Multidimensional Structural Regression Model for Causal Inference under Strongly Ignorable Treatment Assignment

Yonghong Wang
School of Mathematics and Computer, Harbin University, Harbin 150086, China
Email: wangyonghong418@163.com

Abstract  In the study of epidemiology aetiology, we usually cannot measure exposed effect relative to an individual, but under some assumptions, we approximately replace the exposed effect by estimator of population average causal effect. A multidimensional structural regression model for causal inference is established to estimate the population average treatment effect under strongly ignorable treatment assignment. Under the normal distribution, the maximum likelihood estimator for population average treatment effect is proved to be consistent, unbiased and asymptotically normal.

Keywords  Strongly ignorable treatment assignment  Causal inference  Population average treatment effect  Multidimensional structural regression model  Maximum likelihood estimator

1. Introduction

The objective of epidemiology study is to search aetiology, and to measure its causal effect in the light of quantity, thereby prevent from the occurrence of disease[1]. The case-control study, which attempts to find a contrasted or exposed group which can be compared with the treated or exposed one, is the common method that we employ in the research of epidemiology aetiology. Except the difference of being exposed and unexposed, the ideal treated group and the contrasted group are expected to share the rest of the other features, which serves as the guiding principle in the epidemiology research[2-3]. The dummy truth model laid a theoretical foundation for us to demonstrate this principle[4]. Let $U$ be the study population, and denote a generic individual in $U$ by $u \in U$. A variable $E$ is defined on each $u \in U$ so that $(1) E(u) = 1$ if $u$ is exposed to the casual agent of interest and $(0) E(u) = 0$ if $u$ is not so exposed. $Y_1(u)$ and $Y_0(u)$ indicate the diseased status of $u$ as in the exposed and non-exposed cases respectively. Thus we can define the exposed causal effect of the individual $u$ as $Y_1(u) - Y_0(u)$. As we know, in any real epidemiologic study, one does not observe $Y_1(u)$ and $Y_0(u)$ simultaneously. Therefore the causal effect of the exposure to the individual is not attainable. However, in some circumstances, we can get the population average causal effect, which can be defined as $E(Y_1(u) - Y_0(u))$, while $E(\cdot)$ denotes expection or population average (over $U$). Under the circumstances of random tests, for the fact that the independence of the exposed $E$ and other variants, the result should be $E[Y_1, Y_0]$, which indicates that the exposed $E$ is independent of the random Vector $(Y_1, Y_0)$. Thus the population average causal effect is shown as $E(Y_1 - Y_0) = E(Y_1) - E(Y_0) = E(Y_1|E = 1) - E(Y_0|E = 0)$. In other words, the causal effect can be estimated by the difference of the exposed population average diseased condition the non-exposed population average diseased condition[5]. The epidemiology, however, is an

*This work has been supported by the Science Development Foundation of Harbin University(HXKQ200701).
observatory science in nature, so the above formula is not to be established which necessitates the assumption of the state of being ignorable\(^5\)-\(^6\).

**Definition 1.1** Given the covariant \(X\), the exposed \(E\) is called being strongly ignorable. If

\[
\begin{align*}
(1) & \quad (Y_1, Y_0) \perp\!
\perp E \mid X, \text{i.e. the observed covariant } X \text{ is given, the response } \text{variant is independent of the exposed } E. \\
(2) & \quad 0 < \Pr(E = e \mid X) < 1, \text{i.e. for each and every } X, \text{ they may receive various treatment.}
\end{align*}
\]

When the \(E\) is strongly ignorable, by getting the matching-differencing-averaging between the exposed and non-exposed groups, we can get the unbiased estimate of population average causal effect\(^7\). But the above method is complicated in calculation and the additional sparse-data problems, excess covariant, though few, may arise. In addition, the method of ranking the individuals with the estimated propensity score is an advisable way to solve the problem, but the calculation is still very complicated\(^8\)-\(^9\). Under strongly ignorable treatment assignment, a structural regression model for causal inference is established to estimates the population average treatment effect. They are the series of results achieved. with the response variant being one dimensional. The paper will discuss the inference problem with multi-dimensional response variants.

