Journal of Statistical Software

*MMMMMM YYYY, Volume VV, Issue II.* doi: 10.18637/jss.v000.i00

# MIE: An approach to obtaining causal effect estimation in R

**Bo Liu***
Peking University

**Zhixuan Shao***
Peking University

**Xiaohua Zhou**†
Peking University

**Abstract**

*Keywords*: causal effect, survival, treatment, substitution variable, R.

## 1. Introduction

Researches regarding causal effect allure much attention of statisticians due to its various applications in scholarship as well as industry. In the field of biostatistics, one essential application is to determine the causal effect on outcomes a specific treatment may exert. However, the distinguishing charisteristic of ineluctable deaths among the examinees may possibly invalidate the identifiablity of the group we focus on, where patients will survive both with and without the treatment. A number of new methods have been proposed to eliminate the obstacle. One pertinent paper **?** suggests the identifiablity can be vindicated by introducing a substitution variable, and both the estimated causal effect and its asymptic variance are provided. The R package (?) is designed in light of this method.

The paper is organized as follows. Section 2 retraces the model implemented in the package, Section 3 focuses on the data types and functions of the package, and Section 4 provides formulas on calculating variances in a more detailed manner. In Section 5, a specific dataset is generated and tested by the method, and explanations about the results will be analysed in Section 6.

## 2. The Causal Effect Model

*Equally contributed
†Corresponding Author

### 2.1. Survivor Average Causal Effect(SACE)

Consider a trail for a treatment with a single follow-up visit. Let $Z$ be the exposure indicator to the treatment, i.e., $Z = 1$ means the patient receives the treatment and $Z = 0$ means not. Each patient may have two potential survival statuses $S(1)$ and $S(0)$, defined as whether the patient would be alive at the time of follow-up visit had the subject been exposed or not respectively. **?** has defined the causal effect to be $\mathbb{E}(S(1)) - \mathbb{E}(S(0))$, which indicates the difference in the probability of survival between subjects with and without treatement. **?** represents with two letters L and D for 4 potential principle strata of the subjects, where L represents live and D stands for die. For example, if a person will live if treated but will die if not, he belongs to the stratum of LD. **?** extends the causal effect from survival status to the outcomes observed. Similarly, we define $Y(1)$ and $Y(0)$ as the potential outcomes had the subject been exposed or not, respectively. However, these outcomes can only be observed if the subjects are still alive. Therefore, the causal effect of treatment on outcome is defined as

$$\Delta_{\mathrm{LL}} = \mathbb{E}\{Y(1) - Y(0)|G = \mathrm{LL}\}. \tag{1}$$

This is also known as the survivor average causal effect (SACE).

### 2.2. Assumptions of the Model

In real experiments, it is impossible to devide the subjects in the four strata *a priori*. Those alive and treated belong to strata LL or LD, while those alive but not treated belong to LL or DL. The specific stratum LL is unidentifiable.

**?** assumes that

$$S(1) \geq S(0) \text{ a.s.} \tag{2}$$

Let $W$ be the covariates of subjects, and **?** referred as "strong ignorability of treatment assignment" to the assumption

$$Z \perp\!\!\!\perp S(z)|W \text{ for } z = 0, 1. \tag{3}$$

In light of the "missing at random" assumption (**?**), we have

$$Z \perp\!\!\!\perp Y(z)|W, G = \mathrm{LL} \text{ for } z = 0, 1. \tag{4}$$

Let baseline covariates $W = (X, A)$ where $X$ is similar to a confounder while $A$ is a substitution variable, which satisfies the substitution and exclusion assumption (**?**)

$$\begin{cases} A \perp\!\!\!\perp Y(1)|Z = 1, G, X, \\ A \not\perp\!\!\!\perp G|Z = 1, S = 1, X, \end{cases} \tag{5}$$

and the SACE (Eqn.1) is identifiable with assumptions Eqn.2-5 on the variables (**?**).

