On Estimating Causal Effects based on Supplemental Variables

Takahiro Hayashi
Department of Statistics
The Graduate University for Advanced Studies
10-3,Midori-cho, Tachikawa, Tokyo, Japan
hayashi.takahiro@ism.ac.jp

Manabu Kuroki
Department of Data Science
The Institute of Statistical Mathematics
10-3,Midori-cho, Tachikawa, Tokyo, Japan
mkuroki@ism.ac.jp

Abstract

This paper considers the problem of estimating causal effects of a treatment on a response using supplementary variables. Under the assumption that a treatment is associated with a response through a univariate supplementary variable in the framework of linear regression models, Cox (1960) showed that the estimation accuracy of the regression coefficient of the treatment on the response in the single linear regression model can be improved by using the recursive linear regression model based on the supplementary variable from the viewpoint of the asymptotic variance. However, such assumptions may not hold in many practical situations. In this paper, we consider the situation where a treatment is associated with a response through a set of supplementary variables in both linear and discrete models. Then, we show that the estimation accuracy of the causal effect can be improved by using the supplementary variables. Different from Cox (1960), the results of this paper are derived without the assumption of Gaussian error terms in linear models or dichotomous variables in discrete models. The results of this paper help us to obtain the reliable evaluation of causal effects from observed data.

1 Introduction

Supplementary variables are considered as variables that are not of interest in themselves but help us to identify target quantities and/or understand data generating mechanism in practical studies. For example, the instrumental variable (Bowden and Turkington, 1984) is one of supplementary variables because it enables us to identify causal effects under certain assumptions (Angrist et al, 1996) and evaluate the bounds on causal effects under milder conditions (Manski, 2007; Pearl, 2009) in the presence of unobserved confounders. In addition, proxy variables of a treatment, a response and/or covariates are also considered as supplementary variables because the proxy variables enable us to identify the causal effect even when it is difficult to observe the variables of our main interest (Cai and Kuroki, 2008; Kuroki, 2007; Kuroki and Pearl, 2013). Furthermore, intermediate variables are often considered as supplementary variables since they are used to identify various kinds of causal quantities (Pearl, 2001, 2009) as well as to understand data generating mechanisms in mediation analysis (Baron and Kenny, 1986; Imai et al, 2011; MacKinnon, 2008).

In this paper, we focus on the estimation problem of causal effects using a set of supplementary variables including intermediate variables. When data generating mechanism among variables can be described by nonparametric structural equation models and the corresponding directed acyclic graph, Pearl (2009) provided the front door criterion as the graphical identification condition for causal effects based on intermediate variables. In addition, in the framework of Gaussian linear structural equation models, Kuroki (2000) formulated the exact variance of the causal effect based on the front door criterion. Furthermore, Kuroki and Cai (2004) compared some graphical identification conditions in terms of the asymptotic variance of causal effects. On the other hand, under the assumption that a treatment is associated with a response through a univariate intermediate variable, Cox (1960) showed that the estimation accuracy of the regression coefficient of the treatment on the response in the single linear regression model can be improved by using the recursive linear regression model based on the intermediate variable from the viewpoint of the asymptotic variance. Under the assumption that a univariate intermediate variable, Hui and Zhongguo (2008) and Ramsahai (2012) compared the front-door criterion, the back-door criterion (Pearl, 2009) and the ex-

Appearing in Proceedings of the 17^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33.

tended back door criterion (Lauritzen, 2001) in terms of the asymptotic variances of causal effects based on Gaussian linear structural equation models. In addition, Pearl (2012) discussed the same problem as Cox (1960) based on discrete models and stated that the same result as Cox (1960) can be obtained when a treatment, a response and an intermediate variable are dichotomous. However, such assumptions may not hold in many practical situations.

In this paper, we consider the situation where a treatment is associated with a response through a set of supplementary variables in both linear and discrete models. Then, we show that the estimation accuracy of the causal effect can be improved by using the supplementary variables. Different from Cox (1960), the results of this paper are derived without the assumption of Gaussian error terms in linear models or dichotomous variables in discrete models. In addition, we apply our results to an empirical example from process analysis of coating conditions for car bodies in quality control (Kuroki, 2012; Okuno et al, 1986). The results of this paper help us to obtain the reliable evaluation of causal effects from observed data.

2 Causal Bayesian Network

Let $\text{pr}(v_1, \ldots, v_n)$ be a strictly positive (or non-degenerate) joint distribution of a set $\boldsymbol{V} = \{V_1, \cdots, V_n\}$ of variables, $\text{pr}(v_i|v_j)$ the conditional distribution of V_i given $V_j = v_j$ ($V_i, V_j \in \boldsymbol{V}$) and $\text{pr}(v_i)$ the marginal distribution of V_i. Similar notation is used for other distributions. For graph theoretic terminology used in this paper, for example, refer to Pearl (2009).

When a directed acyclic graph $G = (\boldsymbol{V}, \boldsymbol{E})$ with a set $\boldsymbol{V}$ of variables and a set $\boldsymbol{E}$ of arrows is given and the joint distribution of $\boldsymbol{V}$ is factorized recursively according to the graph G as the following equation, the graph is called a Bayesian network:

$$\text{pr}(v_1, \cdots, v_n) = \prod_{i=1}^{n} \text{pr}\{v_i|\text{pa}(v_i)\}, \qquad (1)$$

where $\text{pa}(v_i)$ is a set of parents of V_i in G. When $\text{pa}(v_i)$ is an empty set, $\text{pr}\{v_i|\text{pa}(v_i)\}$ is the marginal distribution $\text{pr}(v_i)$ of v_i.

If a joint distribution is factorized recursively according to the directed acyclic graph G, the conditional independencies implied by the factorization (1) can be read off from G according to the d-separation criterion (Pearl, 1988), that is, if $\boldsymbol{W}_1$ d-separates $\boldsymbol{W}_2$ from $\boldsymbol{W}_3$ in G ($\boldsymbol{W}_1$, $\boldsymbol{W}_2$, $\boldsymbol{W}_3 \subset \boldsymbol{V}$), then $\boldsymbol{W}_2$ is conditionally independent of $\boldsymbol{W}_3$ given $\boldsymbol{W}_1$ in the corresponding recursive factorization (1); See Pearl (1988, 2009).

The recursive factorization (1) can be given causal interpretation, and the arrows in G as representing potential causal influences between the corresponding variables and, alternatively, the absence of arrows represents no direct causal influence between the corresponding variables. In this interpretation, the recursive factorization (1) still holds, but the factors are further assumed to represent autonomous data generating mechanism, that is, each family conditional probability $\text{pr}\{v_i|pa(v_i)\}$ represents a stochastic process by which the values of V_i are assigned in response to the values $\text{pa}(v_i)$ (previously chosen for V_i's parents), and the stochastic variation of this assignment is assumed independent of the variations in all other assignments in the model (Bareinboim et al, 2011). Then, the Bayesian network G is called a causal Bayesian network.