### 2. Multidimensional Structural Regression Model

**Definition 2.1** In Condition of Strongly Ignorable Treatment Assignment, denote multidimensional Structural Regression model as follows:

\[
\begin{align*}
Y_j &= \alpha_t + \beta_t X_{jt} + \epsilon_j \\
X_{jt} &\sim N(\mu_x, \Sigma_x), \epsilon_j \sim N(0, \Sigma_e), X_{jt} \text{ is independent of } \epsilon_j
\end{align*}
\]

Here \(t=0,1\) means two treatment; \(Y_j\) is the random vector of order \(p \times 1\), indicated the response variant of the \(j\)th observation for the \(t\)th treatment; \(X_{jt}\) is the covariant vector of order \(k \times 1\); \(\epsilon_j\) is the random error vector of order \(p \times 1\); \(\alpha_t\) is the parameter vector of order \(p \times 1\); \(\beta_t\) is the parameter matrix of order \(p \times k\); \(\mu_x\) is the parameter vector of order \(k \times 1\); the covariant matrix \(\Sigma_x, \Sigma_e|V_t\) of \(X_{jt}\) and \(\epsilon_j\) are both positive definite.

Under model (2)

\[
E\left(\begin{bmatrix} Y_j \\ Y\end{bmatrix}\right) = \begin{bmatrix} \mu_x \\ \alpha_t + \beta_t \mu_x \end{bmatrix}, A_t = \text{Cov}\left(\begin{bmatrix} Y_j \\ Y\end{bmatrix}\right) = \begin{bmatrix} \Sigma_x & \Sigma_e \beta_t' \\ \beta_t \Sigma_x & \beta_t \Sigma_e \beta_t' + \Sigma_e \end{bmatrix}
\]

We can easily find \(|A_t| = \Sigma_x \cdot |V_t|\); because \(\Sigma_x\) and \(V_t\) are both positive definite, we can know \(\Sigma_x \neq 0, V_t \neq 0\); so \(A_t \neq 0\); therefore \(A_t^{-1}\) is valid, and we denote it as \(B_t\), again based on the principle of block matrix inversion, we can get

\[
B_t = \begin{bmatrix} \Sigma_x^{-1} + \beta_t V_t^{-1} \beta_t & -\beta_t V_t^{-1} \\ V_t^{-1} \beta_t & V_t^{-1} \end{bmatrix}
\]

So

\[
\begin{bmatrix} X_{jt} \\ Y_j \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_x \\ \alpha_t + \beta_t \mu_x \end{bmatrix}, A_t\right)
\]

Where: \(t=0, 1; \quad j=1, \ldots, n_t\).

We also get observation logarithm likelihood function of the sample \((x_{jt}, y_j)\) with the
capacity \( n_i \) \( (t=0, 1, \ldots, n_t) \)

\[
-2 \ln L = \sum_{t=0}^{n_t} n_i \ln |A_i| + \sum_{t=0}^{n_t} \sum_{j=1}^{n} \left( x_{ij} - \mu_x \right) \left( y_{ij} - \alpha_t - \beta_t \mu_x \right)
\]

Furthermore,

\[
-2 \ln L = \sum_{t=0}^{n_t} n_i (\ln |\Sigma_x| + \ln |V_t|) + \sum_{t=0}^{n_t} \sum_{j=1}^{n} \left( x_{ij} \Sigma_x^{-1} x_{ij} - 2x_{ij} \Sigma_x^{-1} \mu_x \right)
\]

\[
+ \mu_x \Sigma_x^{-1} \mu_x + x_{ij} \beta_t V_i^{-1} \beta_t x_{ij} - 2x_{ij} \beta_t V_i^{-1} y_{ij}
\]

\[
+ 2x_{ij} \beta_t V_i^{-1} \alpha_t + y_{ij} V_i^{-1} y_{ij} - 2y_{ij} V_i^{-1} \alpha_t + \alpha_t V_i^{-1} \alpha_t
\]

3. Conclusion

3.1 Two Lemmas

In order to get the likelihood estimates of all parameters in model (2) and the population average causal effect, we give the following two lemmas, of which lemma 3.1 is shown at [10].

