Journal of Mathematical Research & Exposition
Vol.24, No.3, 381–387, August, 2004

Bayes Estimation of Causal Effect for a Counterfactural Model *

XU Jing, ZHENG Zhong-guo
(Dept. of Probability and Statistics, Peking University, Beijing 100871, China)

Abstract: This paper presents the Bayes estimation and empirical Bayes estimation of causal effects in a counterfactual model. It also gives three kinds of prior distribution of the assumptions of replaceability. The experiment shows that empirical Bayes estimation is better than other estimations when not knowing which assumption is true.

Key words: Bayes estimation; causal effect; counterfactual model; intervention; replaceability.

Classification: AMS(2000) 62C12, 62F07/CLC number: O212.1

Document code: A **Article ID**: 1000-341X(2004)03-0381-07

1. Introduction

Causality plays an important role in modern medical, behavioral, social and biological sciences. The central aim of studies in these fields is to elucidate the causal effects among variables, only through which the effects of some actions or strategies can be predicted. The possibility of learning causal relationship from raw data entered the realm of formal treatment and feasible computation in the mid-1980s when the mathematical relationship between graphs and probabilistic dependency came into light. The counterfactual causal model given by Rubin[1] and causal graph model given by Pearl[2,3] constitute the framework of causality. Pearl[4] described that causal relationships can be inferred from nontemporal statistical data if one makes certain assumptions about the underlying process of data generation. As we know, causal relationship cannot be completely determined by the correlation. So it leads to the identification problem in the analysis of causal effects. If the causal effects are identifiable, we can computer them from the observed data. Zheng et al.[5] investigated the identifiability of causal effects of a control variable on a resulting variable in a simple counterfactual model, and presented three assumptions of replaceability and the formulas of computing causal effects. However, any causal assumption cannot be realized by imposing statistical assumption, that is, we cannot determine

*Received date: 2002-10-07
Foundation item: Supported by NNSFC (39930160)
Biography: XU Jing (1977-), female, Ph.D.

which assumption is suitable for a causal model. Therefore it is necessary to look for good estimations of causal effects.

2. Causal model in this paper

We call $M = \langle G, \Theta_G \rangle$ a causal model where G is a directed acyclic graph (DAG) over a set V of variables and Θ_G is the set of probability parameters. Causal model can predict the effect that any external or spontaneous changes have on the distributions.

In this paper, let $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$ denote the domains of X, Y, Z respectively, where $\mathcal{X} = \mathcal{Y} = \mathcal{Z} = \{0, 1\}$ and X is a control variable, Y is an outcome or response variable, and Z is a covariant variable. The causal graph DAG G is given in Fig.1.

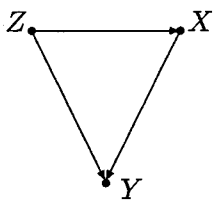

Fig.1 DAG G

Since the joint distribution P of (X, Y, Z) can be estimated from the observed data, we assume that the distribution of (X, Y, Z) is known and the compatibility of P with DAG G in Fig.1 implies that

$$P(x, y, z) = P(z)P(x|z)P(y|x, z).$$

In Pearl's framework, an intervention is to force a subset X of V to be fixed values x, written by $do(X = x)$, thus defining a new distribution over the remaining variables that characterizes the effects of the intervention. Here, we do intervention to control variable X and investigate the causal effects of X on Y. Let variable Y_0 also be binary with domain $\{0, 1\}$ and identical to Y when $do(X = 0)$, so we use $P(Y_0)$ to represent the distribution of Y under the intervention $do(X = 0)$.

Definition 1 *$P(Y_0)$ is called the causal effect of $X = 0$ on Y or causal distribution of Y when $do(X = 0)$.*

Because the method to discuss intervention $do(X = 1)$ is as same as $do(X = 0)$, we only investigate the case of $do(X = 0)$.

3. Identifiability of causal effect

We adopt the following notation to simplify presentation

$$c = P(Z = 1),\ a_0 = P(X = 1|Z = 0),\ a_1 = P(X = 1|Z = 1),$$

$$b_{ij} = P(Y = 1|X = i, Z = j),\ i, j = 0, 1.$$

Then

$$P(Z = 0) = 1 - c \triangleq \bar{c}, \quad P(X = 0|Z = 0) = 1 - a_0 \triangleq \overline{a_0}, \quad P(X = 0|Z = 1) = 1 - a_1 \triangleq \overline{a_1}.$$