### 2.3. Identification of SACE

We illustrate hereafter the main points in the proof by **?**.

$$\mathbb{E}(Y(z)|G = \mathrm{LL}) = \frac{\mathbb{E}_w(\mu_{z,\mathrm{LL},w} \cdot \pi_{\mathrm{LL}|w})}{\mathbb{E}_w(\pi_{\mathrm{LL}|w})}, \tag{6}$$

where $\mu_{z,g,w} = \mathbb{E}(Y(z)|G = g, W = w)$ and $\pi_{g|w} = \mathbb{P}(G = g|W = w)$.
Under Eqn.4,

$$\mu_{z,g,w} = \mathbb{E}(Y(z)|Z = z, G = g, W = w) = \mathbb{E}(Y|Z = z, G = g, W = w). \tag{7}$$

Under Eqn.2 and Eqn.3,

$$\pi_{\mathrm{LL}|w} = \mathbb{P}(S(0) = 1, S(1) = 1|W = w) = \mathbb{P}(S(0) = 1|W = w) = \mathbb{P}(S = 1|Z = 0, W = w). \tag{8}$$

$$\pi_{\mathrm{LD}|w} = \mathbb{P}(S(0) = 0, S(1) = 1|W = w) = \mathbb{P}(S = 1|Z = 1, W = w) - \mathbb{P}(S = 1|Z = 0, W = w). \tag{9}$$

From Eqn.6,

$$\mu_{0,\mathrm{LL},w} = \mathbb{E}(Y|Z = 0, S = 1, W = w). \tag{10}$$

Since $Y, Z, S$ and $W$ are all observable from the data, $\pi_{\mathrm{LL}|w}$, $\pi_{\mathrm{LD}|w}$ and $\mu_{0,\mathrm{LL},w}$ are identifiable. Denote $W$ in the form of $(X, A)$, and define

$$p_{g|z,x,a,s} = \frac{\mathbb{P}(G = g|X = x, A = a)}{\mathbb{P}(S = s|Z = z, X = x, A = a)}. \tag{11}$$

It can be proven that under Eqn.5,

$$\mathbb{E}(Y|Z = 1, S = 1, X = x, A = a) = p_{\mathrm{LL}|1,x,a,1}\mu_{1,\mathrm{LL},x} + (1 - p_{\mathrm{LL},1,x,a,1})\mu_{1,\mathrm{LD},x}. \tag{12}$$

For $\forall x$, there exist different $a_1$ and $a_2$ satisfying Eqn.12 (under Eqn.5), and thus $\mu_{1,\mathrm{LL},x}$ can be identified.

### 2.4. Model Parameterization

For simplicity, we use (generalized) linear model as our assumption.

$$\mu_{0,\mathrm{LL},W} = \mathbb{E}(Y|Z = 0, S = 1, X, A) = \alpha_{10} + X\alpha_{11} + A\alpha_{12}. \tag{13}$$

$$\mu_{1,\mathrm{L}\cdot,W} = \mathbb{E}(Y|Z = 1, S(0) = L, S(1) = 1, X, A) = \alpha_{20} + X\alpha_{21} + L\alpha_{22}. \tag{14}$$

In Eqn.14, $A$ is dropped because under Eqn.5, $A \perp\!\!\!\perp Y(1)|Z = 1, S(0) = L, S(1) = 1, X$. $L$ is not observable when $S(1) = 1$, so we need to figure out later how to estimate $L$.

$$\mathbb{P}(S = 1|Z = 1, W) = \mathrm{expit}(\beta_0 + X\beta_1 + A\beta_2). \tag{15}$$

$$\frac{\mathbb{P}(S = 1|Z = 0, W)}{\mathbb{P}(S = 1|Z = 1, W)} = \mathrm{expit}(\gamma_0 + X\gamma_1 + A\gamma_2). \tag{16}$$

One can verify that these assumptions on the parameterization satisfy the requirements of the model.