First, suppose \( X \) is the matrix of order \( m \times n \), \( f(x) \) is the real-valued function of matrix \( X \), \( \frac{\partial f(x)}{\partial x} \) indicate real-valued function \( f(x) \) settling partial derivation for each element of matrix \( X \). Thus the following lemmas are valid:

Lemma 3.1

\[
(1) \quad \frac{\partial \text{tr}(AXB)}{\partial X} = AB
\]

\[
(2) \quad \frac{\partial \text{tr}(X'AXBX)}{\partial X} = AXB + A' XB'
\]

\[
(3) \quad \frac{\partial \ln |X|}{\partial X} = (X^{-1})'
\]

Of which \( A \) and \( B \) are two matrixes that can make matrix calculation with \( X \), \( \text{tr}(X) \) means the trace of matrix \( X \).

Lemma 3.2 If model (2) is valid, when the covariant matrix and the random error matrix are independent of each other, then

\[
X_i = \frac{1}{n_t} \sum_{j=1}^{n} X_{ij}
\]

and

\[
M_{X_i X_i} = \frac{1}{n_t} \sum_{j=1}^{n} (X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)'
\]

\[
M_{e_i X_i} = \frac{1}{n_t} \sum_{j=1}^{n} (e_{ij} - \bar{e}_i)(X_{ij} - \bar{X}_i)
\]

are independent of each other, where \( \bar{e}_i = \frac{1}{n_t} \sum_{j=1}^{n} e_{ij} \).

Proof: First to prove \( \bar{X}_i \) and \( M_{X_i X_i} \) are independent.
Let \( X_i = \begin{pmatrix} X_{r1} & X_{r2} & \cdots & X_{rk} \\ X_{r1} & X_{r2} & \cdots & X_{rk} \\ \vdots & \vdots & \ddots & \vdots \\ X_{rn1} & X_{rn2} & \cdots & X_{nk} \end{pmatrix} \)

Set \( \bar{X}_i = \frac{1}{n} X_i' \mathbf{1} \) also

\[
M_{X,X_i'} = \frac{1}{n_t} (X_i' - \frac{1}{n_t} X_i' \mathbf{1} \mathbf{1}') (X_i' - \frac{1}{n_t} \mathbf{1} \mathbf{1}' X_i') = \frac{1}{n_t} X_i' (\mathbf{1}_n - \frac{1}{n_t} \mathbf{1} \mathbf{1}') X_i
\]

Which \( \mathbf{1} \) is the column vector of \( n_t \times 1 \).

We can do the vector straight operations to matrix \( X_i \), then \( \text{Cov}(\text{Vec}(X)) = \Sigma_x \otimes \mathbf{1}_n \),
also, as for \( \bar{X}_i \), and each at the element \( m_{ij} \) in \( M_{X,X_i'} \)

follow \( \bar{X}_i = \frac{1}{n_t} (\mathbf{1}_k \otimes \mathbf{1}') \text{Vec}(X_i) \)

\[
m_{ij} = \text{Vec}(X_i') [\frac{1}{2} (E_{ij} + E_{ij}') \otimes (\mathbf{1}_n - \frac{1}{n_t} \mathbf{1} \mathbf{1}')] \text{Vec}(X_i)
\]