Using the same methods as [5], we derive four assumptions of replaceability for the joint distribution P of (X, Y, Z, Y_0), and only under one of them could the causal effect of $X = 0$ on Y be identifiable. This distribution can be expressed by parameters $(c, a_0, a_1, b_{00}, b_{01}, b_{10}, b_{11}, u_{10}, u_{11})$, where $u_{10} = P(Y_0 = 1|X = 1, Z = 0)$ and $u_{11} = P(Y_0 = 1|X = 1, Z = 1)$ are unknown. Our purpose is to know the effect of X on Y, that is to know the value $P(Y = 1|X = 1) - P(Y_0 = 1|X = 1)$ which reflects how much the control variable affects the response variable. $P(Y_0 = 1|X = 1)$ implies the counterfactual probability of $Y = 1$ if we force $X = 0$ given $X = 1$ and it is different from $P(Y = 1|X = 1)$ which means the conditional probability of $Y = 1$ given $X = 1$ and can be computed from data but $P(Y_0 = 1|X = 1)$ cannot. Obviously, the more the difference between $P(Y = 1|X = 1)$ and $P(Y_0 = 1|X = 1)$, the more the effect of X on Y. The difference can be derived from the identifiability of $P(Y_0 = 1)$, for

$$\begin{aligned} P(Y_0 = 1|X = 1) &= \sum_{j=0}^{1} P(Y_0 = 1|X = 1, Z = j)P(Z = j|X = 1) \\ &= \sum_{j=0}^{1} P(Y_0 = 1|X = 1, Z = j)\frac{P(X = 1|Z = j)P(Z = j)}{\sum_{k=0}^{1} P(X = 1|Z = k)P(Z = k)} \\ &= \frac{u_{10}a_0\bar{c} + u_{11}a_1c}{a_0\bar{c} + a_1c}. \end{aligned}$$

Therefore, the identifiability of $P(Y_0 = 1|X = 1)$ is equivalent to the identifiability of $P(Y_0 = 1)$ and

$$P(Y_0 = 1) = b_{00}\overline{a_0}\bar{c} + b_{01}\overline{a_1}c + u_{10}a_0\bar{c} + u_{11}a_1c.$$

We present this result by a theorem.

Theorem 1 *Suppose the joint distribution P of (X, Y, Z, Y_0) satisfies*

(H1) $X \perp Y_0$, *or*

(H2) $X \perp Y_0|Z$, *or*

(H3) $X \perp Y_0|Z = 0$, $Y_0 \perp Z|X = 1$, *or*

(H4) $X \perp Y_0|Z = 1$, $Y_0 \perp Z|X = 1$.

Then $P(Y_0 = 1)$ is identifiable and has the following form:

$$P(Y_0 = 1) = \begin{cases} \frac{b_{00}\overline{a_0}\bar{c} + b_{01}\overline{a_1}c}{\overline{a_0}\bar{c} + \overline{a_1}c} & \text{if } (X \perp Y_0)_P; \\ b_{00}\bar{c} + b_{01}c & \text{if } (X \perp Y_0|Z)_P; \\ b_{00}\bar{c} + b_{01}\overline{a_1}c + b_{00}a_1c & \text{if } (X \perp Y_0|Z = 0)_P \cap (Y_0 \perp Z|X = 1)_P; \\ b_{00}\overline{a_0}\bar{c} + b_{01}a_0\bar{c} + b_{01}c & \text{if } (X \perp Y_0|Z = 1)_P \cap (Y_0 \perp Z|X = 1)_P. \end{cases} \quad (1)$$

Here $\perp$ means independence between two variables.

The proof of this theorem is similar to that of Theorem 1 in [5].

Definition 2 *The assumptions for P given by (2)*

$$\begin{aligned} (X \perp Y_0) \cup (X \perp Y_0|Z) \cup ((X \perp Y_0|Z = 0) \cap (Y_0 \perp Z|X = 1)) \cup \\ ((X \perp Y_0|Z = 1) \cap (Y_0 \perp Z|X = 1)) \end{aligned} \quad (2)$$

is called the replaceable assumptions.

Though the causal effect is identifiable under those assumptions, we cannot determine which assumption is suitable in the model from observed data. Therefore, in a stricter way, the conclusion in Theorem 1 cannot be called identifiable results. Generally people choose one according to experience or objective situation. In the next section, we show how to estimate $P(Y_0 = 1)$ from observed data.

4. Bayes estimation of causal effect

For a causal model $\langle G, \Theta_G \rangle$, our focus is to estimate $P(Y_0 = 1)$ given the data. In order to do this we must suppose that $P(Y_0 = 1)$ is identifiable. Let θ be the parameter of observational variable (X, Y, Z) and $H = i$ represents assumption (Hi) in Theorem 1, $i = 1, 2, 3, 4$. Because we want to use Bayes method, the distribution parameter θ and the assumption H are all random variables. Therefore the variables corresponding to our problem are $(X, Y, Z, Y_0, \theta, H)$ and

$$P(x, y, z, y_0, \theta, i) = P(x, y, z, y_0|\theta, i)h(\theta, i) = P(x, y, z, y_0|\theta)h(\theta, i), \tag{3}$$

where $h(\theta, i)$ is the join distribution of (θ, H). Because (X, Y, Z, Y_0) is independent of the replaceable assumption H, $P(x, y, z, y_0|\theta, i) = P(x, y, z, y_0|\theta)$ in (3).

