since h is known. One idea would be using $\mathbb{E}[Y|X]$ as an approximation. This seemingly crude idea comes from the insights that $\mathbb{E}[Y|X,U]$ is the projection of Y on a subspace spanned by all measurable functions of X and U, and $\mathbb{E}[Y|X]$ is the projection of Y on a subspace spanned by all measurable functions of X. Note that the space of random variables is an infinite-dimensional vector space. Then we have the following intuitive arguments:

- If U and X are highly correlated, it means the angle between X and Y in the vector space is small, which implies the subspace spanned by X and U is close to the subspace spanned by X. In this case, $\mathbb{E}[Y|X]$ is a good approximation of $\mathbb{E}[Y|X,U]$.

- If the correlation between U and X is weak, it means the strength of the confounder is weak. In this case, the bias introduced by the confounder is not significant, and we can fit an ordinary logistic regression to identify the true parameters.

However, the above arguments are not rigorous and practical. The first argument does not imply the subspace spanned by *all measurable function of* X is close to the subspace spanned by *all measurable function of* X *and* U. In fact, $\mathbb{E}[Y|X]$ and $\mathbb{E}[Y|X,U]$ are certainly *nonlinear*, so we cannot guarantee that high correlation between X and U implies a good approximation. The second argument is also impractical because we cannot measure the strength of the confounder in practice.

B Proof of Theorem 1

To show the model specified by (6) and (7) is identifiable, we only need to show there's a one-to-one mapping between β and the conditional distribution function $q(x,z) = \Pr(Y = 1, X \leq x | Z = z)$. Expanding the function, we have:

$$\begin{aligned}
q(x,z) =& P(Y=1, X \leq x | Z = z) \\
=& P\left(Y = 1, U \leq \frac{1}{\gamma}(x - \alpha z) | Z = z\right) \\
=& P\left(Y = 1 | U \leq \frac{1}{\gamma}(x - \alpha z), Z = z\right) P\left(U \leq \frac{1}{\gamma}(x - \alpha z)\right) \\
=& \int_{-\infty}^{(x-\alpha z)/\gamma} h(\beta(\alpha z + \gamma u) + \eta u)\, dF_U(u) \\
=& \int_{-\infty}^{(x-\alpha z/\gamma)} h(\beta\alpha z + (\beta\gamma + \eta) u)\, dF_U(u)
\end{aligned}$$

where F_U is the cumulative distribution function of U.

When U is uniformly distributed, without loss of generality, we assume it follows an improper uniform distribution on $\mathbb{R}$. Then we have

$$\begin{aligned}
q(x,z) \propto& \frac{1}{\beta\gamma + \eta} \int_{-\infty}^{x - \alpha z/\gamma} h(\beta\alpha z + (\beta\gamma + \eta)u)\, d(\beta\alpha z + (\beta\gamma + \eta)u) \\
=& \frac{1}{\beta\gamma + \eta} \ln(1 + \exp(\beta\alpha z + (\beta\gamma + \eta)(x - \alpha z)/\gamma) \\
=& 1/(\beta\gamma + \eta) \ln\left(1 + \exp\left(\left(\beta\alpha - \frac{(\beta\gamma + \eta)\alpha}{\gamma}\right) z + \frac{\beta\gamma + \eta}{\gamma} x\right)\right) \\
=& \frac{1}{\beta\gamma + \eta} \ln\left(1 + \exp\left(-\frac{\eta\alpha}{\gamma} z + (\beta + \frac{\eta}{\gamma}) x\right)\right)
\end{aligned}$$

Note that the coefficient of z and x in the last equation are uniquely determined by the function $q(x,z)$, and α is identifiable by a simple linear regression of X on Z. Thus, $\frac{\eta}{\gamma}$ is uniquely determined and hence β is identifiable.

Acknowledgement

This paper not only reflects my work and exploration in the field of causal inference but also the support and encouragement of many people.

I would like to express my sincere gratitude to my thesis supervisor Yuting Ye and my academic supervisor Yifang Ma for their guidance, encouragement, and patience throughout this research. I am grateful for their support and advice, which are invaluable to my academic journey and life.

Additionally, I woxuld like to thank my friends in the Department of Statistic and Data Science. We have studied together, worked together and shared our lives together. I am grateful for their support and encouragement, which have been a source of strength for me.

Finally, I would like to express my deep gratitude to my family. They have always been there for me, supporting me and encouraging me. I am grateful for their love and care.