Based on the theory of causal Bayesian networks, Pearl (2009) defined a causal effect as a distribution of a response when conducting an external intervention to a treatment, where an 'external intervention' means that a variable is forced to take on some fixed value, regardless of the values of other variables.

Definition 1(Causal effect)
Let $\boldsymbol{V} = \{X, Y\} \cup \boldsymbol{Q}$ ($\{X, Y\} \cap \boldsymbol{Q} = \phi$) be a set of variables represented in a causal Bayesian network G. The causal effect of X on Y is defined by

$$\text{pr}\{y|\text{do}(X = x)\} = \sum_{q} \frac{\text{pr}(x, y, \boldsymbol{q})}{\text{pr}\{x|\text{pa}(x)\}}, \qquad (2)$$

where $\text{do}(X = x)$ means that X is set to a value x by an external intervention. In addition, summation signs are replaced by integrals whenever the summed variables are continuous. □

Given a causal Bayesian network G, in order to evaluate the causal effect of X on Y from a joint factorized distribution of observed variables, it is required to observe not only X and Y but also a set $\boldsymbol{W}$ of other variables, such as confounders. Pearl (2009) provided 'the back door criterion' as one of graphical identifiability criteria for the causal effect, where 'identifiable' means that the causal effect can be determined uniquely from a joint distribution of observed variables.

Definition 2 (Back door criterion)
Suppose that X is a non-descendant of Y in a directed acyclic graph G. If a set $\boldsymbol{W}$ of vertices satisfies the following conditions relative to an ordered pair (X, Y) of vertices, then $\boldsymbol{W}$ is said to satisfy the back door criterion relative to (X, Y):

(i) no vertex in $\boldsymbol{W}$ is a descendant of X;

(ii) $\boldsymbol{W}$ blocks every path between X and Y that contains an arrow pointing to X. □

If a set $\boldsymbol{W}$ of variables satisfies the back door criterion relative to (X, Y) in a causal Bayesian network G, then the causal effect of X on Y is identifiable through the observa-

tion of $\boldsymbol{W} \cup \{X, Y\}$ and is given by the formula

$$\text{pr}\{y|\text{do}(X = x)\} = \sum_{w} \text{pr}(y|x, \boldsymbol{w})\text{pr}(\boldsymbol{w}). \quad (3)$$

3 Analytical Results

3.1 Linear Model

In this section, for two distinct sets $\boldsymbol{S}$ and $\boldsymbol{W}$ of variables, we assume that X is conditionally independent of Y given a set $\boldsymbol{S} \cup \boldsymbol{W}$ of supplementary variables and $\boldsymbol{W}$ satisfies the back door criterion relative to (X, Y). This situation can be described as the graph shown in Fig. 1.

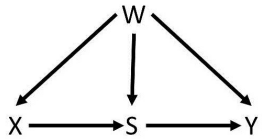

Fig. 1: Causal diagram (1)

When observed data is assumed to be generated according to a linear structural equation model and the corresponding directed acyclic graph shown in Fig. 1, in order to estimate $dE\{Y|\text{do}(X = x)\}/dx$, we consider the following linear regression model:

$$Y = \beta_{y.xw} + \beta_{yx\cdot xw}X + B_{yw.xw}\boldsymbol{W} + \epsilon_{y.xw}, \quad (4)$$

where $\beta_{y.xw}$, $\beta_{yx\cdot xw}$ and $B_{yw.xw}$ are an intercept, the regression coefficients of X and the vector of the regression coefficients of $\boldsymbol{W}$ in the regression model of Y on X and $\boldsymbol{W}$, respectively. Similar notation is used for other regression parameters. In addition, the error term $\epsilon_{y.xw}$ is assumed to follow the distribution with the mean 0 and the variance $\sigma_{yy.xw} > 0$ and $\{X\} \cup \boldsymbol{W}$ is independent of $\epsilon_{y.xw}$. Then, we have $dE\{Y|\text{do}(X = x)\}/dx = \beta_{yx\cdot xw}$.

Noting that X is conditionally independent of Y given $\boldsymbol{S} \cup \boldsymbol{W}$, since we have $\beta_{yx\cdot xw} = B_{ys.xsw}B_{sx.xw} = B_{ys.sw}B_{sx.xw}$, we can also use the following recursive regression model to estimate $B_{ys.sw}$ and $B_{sx.xw}$:

$$\begin{aligned} Y &= \beta_{y.sw} + B_{ys.sw}\boldsymbol{S} + B_{yw.sw}\boldsymbol{W} + \epsilon_{y.sw}, \quad (5) \\ \boldsymbol{S} &= B_{s.xsw} + B_{sx.xsw}X + B_{ss.xsw}\boldsymbol{S} \\ &\quad + B_{sw.xsw}\boldsymbol{W} + \boldsymbol{\epsilon}_{s.xsw}, \quad (6) \end{aligned}$$

where a set $\{\epsilon_{y.sw}\} \cup \boldsymbol{\epsilon}_{s.xsw}$ of error terms follows the multivariate distribution with mean vector $\boldsymbol{0}$ and positive-definite diagonal covariance matrix. $B_{s.xsw}$ and $B_{ss.xsw}$ are the vector of the intercepts and the matrix of the regression coefficients in the regression models of $\boldsymbol{S}$ on X, $\boldsymbol{S}$ and $\boldsymbol{W}$, respectively. In addition, it is assumed that the explanatory variables in the regression models (5) and (6) are independent of error terms which appear in each equation. Furthermore, the elements of $\boldsymbol{S}$ are arranged to satisfy that $B_{ss.xsw}$ is an upper triangular matrix. Here, in the discussion of this section, when it is known from prior knowledge that a treatment is conditionally independent of a response given supplementary variables and it is justified that observed data is generated based on the distribution satisfying such a conditional independence, the distributions of explanatory variables and error terms are not limited to the Gaussian distribution, as far as the ordinary least squares method can be applied to obtaining the unbiased estimators of the regression coefficients. On the other hand, when it is necessary to conduct a statistical test of the conditional independence between the treatment and the response given supplementary variables, it is required to make assumptions about the nature of the error terms and it is commonly assumed that the error terms follow the Gaussian distribution.