At first we consider a simple case which has another assumption — θ is independent of H, that is

$$h(\theta, i) = h(\theta)\pi_i,$$

where $h(\theta)$ is the prior distribution of θ and π_i is the prior distribution of H. Let

$$h(\theta) = h_\alpha(\theta),$$

where α is the unknown parameter in the distribution $h_\alpha(\theta)$, the domain of α is identical to that of θ, and satisfies

$$P(\theta, \alpha) = \begin{cases} 1 & \text{if } \theta = \alpha; \\ 0 & \text{if } \theta \neq \alpha. \end{cases} \tag{4}$$

Equation (4) shows that $h_\alpha(\theta)$ is a discrete distribution and the probability of $\theta = \alpha$ is 1.

Let $\delta = \delta(X, Y, Z)$ be an estimation of $P(Y_0 = 1)$ under $H = i$. Its risk function with square loss is

$$\begin{aligned} R(\theta, i, \delta) &= \int_{\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}\times\mathcal{Y}} (\delta(X, Y, Z) - P(Y_0 = 1))^2 \mathrm{d}P(X, Y, Z, Y_0) \\ &= \int_{\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}} (\delta(X, Y, Z) - P_i(\theta))^2 \mathrm{d}P_\theta(X, Y, Z), \end{aligned}$$

where $P_i(\theta)$ equals to $P(Y_0 = 1)$ under the assumption (Hi) in (1), $i = 1, 2, 3, 4$. For a

fixed α the average risk of δ is

$$\begin{aligned} R(\delta) &= \sum_{i=1}^{4} \int_{\Theta} R(\theta, i, \delta) h_{\alpha}(\theta, i) \mathrm{d}\theta \\ &= \sum_{i=1}^{4} \int_{\Theta} R(\theta, i, \delta) h_{\alpha}(\theta) \pi_i \mathrm{d}\theta = \sum_{i=1}^{4} R(\alpha, i, \delta) \pi_i \\ &= \sum_{i=1}^{4} \int_{\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}} (\delta(X, Y, Z) - P_i(\alpha))^2 \mathrm{d}P_{\alpha}(X, Y, Z) \pi_i \\ &= \int_{\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}} \sum_{i=1}^{4} (\delta(X, Y, Z) - P_i(\alpha))^2 \pi_i \mathrm{d}P_{\alpha}(X, Y, Z). \end{aligned} \tag{5}$$

The Bayes estimation δ_{α} of $P(Y_0 = 1)$ related to the prior distribution $h_{\alpha}(\theta)\pi_i$ is the estimation which makes $R(\delta)$ be minimum, so it is

$$\delta_{\alpha} = \sum_{i=1}^{4} P_i(\alpha) \pi_i. \tag{6}$$

We know that δ_{α} does not depend on data, but on α. Because α is a parameter, it can be estimated from data. Putting the estimation $\hat{\alpha}$ of α into (6), we obtain

$$\hat{\delta} = \sum_{i=1}^{4} P_i(\hat{\alpha}) \pi_i, \tag{7}$$

and this is empirical Bayes estimation of $P(Y_0 = 1)$. The average risk of empirical Bayes estimation is

$$R(\hat{\delta}) = \int_{\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}} \sum_{i=1}^{4} (\hat{\delta}(X, Y, Z) - P_i(\alpha))^2 \pi_i \mathrm{d}P_{\alpha}(X, Y, Z). \tag{8}$$

On the other hand, let $\tilde{\delta}$ be another estimation of $P(Y_0 = 1)$. Then its average risk is

$$R(\tilde{\delta}) = \int_{\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}} \sum_{i=1}^{4} (\tilde{\delta}(X, Y, Z) - P_i(\alpha))^2 \pi_i \mathrm{d}P_{\alpha}(X, Y, Z).$$

In Section 5, we perform experiments to compare $\hat{\delta}$ and $\tilde{\delta}$ by comparing their average risks.