Of which \( \otimes \) indicates the Kronecker product of matrix, \( E_{ij} \) indicates matrix of order \( k \times k \) with \( (i,j) \)\(^{th}\) is 1, and the rest is 0. Now we can draw the following conclusion by making use of the characteristic of multidimensional normal random variant and the Kronecker product of the matrix:

\[
\frac{1}{n_t} (\mathbf{1}_k \otimes \mathbf{1}') (\Sigma_x \otimes \mathbf{1}_n) [\frac{1}{2} (E_{ij} + E_{ij}') \otimes (\mathbf{1}_n - \frac{1}{n_t} \mathbf{1} \mathbf{1}')] = 0
\]

Thus we can prove \( m_{ij} \) and \( \bar{X}_i \) are independent, and it’s valid to any \( i,j=1, \ldots, k \). so we have proved \( \bar{X}_i = \frac{1}{n_t} \sum_{j=1}^{n_t} X_{ij} \) and \( M_{X,X_i'} = \frac{1}{n_t} \sum_{j=1}^{n_t} (X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)' \) are independent of each other.

Suppose \( e_i = \begin{pmatrix} e_{r1} & e_{r12} & \cdots & e_{r1p} \\ e_{r21} & e_{r22} & \cdots & e_{r2p} \\ \vdots & \vdots & \ddots & \vdots \\ e_{rn1} & e_{rn2} & \cdots & e_{rn,p} \end{pmatrix} \)

\( Z_i = (X_i, e_i) \), in the same way , and by mutual independence of \( X_i \) and \( e_i \), we can prove \( \bar{X}_i \) and \( M_{e,e_i} \) are also independent of each other.

Lemma3.2 is proved!
3.2 Estimate of Parameter

**Theorem 3.1** Suppose model (2) is valid, we can get the maximum likelihood estimates of all the parameter:

\[
\hat{\alpha}_i = \bar{y}_i - M_{y_i x_i, (M_{x_i}^{-1})^{-1}}x_i; \quad \hat{\beta}_i = M_{y_i x_i, (M_{x_i}^{-1})^{-1}}x_i; \quad \hat{\mu}_i = \frac{1}{n_0 + n_1} \sum_{j=0}^{n_1} x_{ij};
\]

\[
\hat{\Sigma}_x = \frac{1}{n_0 + n_1} \sum_{i=0}^{n_1} \sum_{j=1}^{n} (x_{ij} - \hat{\mu}_i)(x_{ij} - \hat{\mu}_i)'; \quad \hat{V}_i = \frac{1}{n_1} \sum_{j=1}^{n} (y_{ij} - \hat{\alpha}_i - \hat{\beta}_i x_{ij})(y_{ij} - \hat{\alpha}_i - \hat{\beta}_i x_{ij})'.
\]

Of which \( \bar{y}_i = \frac{1}{n_1} \sum_{j=1}^{n} y_{ij} \), \( M_{y_i x_i} = \frac{1}{n_1} \sum_{j=1}^{n} (y_{ij} - \bar{y}_i)(x_{ij} - \bar{x}_i)'; \quad \) The form of \( \bar{x}_i \) and \( M_{x_i} \) is shown in lemma 3.2.

**Proof:** First, we get the score function of \( \hat{\alpha}_i \):

\[
\frac{\partial(-2 \ln L)}{\partial \alpha_i} = \sum_{j=1}^{n_1} \frac{\partial}{\partial \alpha_i} (2x_{ij}'\beta_i V^{-1}_i \alpha_i - 2y_{ij}'V^{-1}_i \alpha_i + \alpha_i V^{-1}_i \alpha_i)
\]

\[
= \sum_{j=1}^{n_1} \left[ \frac{\partial}{\partial \alpha_i} 2tr(x_{ij}'\beta_i V^{-1}_i \alpha_i) - 2 \frac{\partial}{\partial \alpha_i} tr(y_{ij}'V^{-1}_i \alpha_i) + \frac{\partial}{\partial \alpha_i} tr(\alpha_i V^{-1}_i \alpha_i) \right]
\]

\[
= \sum_{j=1}^{n_1} (2V^{-1}_i \beta_i x_{ij} - 2V^{-1}_i \alpha_i) = 0
\]