Equation (6) can be rewritten by

$$\begin{aligned} \boldsymbol{S} &= (I - B_{ss.xsw})^{-1}B_{s.xsw} + (I - B_{ss.xsw})^{-1}B_{sx.xsw}X \\ &\quad + (I - B_{ss.xsw})^{-1}B_{sw.xsw}\boldsymbol{W} \\ &\quad + (I - B_{ss.xsw})^{-1}\boldsymbol{\epsilon}_{s.xsw}, \\ &= B_{s.xw} + B_{sx.xw}X + B_{sw.xw}\boldsymbol{W} + \boldsymbol{\epsilon}_{s.xw}, \end{aligned}$$

where $B_{s.xw} = (I - B_{ss.xsw})^{-1}B_{s.xsw}$, $B_{sx.xw} = (I - B_{ss.xsw})^{-1}B_{sx.xsw}$, $B_{sw.xw} = (I - B_{ss.xsw})^{-1}B_{sw.xsw}$ and $\boldsymbol{\epsilon}_{s.xw} = (I - B_{ss.xsw})^{-1}\boldsymbol{\epsilon}_{s.xsw}$. Thus, $B_{s.xw}$, $B_{sx.xw}$ and $B_{sw.xw}$ are estimable by using the ordinary least square method based on the regression model of the element of $\boldsymbol{S}$ on X and $\boldsymbol{W}$. Here, it is noted that the elements of $\boldsymbol{\epsilon}_{s.xw}$ may not be independent of each other. On the other hand, instead of equation (5), we can also use

$$\begin{aligned} Y &= \beta_{y.xsw} + B_{ys.xsw}\boldsymbol{S} + \beta_{yx\cdot xsw}X \\ &\quad + B_{yw.xsw}\boldsymbol{W} + \epsilon_{y.xsw} \quad (7) \end{aligned}$$

to estimate $B_{ys.xsw}$, which is consistent with $B_{ys.sw}$ since X is conditionally independent of Y given $\boldsymbol{S} \cup \boldsymbol{W}$.

Let $\hat{\beta}_{yx.xw}$ be the ordinary least square estimator of $\beta_{yx.w}$, with a similar notation used for other ordinary least square estimators. Then, for $\hat{\beta}_{yx.xw}$, $\hat{B}_{ys.sw}\hat{B}_{sx.xw}$ and $\hat{B}_{ys.xsw}\hat{B}_{sx.xw}$, if their variances exist, both

$$\text{var}(\hat{\beta}_{yx.xw}) \geq \text{var}(\hat{B}_{ys.sw}\hat{B}_{sx.xw}) \quad (8)$$

and

$$\text{var}(\hat{B}_{ys.xsw}\hat{B}_{sx.xw}) \geq \text{var}(\hat{B}_{ys.sw}\hat{B}_{sx.xw}) \quad (9)$$

hold. The proof is provided in Appendix I.

Since the result is based on both the exact variance and several supplementary variables, the result can be considered as the improvement of Cox (1960) whose discussion is based on the asymptotic variance and a univariate supplementary variable. To our surprise, as far as no perfect multi-collinearity occurs, even if some elements of

$\{X\} \cup \boldsymbol{S} \cup \boldsymbol{W}$ are highly associated, the estimation accuracy of $\hat{B}_{ys.sw}\hat{B}_{sx.xw}$ is superior to that of $\hat{\beta}_{yx.xw}$. However, $\text{var}(\hat{\beta}_{yx.xw}) - \text{var}(\hat{B}_{ys.xsw}\hat{B}_{sx.xw})$ may not be nonnegative.

3.2 Discrete Model

In this section, we assume that $\boldsymbol{S} \cup \boldsymbol{W} \cup \{X, Y\}$ follows the multinomial distribution $\text{MN}\{N, \text{pr}(x, y, \boldsymbol{s}, \boldsymbol{w})\}$ with sample size N and cell probabilities $\text{pr}(x, y, \boldsymbol{s}, \boldsymbol{w}) > 0$ for any $x, y, \boldsymbol{s}$ and $\boldsymbol{w}$. When X is conditionally independent of Y given a set $\boldsymbol{S} \cup \boldsymbol{W}$ of supplementary variables and $\boldsymbol{W}$ satisfies the back door criterion relative to (X, Y), as shown in Fig. 1, we consider the following three quantities for the causal effect of X on Y:

$$\text{pr}\{y|\text{do}(X = x)\} = \sum_{w} \text{pr}(y|x, \boldsymbol{w})\text{pr}(\boldsymbol{w}),$$

$$\text{pr}\{y|\text{do}(X = x); \boldsymbol{s}\} = \sum_{s,w} \text{pr}(y|\boldsymbol{s}, \boldsymbol{w})\text{pr}(\boldsymbol{s}|x, \boldsymbol{w})\text{pr}(\boldsymbol{w}),$$

$$\begin{aligned}&\text{pr}\{y|\text{do}(X = x); x', \boldsymbol{s}\}\\ &= \sum_{s,w} \text{pr}(y|x', \boldsymbol{s}, \boldsymbol{w})\text{pr}(\boldsymbol{s}|x, \boldsymbol{w})\text{pr}(\boldsymbol{w})\end{aligned}$$

for $x \neq x'$. Here, $\text{pr}\{y|\text{do}(X = x); \boldsymbol{s}\}$ is the causal effect of X on Y when the information on $\boldsymbol{S}$ is used, and $\text{pr}\{y|\text{do}(X = x); x', \boldsymbol{s}\}$ is the causal effect of X on Y when the information on both $\boldsymbol{S}$ and $X = x'$ are used . Then, it is obvious that $\text{pr}\{y|\text{do}(X = x)\} = \text{pr}\{y|\text{do}(X = x); \boldsymbol{s}\} = \text{pr}\{y|\text{do}(X = x); x', \boldsymbol{s}\}$ holds from the assumptions.

Letting n_{xysw} represents the number of subjects in cell $(X, Y, \boldsymbol{S}, \boldsymbol{W}) = (x, y, \boldsymbol{s}, \boldsymbol{w})$, with a similar notation used for other cells, $\text{pr}(y|x, \boldsymbol{w})$, $\text{pr}(y|x, \boldsymbol{s}, \boldsymbol{w})$, $\text{pr}(y|\boldsymbol{s}, \boldsymbol{w})$, $\text{pr}(\boldsymbol{s}|x, \boldsymbol{w})$ and $\text{pr}(\boldsymbol{w})$ are estimated by

$$\hat{\text{pr}}(y|x, \boldsymbol{w}) = \frac{n_{xyw}}{n_{xw}}, \quad \hat{\text{pr}}(y|x, \boldsymbol{s}, \boldsymbol{w}) = \frac{n_{xysw}}{n_{xsw}},$$

$$\hat{\text{pr}}(y|\boldsymbol{s}, \boldsymbol{w}) = \frac{n_{ysw}}{n_{sw}}, \quad \hat{\text{pr}}(\boldsymbol{s}|x, \boldsymbol{w}) = \frac{n_{xsw}}{n_{xw}},$$

$$\hat{\text{pr}}(\boldsymbol{w}) = \frac{n_w}{N},$$

respectively, where $N = \sum_{x,y,s,w} n_{xysw}$. Then, both

$$\text{a.var}\left[\hat{\text{pr}}\{y|\text{do}(X = x)\}\right] \geq \text{a.var}\left[\hat{\text{pr}}\{y|\text{do}(X = x); \boldsymbol{s}\}\right] \tag{10}$$

and

$$\begin{aligned}&\text{a.var}\left[\hat{\text{pr}}\{y|\text{do}(X = x); x', \boldsymbol{s}\}\right]\\ &\quad \geq \text{a.var}\left[\hat{\text{pr}}\{y|\text{do}(X = x); \boldsymbol{s}\}\right]\end{aligned} \tag{11}$$

hold for $x \neq x'$, where $\text{a.var}(\cdot)$ indicates the asymptotic variance. The proof is provided in Appendix II. Here, it is noted that there is no qualitative inequality relationship between $\hat{\text{pr}}\{y|\text{do}(X = x)\}$ and $\hat{\text{pr}}\{y|\text{do}(X = x); x', \boldsymbol{s}\}$.