## 3. Parameter Estimation

### 3.1. Data

The data input should include multiple observations on random variables $Z, S, Y, X$ and $A$. Suppose the $i$-th observation can be denoted by $(Z^i, S^i, Y^i, X^i, A^i)$, and they denote respectively:

- $Z^i$: an integer valued 0 or 1, exposure indicator of the $i$-th subject;
- $S^i$: an integer valued 0 or 1, survival indicator of the $i$-th subject;
- $Y^i$: a numeric vector, indicating the survival outcome of the $i$-th subject;
  - $Y^i$ is discarded where $S^i = 0$.
  - A warning will be raised if $Y^i$ is NA where $S^i = 1$.
- $X^i$: a numeric row vector, the confounder of covariate $W$ of the $i$-th subject.
- $A^i$: a numeric row vector, the substitution variable of covariate $W$ of the $i$-th subject.

Accordingly, the data type of observed data matrix $\mathbf{Z}, \mathbf{S}, \mathbf{Y}, \mathbf{X}$ and $\mathbf{A}$ are defined as Table 3.2.

| Variable | Data Type | Size |
|---|---|---|
| $\mathbf{Z}$ | Logical Column Vector | $n$ |
| $\mathbf{S}$ | Logical Column Vector | $n$ |
| $\mathbf{Y}$ | Numeric Matrix | $n \times k$ |
| $\mathbf{X}$ | Numeric Matrix | $n \times p$ |
| $\mathbf{A}$ | Numeric Matrix | $n \times q$ |

Table 1: Data Types of Observed Variables

### 3.2. Parameters

In Sec. 2.4, we put forward four models with parameters $\alpha_0$, $\alpha_1$, $\beta$ and $\gamma$. Here we put it in the matrix form.

Let

$$\alpha_1 = \begin{pmatrix} \alpha_{10} \\ \alpha_{11} \\ \alpha_{12} \end{pmatrix}, \alpha_2 = \begin{pmatrix} \alpha_{20} \\ \alpha_{21} \\ \alpha_{22} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}, \gamma = \begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \end{pmatrix}, \tag{17}$$

be the parameters, and let

$$\begin{aligned} \mathbf{W} &= (\mathbf{X} \quad \mathbf{A}), \\ \widetilde{\mathbf{W}} &= (\mathbf{1}_n \quad \mathbf{W}), \\ \widetilde{\mathbf{X}} &= (\mathbf{1}_n \quad \mathbf{X}). \end{aligned} \tag{18}$$

be the observed data matrix of random variables $W$, $\widetilde{W}$ and $\widetilde{X}$ respectively, then our parametric model can be rewritten as

$$\begin{aligned}\mathbb{E}(Y|Z=0,S=1,X,A) &= \widetilde{W}\alpha_1,\\ \mathbb{E}(Y|Z=1,S(0)=L,S(1)=1,X,A) &= (\widetilde{X}\ \ L)\alpha_2,\\ \mathbb{P}(S=1|Z=1,W) &= \text{expit}(\widetilde{W}\beta),\\ \frac{\mathbb{P}(S=1|Z=0,W)}{\mathbb{P}(S=1|Z=1,W)} &= \text{expit}(\widetilde{W}\gamma).\end{aligned} \tag{19}$$

| Parameter | Data Type | Size |
|---|---|---|
| $\alpha_1$ | Numeric Matrix | $(p+q+1)\times k$ |
| $\alpha_2$ | Numeric Matrix | $(p+2)\times k$ |
| $\beta$ | Numeric Vector | $p+q+1$ |
| $\gamma$ | Numeric Vector | $p+q+1$ |

Table 2: Data Types of Parameters

## 3.3. Estimating the Parameters

*Estimating* $\alpha_1$

Since

$$\mathbb{E}(Y|Z=0,S=1,X,A) = \widetilde{W}\alpha_1,$$

we apply the ordinary least square method (OLS) to estimate $\alpha_1$. Let $\mathbf{Y}_{01}$ and $\widetilde{\mathbf{W}}_{01}$ be the observed data of $Y$ and $\widetilde{W}$ in the subset where $Z=0$ and $S=1$, and the estimator of $\alpha_1$ can be

$$\widehat{\alpha}_1 = (\widetilde{\mathbf{W}}_{01}^{\mathrm{T}}\widetilde{\mathbf{W}}_{01})^{-1}\widetilde{\mathbf{W}}_{01}^{\mathrm{T}}\mathbf{Y}_{01}. \tag{20}$$