We have discussed the case that H is independent of θ. Now we consider the other case, that is, H is not independent of θ

$$h(\theta, i) = h(\theta) \pi_i(\theta).$$

Let $\delta = \delta(X, Y, Z)$ be an estimation of $P(Y_0 = 1)$. Its average risk function with square loss is

$$R(\theta, i, \delta) = \int_{\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}} (\delta(X, Y, Z) - P_i(\theta))^2 \mathrm{d}P_{\theta}(X, Y, Z),$$

and so the average risk of δ is

$$R(\delta) = \sum_{i=1}^{4} \int_{\Theta} \int_{\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}} (\delta(X,Y,Z) - P_i(\theta))^2 h(\theta)\pi_i(\theta) \mathrm{d}P_\theta(X,Y,Z)\mathrm{d}\theta.$$

As discussed above, let $h(\theta) = h_\alpha(\theta)$. Then the Bayes estimation of $P(Y_0 = 1)$ is

$$\delta_\alpha = \sum_{i=1}^{4} P_i(\alpha)\pi_i(\alpha),$$

and empirical Bayes estimation is

$$\hat{\delta} = \sum_{i=1}^{4} P_i(\hat{\alpha})\pi_i(\hat{\alpha}),$$

where $\hat{\alpha}$ is the estimation of α.

As mentioned in Section 2, the four assumptions of replaceability are not testable from the observed data. Now we give three kinds of choice of the prior distribution of H.

(Choice 1) H has union distribution, that is

$$\pi_i = 1/4, \quad i = 1,2,3,4. \tag{9}$$

(Choice 2) Firstly, compute every variance of $P_i(X,Y,Z)$ which is the estimation of $P_i(\alpha)$ under the assumption (Hi), i=1,2,3,4

$$\sigma_i^2(\alpha) = \int_{\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}} [P_i(X,Y,Z) - P_i(\alpha)]^2 \mathrm{d}P_\alpha(X,Y,Z), \quad i = 1,2,3,4. \tag{10}$$

Then let

$$\pi_i(\alpha) = \frac{\dfrac{1}{\sigma_i^2(\alpha)}}{\sum_{j=1}^{4} \dfrac{1}{\sigma_j^2(\alpha)}}, \quad i = 1,2,3,4. \tag{11}$$

Because α is unknown we can use its estimation $\hat{\alpha}$ instead of α in (10).

(Choice 3) Let $\pi_i = 1$ when the variance of $P_i(X,Y,Z)$ is the minimum one among four variances, that is

$$\pi_i(\hat{\alpha}) = \begin{cases} 1, & \text{if } \sigma_i^2(\hat{\alpha}) = \min\{\sigma_j^2(\hat{\alpha}), j = 1,2,3,4\} \\ 0, & \text{otherwise.} \end{cases}$$

Thus the empirical Bayes estimation is

$$\hat{\delta} = P_k(\hat{\alpha}), \quad \text{where} \quad k = \arg\min_i \{\sigma_i^2(\hat{\alpha}), i = 1,2,3,4\}$$

5. Experiment results

We repeat sampling parameters of this model for 50 times and obtain data sets of sample size 300 for every time.We can compute the ratios of average risks of empirical Bayes estimation $\hat{\delta}$ to the other four estimations $P_i(X,Y,Z)$ under three choices of π_i and count how many ratios are smaller than 1, $i=1,2,3,4$. The following table lists the result of the experiment, which reflects what percent the ratios are smaller than 1, where $R(\cdot)$ is the average risk.

choice	$R(\hat{\delta})/R(P_1(\hat{\alpha}))$	$R(\hat{\delta})/R(P_2(\hat{\alpha}))$	$R(\hat{\delta})/R(P_3(\hat{\alpha}))$	$R(\hat{\delta})/R(P_4(\hat{\alpha}))$
1	62%	72%	86%	90%
2	70%	90%	90%	96%
3	66%	62%	78%	82%

We see that most of the ratios are smaller than 1, that is, the empirical Bayes estimation is a better estimation.

References:

[1] RUBIN D B. *Bayes inference for causal effects: the role of randomization* [J]. Ann. Statist., 1978, **6**: 34–58.

[2] VERMA T, PEARL J. *Causal networks: semantics and expressiveness* [C]. Uncertainty in Artificial Intelligence, 1990, 69–76.

[3] PEARL J. *Causality: Models, Reasoning and Inference* [M]. Cambridge University Press, Cambridge, 2000.

[4] PEARL J. *Probabilistic Reasoning in Intelligent Systems* [M]. Morgan Kaufmann, San Mateo, 1988.

[5] ZHENG Z G, ZHANG Y Y, TONG X W. *Identifiability of causal effect for a simple causal model* [J]. Science in China, Ser. A, 2002, **45**: 335–341.

一个虚拟事实模型中因果效应的 Bayes 估计

许 静, 郑 忠 国

(北京大学数学学院概率统计系, 北京 100871)

摘 要: 给出了一个虚拟事实模型中因果效应的 Bayes 估计和经验 Bayes 估计, 提供了三种可替换性假设的先验分布的选择方法, 并用实验说明, 在不知道取哪个可替换性假设的情况下, 经验 Bayes 估计要优于其他的估计.

关键词: Bayes 估计; 因果效应; 虚拟事实模型; 干预; 可替换性.