From Sections 3.1 and 3.2, the supplementary variables can improve the estimation accuracies of causal effects if we use the conditional distribution of Y given the supplementary variables only to estimate the causal effects, under the assumption that a treatment is conditionally independent of a response given supplementary variables. However, it is difficult to provide the qualitative judgment whether or not the estimation accuracies can be improved by using the supplementary variables if the assumption does not hold. In addition, when we apply our results to real data analysis with the finite sample size, it is noted that if the difference between the two variances is very small then the inequality relationships may be reversed because of sampling variability.

4 Simulation Experiments

4.1 Linear Model

In this section, through numerical experiments, we examine the inequality relationships between the estimation accuracies stated in Section 3.1. For simplicity, we consider the causal Bayesian network shown in Fig. 2 and the corresponding linear structural equation model in which X is conditionally independent of Y given S and an empty set satisfies the back door criterion relative to (X, Y), that is,

$$Y = S + \epsilon_{y.s}, S = X + \epsilon_{s.x}, X = \epsilon_x.$$

This situation can be considered as a simple version of Fig. 1.

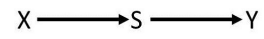

Fig. 2: Causal diagram (2)

Under the setting, we consider the following two cases:

Case 1 (Symmetric distribution): $\epsilon_{y.s}$ and ϵ_x follow the standard normal distribution $N(0, 1)$ independently, but $\epsilon_{s.x}$ follows the normal distribution $N(0, \sigma_{ss.x})$ with the mean 0 and the variance $\sigma_{ss.x}$ and $\epsilon_{s.x}$ is marginally independent of $\{\epsilon_{y.s}, \epsilon_x\}$.

Case 2 (Asymmetric distribution): $\epsilon_{y.s}$ and ϵ_x follow the exponential distribution with the location -1 and the scale 1 independently, but $\epsilon_{s.x}$ follows the exponential distribution with the location $-\lambda$ and the scale λ and $\epsilon_{s.x}$ is marginally independent of $\{\epsilon_{y.s}, \epsilon_x\}$ for $\lambda > 0$.

We set $\sigma_{ss.x}$ to 0.01, 0.500 and 1.000 in Case 1 and λ^2 to 0.01, 0.500 and 1.000 in Case 2 respectively. It is reasonable to consider that the multi-collinearity has a serious effect on the estimation accuracies of the causal effect when the parameters $\sigma_{ss.x}$ and λ are small, but not when

Table 1: Simulation Results (Linear Model)

	$\sigma_{ss.x}=0.01$			$\sigma_{ss.x}=0.5$			$\sigma_{ss.x}=1.000$		
Case 1	$\hat{\beta}_{ys.s}\hat{\beta}_{sx.x}$	$\hat{\beta}_{yx.x}$	$\hat{\beta}_{ys.xs}\hat{\beta}_{sx.x}$	$\hat{\beta}_{ys.s}\hat{\beta}_{sx.x}$	$\hat{\beta}_{yx.x}$	$\hat{\beta}_{ys.xs}\hat{\beta}_{sx.x}$	$\hat{\beta}_{ys.s}\hat{\beta}_{sx.x}$	$\hat{\beta}_{yx.x}$	$\hat{\beta}_{ys.xs}\hat{\beta}_{sx.x}$
$N=10$	0.372	0.375	4.197	0.396	0.463	0.663	0.454	0.533	0.573
$N=50$	0.146	0.146	1.460	0.157	0.178	0.232	0.176	0.204	0.208

	$\lambda^2=0.01$			$\lambda^2=0.5$			$\lambda^2=1.000$		
Case 2	$\hat{\beta}_{ys.s}\hat{\beta}_{sx.x}$	$\hat{\beta}_{yx.x}$	$\hat{\beta}_{ys.xs}\hat{\beta}_{sx.x}$	$\hat{\beta}_{ys.s}\hat{\beta}_{sx.x}$	$\hat{\beta}_{yx.x}$	$\hat{\beta}_{ys.xs}\hat{\beta}_{sx.x}$	$\hat{\beta}_{ys.s}\hat{\beta}_{sx.x}$	$\hat{\beta}_{yx.x}$	$\hat{\beta}_{ys.xs}\hat{\beta}_{sx.x}$
$N=10$	4.619	4.631	4.620	0.690	0.813	0.746	0.546	0.655	0.720
$N=50$	1.548	1.555	1.548	0.234	0.265	0.245	0.187	0.217	0.221

Table 2: Simulation Results (Discrete Model)

Case 1	$\hat{\text{pr}}\{y_1\|\text{do}(X=x_1);\boldsymbol{s}\}$	$\hat{\text{pr}}\{y_1\|\text{do}(X=x_1)\}$	$\hat{\text{pr}}\{y_1\|\text{do}(X=x_1);x_0,\boldsymbol{s}\}$
$N=50$	0.074	0.101	0.105
$N=100$	0.053	0.071	0.073

Case 2	$\hat{\text{pr}}\{y_1\|\text{do}(X=x_1);\boldsymbol{s}\}$	$\hat{\text{pr}}\{y_1\|\text{do}(X=x_1)\}$	$\hat{\text{pr}}\{y_1\|\text{do}(X=x_1);x_0,\boldsymbol{s}\}$
$N=50$	0.086	0.101	0.263
$N=100$	0.061	0.071	0.176

these parameters are large. In order to verify the properties of the variances of the estimators of $\beta_{ys.s}\beta_{sx.x}$, $\beta_{yx.x}$ and $\beta_{ys\cdot x}\beta_{sx.x}$ through the ordinary least square method in Section 3.1, we did simulation experiments based on these settings in sample sizes $N=10$ and 50.

Table 1 reports the standard errors of $\hat{\beta}_{ys.s}\hat{\beta}_{sx.x}$, $\hat{\beta}_{yx.x}$ and $\hat{\beta}_{ys.xs}\hat{\beta}_{sx.x}$ from 5000 replications. From Table 1, for each case, the standard errors of $\hat{\beta}_{ys.s}\hat{\beta}_{sx.x}$ are smaller than those of $\hat{\beta}_{yx.x}$ and $\hat{\beta}_{ys\cdot xs}\hat{\beta}_{sx.x}$, which is consistent with the results. On the other hand, although $\hat{\beta}_{yx.x}$ provides better estimation accuracies than $\hat{\beta}_{ys\cdot xs}\hat{\beta}_{sx.x}$, the quantitative difference between them is smaller when the variance of $\epsilon_{s.x}$ is larger. In addition, when the variance of $\epsilon_{s.x}$ is small, the difference between the estimation accuracies of $\hat{\beta}_{ys.s}\hat{\beta}_{sx.x}$ and $\hat{\beta}_{yx.x}$ is small. However, when the variance of $\epsilon_{s.x}$ is large, the difference between the estimation accuracies of $\hat{\beta}_{ys.s}\hat{\beta}_{sx.x}$ and $\hat{\beta}_{yx.x}$ is large.