We can get \( \beta_i = \bar{y}_i - \beta_i \bar{x}_i \). And the score function of \( \hat{\beta}_i \) is as follows:

\[
\frac{\partial(-2 \ln L)}{\partial \beta_i} = \sum_{j=1}^{n_1} (2V^{-1}_i \beta_i x_{ij} x_{ij}' - 2V^{-1}_i y_{ij} x_{ij}' + 2V^{-1}_i \alpha_i x_{ij}') = 0
\]

Simplified as

\[
\sum_{j=1}^{n_1} (\beta_i x_{ij} x_{ij}' - y_{ij} x_{ij}' + \alpha_i x_{ij}') = 0
\]

Fit \( \beta_i \) into above formula of \( \hat{\alpha}_i \), we set \( \hat{\alpha}_i = \bar{y}_i - M_{y_i x_i, (M_{x_i}^{-1})^{-1}}x_i \).

In the same way, we can get the maximum likelihood estimates of the other three parameters.

Theorem 3.1 is proved!

In fact, as for the multidimensional structural regression model with the response variants as shown in model (2), we can see from the above demonstration that the maximum likelihood estimates of parameter \( \alpha_i \) and \( \beta_i \) do not depend on the selection of the covariant matrix and random error matrix. Moreover, we can also see that if we want to estimate the parameter that measure relationship between the covariant and the \( j \)th \((j=1, \ldots, n_1) \) response variable, we should only note the \( j \)th response variable. i.e. the model (2) can be divided into \( p \) one dimensional models, which is evident in Jinhua’s paper (2000).

3.3 Population Average Causal Effect

**Definition 3.1** provides the covariant X, if the assumption of Strongly Ignorable condition (1) is fulfilled, then treatment \( E=1 \) to \( E=0 \) average causal effect \( ATE \) is

\[
ATE = E(Y_1) - E(Y_0) = E_x(E(Y_1) - E(Y_0) | X = x)
\]
Theorem 3.2 suppose model (2) is valid, and treatment assignment variant E is strongly ignorable, then

1) The maximum likelihood estimator of population average treatment effect $ATE$ is

$$
\hat{ATE} = \bar{Y}_1 - \bar{Y}_0 - \frac{n_0\hat{\beta}_1 + n_1\hat{\beta}_0}{n_0 + n_1}(\bar{X}_1 - \bar{X}_0)
$$

2) $\hat{ATE}$ is consistent unbiased estimator of $ATE$.

3) With the case of the large sample, $\hat{ATE}$ is close to normal distribution $N(ATE, \Gamma)$.

Proof: (1) with model (2), if treatment assignment variant E is strongly ignorable, then the population average treatment effect $ATE = (\alpha_1 - \alpha_0) + (\beta_1 - \beta_0)\mu_x$

Based on theorem 3.1 and the invariance of maximum likelihood estimator, we can know

$$
\hat{ATE} = (\hat{\alpha}_1 - \hat{\alpha}_0) + (\hat{\beta}_1 - \hat{\beta}_0)\hat{\mu}_x = \bar{Y}_1 - \bar{Y}_0 - \frac{n_0\hat{\beta}_1 + n_1\hat{\beta}_0}{n_0 + n_1}(\bar{X}_1 - \bar{X}_0)
$$

(2) Consistency: we know based on the model (2)

$$
\hat{\beta}_t = M_{y_t|x_t}(M_{x_t|x_t})^{-1} = \beta_t + (M_{e_t|x_t})(M_{x_t|x_t})^{-1}
$$

Based on the law of large numbers and the nature of the random variant order converging by probability, we can know

$$
M_{x_t|x_t} \xrightarrow{p} \Sigma_x, M_{e_t|x_t} \xrightarrow{p} 0
$$

Then $\hat{\beta}_t \xrightarrow{p} \beta_t$

In the same way, $\bar{X}_t \xrightarrow{p} \mu_x, \bar{Y}_t \xrightarrow{p} \alpha_t + \beta_t\mu_x$

So far, again based on the nature of the random variant order converging by probability, we get $\hat{ATE} \xrightarrow{p} ATE$, i.e. $\hat{ATE}$ is consistent estimate of the population average treatment effect $ATE$.