*Estimating* $\alpha_2$

It may be natural to apply here the same method how we estimate $\alpha_1$, but it can be rather tricky because $L$ is not observable. By doing simple algebra, one can avoid the problem by taking expectation over $L$.

$$\begin{aligned}\mathbb{E}(Y|Z=1,S(1)=1,X,A) =& \mathbb{E}(Y|Z=1,S(0)=0,S(1)=1,X,A)\mathbb{P}(S(0)=0|Z=1,S(1)=1,W)\\ &+\mathbb{E}(Y|Z=1,S(0)=1,S(1)=1,X,A)\mathbb{P}(S(0)=1|Z=1,S(1)=1,W)\\ =&(\widetilde{X}\ \ 0)\alpha_2\mathbb{P}(S(0)=0|Z=1,S(1)=1,W)\\ &+(\widetilde{X}\ \ 1)\alpha_2\mathbb{P}(S(0)=1|Z=1,S(1)=1,W)\\ =&(\widetilde{X}\ \ \mathbb{P}(S(0)=1|Z=1,S(1)=1,W))\\ =&\left(\widetilde{X}\ \ \frac{\mathbb{P}(S=1|Z=0,W)}{\mathbb{P}(S=1|Z=1,W)}\right)\\ =&\left(\widetilde{X}\ \ \text{expit}(\widetilde{W}\gamma)\right).\end{aligned} \tag{21}$$

Let $\mathbf{Y}_{11}$, $\widetilde{\mathbf{X}}_{11}$ and $\widetilde{\mathbf{W}}_{11}$ be the observed data of $Y$, $\widetilde{X}$ and $\widetilde{W}$ in the subset where $Z=1$ and $S=1$, $\widehat{\gamma}$ be the estimator of $\gamma$. Denote $\widetilde{\mathbf{XL}}_{11}=\left(\widetilde{\mathbf{X}}_{11}\ \ \mathrm{expit}(\widetilde{\mathbf{W}}_{11}\widehat{\gamma})\right)$, and the estimator of $\alpha_2$ can be

$$\widehat{\alpha}_2=(\widetilde{\mathbf{XL}}_{11}^{\mathrm{T}}\widetilde{\mathbf{XL}}_{11})^{-1}\widetilde{\mathbf{XL}}_{11}^{\mathrm{T}}\mathbf{Y}_{11}. \tag{22}$$

*Estimating $\beta$ and $\gamma$*

Note that the model involving $\beta$ and $\gamma$ is equivalent to

$$\begin{aligned} S|Z=1,W &\sim \mathcal{B}(1,\mathrm{expit}(\widetilde{W}\beta)),\\ S|Z=0,W &\sim \mathcal{B}(1,\mathrm{expit}(\widetilde{W}\beta)\mathrm{expit}(\widetilde{W}\gamma)). \end{aligned} \tag{23}$$

Define

$$\begin{aligned} L(\beta,\gamma|z,s,w) &= \mathbb{P}(S=s|\beta,\gamma,Z=z,W=w)\\ &= \begin{cases} \mathrm{expit}(\widetilde{w}\beta), & z=1,s=1\\ 1-\mathrm{expit}(\widetilde{w}\beta), & z=1,s=0\\ \mathrm{expit}(\widetilde{w}\beta)\mathrm{expit}(\widetilde{w}\gamma), & z=0,s=1\\ 1-\mathrm{expit}(\widetilde{w}\beta)\mathrm{expit}(\widetilde{w}\gamma). & z=0,s=0 \end{cases} \end{aligned} \tag{24}$$

and

$$\begin{aligned} D_\beta L(\beta,\gamma|z,s,w) &= \frac{\partial}{\partial\beta}\ln L(\beta,\gamma|z,s,w)\\ &= \begin{cases} (1-\mathrm{expit}(\widetilde{w}\beta))\widetilde{w}, & z=1,s=1\\ -\mathrm{expit}(\widetilde{w}\beta)\widetilde{w}, & z=1,s=0\\ (1-\mathrm{expit}(\widetilde{w}\beta))\widetilde{w}, & z=0,s=1\\ -\dfrac{(1-\mathrm{expit}(\widetilde{w}\beta))}{\mathrm{expit}^{-1}(\widetilde{w}\beta)\mathrm{expit}^{-1}(\widetilde{w}\gamma)-1}\widetilde{w}. & z=0,s=0 \end{cases} \end{aligned} \tag{25}$$