4.2 Discrete Model

In this section, we examine the inequality relationships between the estimation accuracies stated in Section 3.2. For dichotomous variables X, Y and S ($x \in \{x_0, x_1\}, y \in \{y_0, y_1\}, s \in \{s_0, s_1\}$), under the setting $\text{pr}(x_1)=0.5$, $\text{pr}(y_1|s_1)=0.6$ and $\text{pr}(y_1|s_0)=0.3$, when X is conditionally independent of Y given S and an empty set satisfies the back door criterion relative to (X,Y), as shown in Fig. 2, we consider the following two cases:

> Case 1 (Independence between X and S): $\text{pr}(s_1|x_1)=0.70$ and $\text{pr}(s_1|x_0)=0.70$,
>
> Case 2 (Dependence between X and S): $\text{pr}(s_1|x_1)=0.70$ and $\text{pr}(s_1|x_0)=0.10$.

In order to verify the properties of the variances of the estimators of $\text{pr}\{y_1|\text{do}(X=x_1)\}$, $\text{pr}\{y_1|\text{do}(X=x_1);\boldsymbol{s}\}$ and $\text{pr}\{y_1|\text{do}(X=x_1);x_0,\boldsymbol{s}\}$, we did simulation experiments based on these settings in sample sizes $N=50$ and 100.

Table 2 reports the standard errors of $\hat{\text{pr}}\{y_1|\text{do}(X=x_1)\}$, $\hat{\text{pr}}\{y_1|\text{do}(X=x_1);\boldsymbol{s}\}$ and $\hat{\text{pr}}\{y_1|\text{do}(X=x_1);x_0,\boldsymbol{s}\}$ from 5000 replication. From Table 2, for each case, the standard errors of $\hat{\text{pr}}\{y_1|\text{do}(X=x_1);\boldsymbol{s}\}$ are smaller than those of $\hat{\text{pr}}\{y_1|\text{do}(X=x_1)\}$ and $\hat{\text{pr}}\{y_1|\text{do}(X=x_1);x_0,s\}$, which is consistent with the results in Section 3.2. On the other hand, although $\hat{\text{pr}}\{y_1|\text{do}(X=x_1)\}$ provides better estimation accuracies than $\hat{\text{pr}}\{y_1|\text{do}(X=x_1);x_0,\boldsymbol{s}\}$, the quantitative difference between them is small when X is independent of S (Case 1) compared to the case where X is not independent of S (Case 2). In addition, the difference between the estimation accuracies of $\hat{\text{pr}}\{y_1|\text{do}(X=x_1);\boldsymbol{s}\}$ and $\hat{\text{pr}}\{y_1|\text{do}(X=x_1)\}$ is not large in Case 2, compared to Case 1.

5 Application

We illustrate our results by using the data from a study of setting up coating conditions for car bodies, reported by Okuno et al. (1986). According to Okuno et al. (1986), car bodies are coated in order to increase both the rust protection quality and the visual appearance, and a certain level of the coating thickness must be ensured in the process. Okuno et al. (1986) collected nonexperimental data in the coating process, in order to examine the process conditions and to increase the transfer efficiency. For details, refer to Okuno et al. (1986) and Kuroki (2012). The sample size is $N=38$ and the variables of our interest are Dilution ratio (X), Degree of viscosity (S)
Temperature (W_1), Degree of moisture (W_2)
Transfer efficiency (Y).
The sample correlation matrix is provided in Table 3. Since

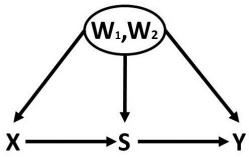

Fig. 3: Causal diagram of process analysis

Table 3: Correlation Matrix extracted from Kuroki (2012).

	X	S	W_1	W_2	Y
X	1.000	-0.678	0.145	-0.496	-0.198
S	-0.678	1.000	-0.509	0.684	0.463
W_1	0.145	-0.509	1.000	-0.571	-0.431
W_2	-0.496	0.684	-0.571	1.000	0.282
Y	-0.198	0.463	-0.431	0.282	1.000

it is assumed that $\{W_1, W_2\}$ satisfies the back door criterion relative to (X, Y) and the observed data is generated by the Gaussian linear structural equation model according to Kuroki (2012), $dE\{Y|\text{do}(X = x)\}/dx$ is identifiable and is given by $\beta_{yx.xw_1w_2}$. In addition, under the same assumptions as Kuroki (2012), the statistical t test for the no-partial correlation between X and Y given $\{S, W_1, W_2\}$ on Table 3 yields t-value $= 0.440$ with 33 degrees of freedom, which gives a p-value of 0.669. Thus, it would be reasonable to assume that X is conditionally independent of Y given $\{S, W_1, W_2\}$ as shown in Fig.3 and $dE\{Y|\text{do}(X = x)\}/dx$ can be estimated by $\hat{\beta}_{yx.xw_1w_2}$, $\hat{\beta}_{ys.sw_1w_2}\hat{\beta}_{sx.xw_1w_2}$ and $\hat{\beta}_{ys.xsw_1w_2}\hat{\beta}_{sx.xw_1w_2}$. Then, the standard errors estimated using the bootstrap methods with 5000 replicates are var$(\hat{\beta}_{yx.xw_1w_2}) = 0.180$, var$(\hat{\beta}_{ys.xsw_1w_2}\hat{\beta}_{sx.xw_1w_2}) = 0.145$ and var$(\hat{\beta}_{ys.sw_1w_2} \times\hat{\beta}_{sx.xw_1w_2}) = 0.111$, which is consistent with the results in Section 3.1. In addition, var$(\hat{\beta}_{yx.xw_1w_2})$ is larger than var$(\hat{\beta}_{ys.xsw_1w_2}\hat{\beta}_{sx.xw_1w_2})$ in this case study. Thus, the results in this case study show that it would be better to use supplementary variables $\{W_1, W_2, S\}$ to improve the estimation accuracy of the causal effect.

6 Discussion

The reliable evaluation of causal effects gains increasing interest in practical science. In many situations, not only a treatment and a response are measured, but also some supplementary variables are measured. This paper discussed the role of supplementary variables in the estimation problem of causal effects, and showed that if a treatment is associated with a response variable through some supplementary variables then supplementary variables enable us to improve the estimation accuracies without amplifying the bias related to the target quantities. Noting that the proposed assumption of the conditional independence can be tested from observed data under the given distributional assumption and thus it is not always necessary to have prior knowledge that the treatment has no direct effect on the response, the results of the paper have a practical advantage in the sense that the estimation accuracy can be improved by using supplementary variables if the assumption is statistically affirmative . On the other hand, if we have such prior knowledge, the results of this paper help us to judge from graph structures under what situation the estimation accuracy of the causal effect can be improved by supplementary variables, and to obtain the reliable evaluation of causal effects from observed data.