Unbiased: Through the demonstration of the consistency, we know

$$
\hat{\beta} = \beta_t + \left[\frac{1}{n_t} \sum_{j=1}^{n_t} (e_{y_j} - \bar{e}_t)(X_{y_j} - \bar{X}_t)^\top\right](M_{x_t|x_t})^{-1}
$$

Whereupon $E(\hat{\beta}) = E(\beta_t + \left[\frac{1}{n_t} \sum_{j=1}^{n_t} (e_{y_j} - \bar{e}_t)(X_{y_j} - \bar{X}_t)^\top\right](M_{x_t|x_t})^{-1})$

$$
= \beta_t + E \left[\frac{1}{n_t} \sum_{j=1}^{n_t} (e_{y_j} - \bar{e}_t)(X_{y_j} - \bar{X}_t)^\top\right](M_{x_t|x_t})^{-1}|X_t = x_t
$$

Therefore $E(\bar{Y}_t) = \alpha_t + \beta_t\mu_x$, so based on lemma 3.2, we set
\[
E(\widehat{ATE}) = E(\bar{X} - \bar{Y}) - E\left(\frac{n_0\hat{\beta}_1 + n_1\hat{\beta}_0}{n_0 + n_1}E(\bar{X} - \bar{X}_0)\right)
\]

\[
= (\alpha_1 - \alpha_0) + (\beta_1 - \beta_0)\mu_x = ATE
\]

(3) Asymptotically normality: Based on above demonstration, we know \(E(\widehat{ATE}) = ATE\)

Also because of
\[
\widehat{ATE} \approx (\bar{X} - \bar{Y}) - \frac{n_0\hat{\beta}_1 + n_1\hat{\beta}_0}{n_0 + n_1}(\bar{X} - \bar{X}_0)
\]

\[
= [\alpha_1 + \frac{n_1(\beta_1 - \beta_0)}{n_0 + n_1}\bar{X}_1 + \bar{\varepsilon}_1] - [\alpha_0 + \frac{n_0(\beta_1 - \beta_0)}{n_0 + n_1}\bar{X}_0 + \bar{\varepsilon}_0]
\]

Moreover \(\text{Var}(\bar{X}_1) = \frac{1}{n_1}\sum_x\), \(\text{Var}(\bar{\varepsilon}_i) = \frac{1}{n_1}V_i\)

Therefore
\[
\text{Var}(\alpha_1 + \frac{n_1(\beta_1 - \beta_0)}{n_0 + n_1}\bar{X}_1 + \bar{\varepsilon}_1) = \frac{n_1}{(n_0 + n_1)^2}(\beta_1 - \beta_0)\sum_x(\beta_1 - \beta_0) + \frac{1}{n_1}V_1
\]

\[
\text{Var}(\alpha_0 + \frac{n_0(\beta_1 - \beta_0)}{n_0 + n_1}\bar{X}_0 + \bar{\varepsilon}_0) = \frac{n_0}{(n_0 + n_1)^2}(\beta_1 - \beta_0)\sum_x(\beta_1 - \beta_0) + \frac{1}{n_0}V_0
\]

So
\[
\text{Var}(\widehat{ATE}) = \frac{1}{n_1}V_1 + \frac{1}{n_0}V_0 + \frac{1}{n_0 + n_1}(\beta_1 - \beta_0)\sum_x(\beta_1 - \beta_0) \overset{\Delta}{=} \Gamma
\]

So based on the Central Limited Theorem, we know \(\widehat{ATE} \sim AN(ATE, \Gamma)\).

Theorem 3.2 is proved!

References