This paper assumed that a treatment is conditionally independent of a response given a set of supplementary variables from prior knowledge. Even when we know that such an assumption holds from prior knowledge, if the variances of the causal effects do not exist in linear models and zero frequencies are included in discrete models because of small sample size, it may be difficult to know whether or not our results hold. On the other hand, if we do not know whether or not such an assumption holds from prior knowledge, the conditional independence should be checked based on statistical hypothetical tests and thus it would be required to construct test statistics based on small sample size, which is a future work. In addition, the discussion of this paper is related to the z-identifiability problem which occurs when the researcher aims to identify causal effects under a situation where observed data might not be enough but randomized experiments over supplementary variables are available (Bareinboim and Pearl, 2012). We did not discuss the estimation problem based on the z-identifiability conditions, which would be another future work.

Appendices

Appendix I

For $\hat{\beta}_{yx\cdot xw}$ and $\hat{B}_{ys\cdot sw}\hat{B}_{sx\cdot xw}$, we have

$$\text{var}(\hat{\beta}_{yx\cdot xw}) = \sigma_{yy\cdot xw}E\left(S^{-1}_{xx\cdot w}\right)$$

and

$$\begin{aligned}\text{var}(\hat{B}_{ys\cdot sw}\hat{B}_{sx\cdot xw}) = \sigma_{yy\cdot sw}E\left(\hat{B}'_{sx\cdot xw}S^{-1}_{ss\cdot w}\hat{B}_{sx\cdot xw}\right)\\ +B_{ys\cdot sw}\Sigma_{ss\cdot xw}B'_{ys\cdot sw}E\left(S^{-1}_{xx\cdot w}\right),\end{aligned}$$

where $S_{ss\cdot w}$ is the sum of squared matrix of $\boldsymbol{S}$ given $\boldsymbol{W}$ and similar notation is used for other matrices. Here, let $\Sigma_{ss.xw}$ be a conditional covariance matrix of $\boldsymbol{S}$ given $\{X\} \cup \boldsymbol{W}$. Then, noting that both $\sigma_{yy.xw} = \sigma_{yy.sw} + B_{ys.sw}\Sigma_{ss.xw}B'_{ys.sw}$ and $\beta_{yx\cdot xw} = B_{ys.sw}B_{sx.xw}$ hold, from $\hat{B}_{sx\cdot xw} = S_{sx.w}S^{-1}_{xx.w}$ and $S_{xx\cdot sw} = S_{xx\cdot w} - S_{xs\cdot w}S^{-1}_{ss\cdot w}S_{sx\cdot w}$, we have

$$\begin{aligned}&\text{var}(\hat{\beta}_{yx\cdot xw}) - \text{var}(\hat{B}_{ys\cdot sw}\hat{B}_{sx\cdot xw})\\ &= \sigma_{yy\cdot sw}E\left(S^{-1}_{xx\cdot w} - \hat{B}'_{sx\cdot xw}S^{-1}_{ss\cdot w}\hat{B}_{sx\cdot xw}\right)\\ &= \sigma_{yy\cdot sw}E\left\{S^{-2}_{xx\cdot w}(S_{xx\cdot w} - S_{xs\cdot w}S^{-1}_{ss\cdot w}S_{sx\cdot w})\right\}\\ &= \sigma_{yy\cdot sw}E\left(S^{-2}_{xx\cdot w}S_{xx\cdot sw}\right),\end{aligned}$$

which shows that var$(\hat{\beta}_{yx\cdot xw}) -$ var$(\hat{B}_{ys\cdot sw}\hat{B}_{sx\cdot xw})$ is non-negative.

Similarly, for $\hat{B}_{ys\cdot xsw}\hat{B}_{sx\cdot xw}$, we have

$$\begin{aligned}\mathrm{var}(\hat{B}_{ys\cdot xsw}\hat{B}_{sx\cdot xw}) &= \sigma_{yy\cdot sw}E\left(\hat{B}_{sx\cdot xw}S^{-1}_{ss\cdot xw}\hat{B}'_{sx\cdot xw}\right)\\ &+B_{ys\cdot sw}\Sigma_{ss\cdot xw}B'_{ys\cdot sw}E\left(S^{-1}_{xx\cdot w}\right).\end{aligned}$$

Thus, noting that $S_{ss\cdot w}-S_{ss\cdot xw}$ is a positive semi-definite matrix, we have

$$\begin{aligned}&\mathrm{var}(\hat{B}_{ys\cdot xsw}\hat{B}_{sx\cdot xw})-\mathrm{var}(\hat{B}_{ys\cdot sw}\hat{B}_{sx\cdot xw})\\ &= \sigma_{yy\cdot sw}E\left\{\hat{B}_{sx\cdot xw}(S^{-1}_{ss\cdot xw}-S^{-1}_{ss\cdot w})\hat{B}'_{sx\cdot xw}\right\}\geq 0.\end{aligned}$$

Appendix II

For a.var[p̂r$\{y|\mathrm{do}(X=x)\}$] and a.var[p̂r$\{y|\mathrm{do}(X=x);\boldsymbol{s}\}$], let $\mathrm{pr}(y_0|x,\boldsymbol{w})=1-\mathrm{pr}(y_1|x,\boldsymbol{w})$, $\mathrm{pr}(y_0|\boldsymbol{s},\boldsymbol{w})=1-\mathrm{pr}(y_1|\boldsymbol{s},\boldsymbol{w})$ and $\mathrm{pr}(y_0|x',\boldsymbol{s},\boldsymbol{w})=1-\mathrm{pr}(y_1|x',\boldsymbol{s},\boldsymbol{w})$. In addition, if the denominators of estimated conditional probabilities are zero, then they are considered as zero in this paper. From the variance basic formula, we have

$$\begin{aligned}&\mathrm{var}\left\{\sum_w \hat{\mathrm{pr}}(y_1|x,\boldsymbol{w})\hat{\mathrm{pr}}(\boldsymbol{w})\right\}\\ &-\mathrm{var}\left\{\sum_{s,w}\hat{\mathrm{pr}}(y_1|\boldsymbol{s},\boldsymbol{w})\hat{\mathrm{pr}}(\boldsymbol{s}|x,\boldsymbol{w})\hat{\mathrm{pr}}(\boldsymbol{w})\right\}\\ &=\sum_w \mathrm{pr}(y_1|x,\boldsymbol{w})\mathrm{pr}(y_0|x,\boldsymbol{w})E\left\{\frac{\hat{\mathrm{pr}}(\boldsymbol{w})^2}{n_{xw}}\right\}\\ &+\mathrm{var}\left\{\sum_w \mathrm{pr}(y_1|x,\boldsymbol{w})\hat{\mathrm{pr}}(\boldsymbol{w})\right\}\\ &-\mathrm{var}\left\{\sum_{s,w}\mathrm{pr}(y_1|\boldsymbol{s},\boldsymbol{w})\hat{\mathrm{pr}}(\boldsymbol{s}|x,\boldsymbol{w})\hat{\mathrm{pr}}(\boldsymbol{w})\right\}\\ &-\sum_{s,w}\mathrm{pr}(y_1|\boldsymbol{s},\boldsymbol{w})\mathrm{pr}(y_0|\boldsymbol{s},\boldsymbol{w})E\left\{\frac{\hat{\mathrm{pr}}(\boldsymbol{s}|x,\boldsymbol{w})^2\hat{\mathrm{pr}}(\boldsymbol{w})^2}{n_{sw}}\right\}\\ &\simeq\sum_w \mathrm{pr}(y_1|x,\boldsymbol{w})\mathrm{pr}(y_0|x,\boldsymbol{w})E\left\{\frac{\hat{\mathrm{pr}}(\boldsymbol{w})^2}{n_{xw}}\right\}\\ &+\mathrm{var}\left\{\sum_w \mathrm{pr}(y_1|x,\boldsymbol{w})\hat{\mathrm{pr}}(\boldsymbol{w})\right\}\\ &-\mathrm{var}\left\{\sum_{s,w}\mathrm{pr}(y_1|\boldsymbol{s},\boldsymbol{w})\mathrm{pr}(\boldsymbol{s}|x,\boldsymbol{w})\hat{\mathrm{pr}}(\boldsymbol{w})\right\}\\ &-E\left[\mathrm{var}\left\{\sum_{s,w}\mathrm{pr}(y_1|\boldsymbol{s},\boldsymbol{w})\hat{\mathrm{pr}}(\boldsymbol{s}|x,\boldsymbol{w})\hat{\mathrm{pr}}(\boldsymbol{w})\right\}\right]\\ &-\sum_{s,w}\mathrm{pr}(y_1|\boldsymbol{s},\boldsymbol{w})\mathrm{pr}(y_0|\boldsymbol{s},\boldsymbol{w})E\left\{\frac{\hat{\mathrm{pr}}(\boldsymbol{s}|x,\boldsymbol{w})^2\hat{\mathrm{pr}}(\boldsymbol{w})^2}{n_{sw}}\right\}\\ &=\sum_w \mathrm{pr}(y_1|x,\boldsymbol{w})\mathrm{pr}(y_0|x,\boldsymbol{w})E\left\{\frac{\hat{\mathrm{pr}}(\boldsymbol{w})^2}{n_{xw}}\right\}\\ &-\sum_{s,w}\mathrm{pr}(y_1|\boldsymbol{s},\boldsymbol{w})\mathrm{pr}(y_0|\boldsymbol{s},\boldsymbol{w})E\left\{\frac{\hat{\mathrm{pr}}(\boldsymbol{s}|x,\boldsymbol{w})^2\hat{\mathrm{pr}}(\boldsymbol{w})^2}{n_{sw}}\right\}\end{aligned}$$

$$\begin{aligned}&-\sum_{s,w}\mathrm{pr}(y_1|\boldsymbol{s},\boldsymbol{w})^2\mathrm{pr}(\boldsymbol{s}|x,\boldsymbol{w})\\ &\times(1-\mathrm{pr}(\boldsymbol{s}|x,\boldsymbol{w}))E\left\{\frac{\hat{\mathrm{pr}}(\boldsymbol{w})^2}{n_{xw}}\right\}\\ &+\sum_{s\neq s',w}\mathrm{pr}(y_1|\boldsymbol{s},\boldsymbol{w})\mathrm{pr}(y_1|s',\boldsymbol{w})\mathrm{pr}(\boldsymbol{s}|x,\boldsymbol{w})\\ &\times\mathrm{pr}(s'|x,\boldsymbol{w})E\left\{\frac{\hat{\mathrm{pr}}(\boldsymbol{w})^2}{n_{xw}}\right\}\\ &=\sum_w \mathrm{pr}(y_1|x,\boldsymbol{w})\mathrm{pr}(y_0|x,\boldsymbol{w})E\left\{\frac{\hat{\mathrm{pr}}(\boldsymbol{w})^2}{n_{xw}}\right\}\\ &-\sum_{s,w}\mathrm{pr}(y_1|\boldsymbol{s},\boldsymbol{w})\mathrm{pr}(y_0|\boldsymbol{s},\boldsymbol{w})E\left\{\frac{\hat{\mathrm{pr}}(\boldsymbol{s}|x,\boldsymbol{w})^2\hat{\mathrm{pr}}(\boldsymbol{w})^2}{n_{sw}}\right\}\\ &-\sum_{s,w}\mathrm{pr}(y_1|\boldsymbol{s},\boldsymbol{w})^2\mathrm{pr}(\boldsymbol{s}|x,\boldsymbol{w})E\left\{\frac{\hat{\mathrm{pr}}(\boldsymbol{w})^2}{n_{xw}}\right\}\\ &+\sum_w \mathrm{pr}(y_1|x,\boldsymbol{w})^2E\left\{\frac{\hat{\mathrm{pr}}(\boldsymbol{w})^2}{n_{xw}}\right\}\\ &=\sum_w \mathrm{pr}(y_1|x,\boldsymbol{w})E\left\{\frac{\hat{\mathrm{pr}}(\boldsymbol{w})^2}{n_{xw}}\right\}\\ &-\sum_{s,w}\mathrm{pr}(y_1|\boldsymbol{s},\boldsymbol{w})\mathrm{pr}(y_0|\boldsymbol{s},\boldsymbol{w})E\left\{\frac{\hat{\mathrm{pr}}(\boldsymbol{s}|x,\boldsymbol{w})^2\hat{\mathrm{pr}}(\boldsymbol{w})^2}{n_{sw}}\right\}\\ &-\sum_{s,w}\mathrm{pr}(y_1|\boldsymbol{s},\boldsymbol{w})^2\mathrm{pr}(\boldsymbol{s}|x,\boldsymbol{w})E\left\{\frac{\hat{\mathrm{pr}}(\boldsymbol{w})^2}{n_{xw}}\right\}\\ &=\sum_{s,w}\mathrm{pr}(y_1|\boldsymbol{s},\boldsymbol{w})\mathrm{pr}(y_0|\boldsymbol{s},\boldsymbol{w})\mathrm{pr}(\boldsymbol{s}|x,\boldsymbol{w})E\left\{\frac{\hat{\mathrm{pr}}(\boldsymbol{w})^2}{n_{xw}}\right\}\\ &-\sum_{s,w}\mathrm{pr}(y_1|\boldsymbol{s},\boldsymbol{w})\mathrm{pr}(y_0|\boldsymbol{s},\boldsymbol{w})E\left\{\frac{\hat{\mathrm{pr}}(\boldsymbol{s}|x,\boldsymbol{w})^2\hat{\mathrm{pr}}(\boldsymbol{w})^2}{n_{sw}}\right\}\\ &=\sum_w\left[\sum_s \mathrm{pr}(y_1|\boldsymbol{s},\boldsymbol{w})\mathrm{pr}(y_0|\boldsymbol{s},\boldsymbol{w})\right.\\ &\left.\times\left[\mathrm{pr}(\boldsymbol{s}|x,\boldsymbol{w})E\left\{\frac{\hat{\mathrm{pr}}(\boldsymbol{w})^2}{n_{xw}}\right\}-E\left\{\frac{\hat{\mathrm{pr}}(\boldsymbol{s}|x,\boldsymbol{w})^2\hat{\mathrm{pr}}(\boldsymbol{w})^2}{n_{sw}}\right\}\right]\right]\\ &=\sum_{s,w}\mathrm{pr}(y_1|\boldsymbol{s},\boldsymbol{w})\mathrm{pr}(y_0|\boldsymbol{s},\boldsymbol{w})\\ &\times E\left\{\frac{n_{sw}-n_{xsw}}{n_{xw}n_{sw}}\hat{\mathrm{pr}}(\boldsymbol{s}|x,\boldsymbol{w})\hat{\mathrm{pr}}(\boldsymbol{w})^2\right\}\geq 0,\end{aligned}$$

which shows that

$$\text{a.var}\left[\hat{\mathrm{pr}}\{y|\mathrm{do}(X=x)\}\right]\geq\text{a.var}\left[\hat{\mathrm{pr}}\{y|\mathrm{do}(X=x);\boldsymbol{s}\}\right]$$

holds.

Similarly, for a.var[p̂r$\{y|\mathrm{do}(X=x);x',\boldsymbol{s}\}$] and a.var[p̂r$\{y|\ \mathrm{do}(X=x);\ \boldsymbol{s}\}$], we have

$$\begin{aligned}&\mathrm{var}\left\{\sum_{s,w}\hat{\mathrm{pr}}(y_1|x',\boldsymbol{s},\boldsymbol{w})\hat{\mathrm{pr}}(\boldsymbol{s}|x,\boldsymbol{w})\hat{\mathrm{pr}}(\boldsymbol{w})\right\}\\ &-\mathrm{var}\left\{\sum_{s,w}\hat{\mathrm{pr}}(y_1|\boldsymbol{s},\boldsymbol{w})\hat{\mathrm{pr}}(\boldsymbol{s}|x,\boldsymbol{w})\hat{\mathrm{pr}}(\boldsymbol{w})\right\}\end{aligned}$$

$$
\begin{aligned}
&\simeq \sum_{s,w} \mathrm{pr}(y_1|x', \boldsymbol{s}, \boldsymbol{w})\mathrm{pr}(y_0|x', \boldsymbol{s}, \boldsymbol{w}) \\
&\times E\left\{\frac{\hat{\mathrm{pr}}(\boldsymbol{s}|x, \boldsymbol{w})^2\hat{\mathrm{pr}}(\boldsymbol{w})^2}{n_{x'sw}}\right\} \\
&- \sum_{s,w} \mathrm{pr}(y_1|\boldsymbol{s}, \boldsymbol{w})\mathrm{pr}(y_0|\boldsymbol{s}, \boldsymbol{w}) E\left\{\frac{\hat{\mathrm{pr}}(\boldsymbol{s}|x, \boldsymbol{w})^2\hat{\mathrm{pr}}(\boldsymbol{w})^2}{n_{sw}}\right\} \\
&= \sum_{s,w} \mathrm{pr}(y_1|\boldsymbol{s}, \boldsymbol{w})\mathrm{pr}(y_0|\boldsymbol{s}, \boldsymbol{w}) \\
&\times E\left\{\frac{\hat{\mathrm{pr}}(\boldsymbol{s}|x, \boldsymbol{w})^2\hat{\mathrm{pr}}(\boldsymbol{w})^2}{n_{x'sw}} - \frac{\hat{\mathrm{pr}}(\boldsymbol{s}|x, \boldsymbol{w})^2\hat{\mathrm{pr}}(\boldsymbol{w})^2}{n_{sw}}\right\} \geq 0,
\end{aligned}
$$

which shows that

$$\text{a.var}\left[\hat{\mathrm{pr}}\{y|\mathrm{do}(X = x); x', \boldsymbol{s}\}\right] \geq \text{a.var}\left[\hat{\mathrm{pr}}\{y|\mathrm{do}(X = x); \boldsymbol{s}\}\right]$$

holds.

Acknowledgements

We thank three reviewers whose comments significantly improved the presentation of the paper. This paper was partially supported by the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, **91**, 444–455.

Bareinboim, E., Brito, C. and Pearl, J. (2011). Local Characterizations of Causal Bayesian Networks. *Proceedings of the GKR-22nd International Joint Conference on Artificial Intelligence*, 1-17.

Bareinboim, E. and Pearl, J. (2012). Causal inference by surrogate experiments: z-identifiability. *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, 113–120.

Baron, R. M. and Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology*, **51**, 1173-1182.

Bowden, R. J. and Turkington, D. A. (1984). *Instrumental Variables*. Cambridge University Press.

Cai, Z. and Kuroki, M. (2008). On identifying total effects in the presence of latent variables and selection bias. *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, 62-69.

Cox, D. R. (1960). Regression analysis when there is prior information about supplementary variables. *Journal of the Royal Statistical Society. Series B*, **22**, 172–176.

Hui, H. and Zhongguo, Z. (2008). Comparing identifiability criteria for causal effects in Gaussian causal models (In Chinese). *Acta Mathematica Scientia A*, **28**, 808–817.

Imai, K., Keele, L., Tingley, D. and Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review*, **105**, 765–789.

Kuroki, M. (2000). Selection of post-treatment variables for estimating total effect from empirical research. *Journal of the Japanese Statistical Society*, **30**, 129-142.

Kuroki, M. (2007). Graphical identifiability criteria for causal effects in studies with an unobserved treatment/response variable. *Biometrika*, **94**, 37–47.

Kuroki, M. (2012). Optimizing a control plan using causal diagram with an application to statistical process analysis. *Journal of Applied Statistics*, **39**, 673-694.

Kuroki, M. and Cai, Z. (2004). Selection of identifiability criteria for total effects by using path diagrams. *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, 333–340.

Kuroki, M. and Pearl, J. (2013). Measurement bias and effect restoration in causal inference. *Biometrika*, Accepted.

Lauritzen, S. L. (2001). Causal inference from graphical models. *Complex Stochastic Systems*. Chapman and Hall/CRC, 63–107.

MacKinnon, D. P. (2008). *Introduction to Statistical Mediation Analysis*. Erlbaum.

Manski, C. F. (2007). *Identification for Prediction and Decision*. Harvard University Press.

Okuno, T., Katayama, Z., Kamigori, N., Itoh, T., Irikura, N. and Fujiwara, N. (1986). *Multivariate Data Analysis in Industry* (In Japanese). JUSE Press.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*, Morgan-Kaufmann.

Pearl, J. (2001). Direct and indirect effects. *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*, 411–420.

Pearl, J. (2009). *Causality: Models of Reasoning and Inference, The 2nd Edition*. Cambridge University Press.

Pearl, J. (2012). Some thoughts concerning transfer learning, with applications to meta-analysis and data-sharing estimation. UCLA Cognitive Systems Laboratory, Technical Report (R-387).

Ramsahai, R. R. (2012). Supplementary variables for causal estimation. *Causality: Statistical Perspectives and Applications*. John Wiley and Sons, 218–233.