$$\begin{aligned} D_\gamma L(\beta,\gamma|z,s,w) &= \frac{\partial}{\partial\gamma}\ln L(\beta,\gamma|z,s,w)\\ &= \begin{cases} 0, & z=1\\ (1-\mathrm{expit}(\widetilde{w}\gamma))\widetilde{w}, & z=0,s=1\\ -\dfrac{(1-\mathrm{expit}(\widetilde{w}\gamma))}{\mathrm{expit}^{-1}(\widetilde{w}\beta)\mathrm{expit}^{-1}(\widetilde{w}\gamma)-1}\widetilde{w}. & z=0,s=0 \end{cases} \end{aligned} \tag{26}$$

Using maximum likelihood estimation,

$$(\widehat{\beta},\widehat{\gamma})=\arg\max\sum_{i=1}^{n}\ln L(\beta,\gamma|z^i,s^i,w^i). \tag{27}$$

# 4. Illustrations

For a simple illustration of basic Poisson and NB count regression the `quine` data from the **MASS** package is used. This provides the number of `Days` that children were absent from school in Australia in a particular year, along with several covariates that can be employed as regressors. The data can be loaded by

```
R> data("quine", package = "MASS")
```

and a basic frequency distribution of the response variable is displayed in Figure **??**.

> For code input and output, the style files provide dedicated environments. Either the "agnostic" `{CodeInput}` and `{CodeOutput}` can be used or, equivalently, the environments `{Sinput}` and `{Soutput}` as produced by `Sweave()` or **knitr** when using the `render_sweave()` hook. Please make sure that all code is properly spaced, e.g., using `y = a + b * x` and *not* `y=a+b*x`. Moreover, code input should use "the usual" command prompt in the respective software system. For R code, the prompt `"R> "` should be used with `"+ "` as the continuation prompt. Generally, comments within the code chunks should be avoided – and made in the regular LaTeX text instead. Finally, empty lines before and after code input/output should be avoided (see above).

As a first model for the `quine` data, we fit the basic Poisson regression model. (Note that JSS prefers when the second line of code is indented by two spaces.)

```
R> m_pois <- glm(Days ~ (Eth + Sex + Age + Lrn)^2, data = quine,
+    family = poisson)
```

To account for potential overdispersion we also consider a negative binomial GLM.

```
R> library("MASS")
R> m_nbin <- glm.nb(Days ~ (Eth + Sex + Age + Lrn)^2, data = quine)
```

In a comparison with the BIC the latter model is clearly preferred.

```
R> BIC(m_pois, m_nbin)

       df      BIC
m_pois 18 2046.851
m_nbin 19 1157.235
```

Hence, the full summary of that model is shown below.

```
R> summary(m_nbin)

Call:
glm.nb(formula = Days ~ (Eth + Sex + Age + Lrn)^2, data = quine,
    init.theta = 1.60364105, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0857  -0.8306  -0.2620   0.4282   2.0898

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.00155    0.33709   8.904  < 2e-16 ***
```

```
EthN         -0.24591    0.39135  -0.628  0.52977
SexM         -0.77181    0.38021  -2.030  0.04236 *
AgeF1        -0.02546    0.41615  -0.061  0.95121
AgeF2        -0.54884    0.54393  -1.009  0.31296
AgeF3        -0.25735    0.40558  -0.635  0.52574
LrnSL         0.38919    0.48421   0.804  0.42153
EthN:SexM     0.36240    0.29430   1.231  0.21818
EthN:AgeF1   -0.70000    0.43646  -1.604  0.10876
EthN:AgeF2   -1.23283    0.42962  -2.870  0.00411 **
EthN:AgeF3    0.04721    0.44883   0.105  0.91622
EthN:LrnSL    0.06847    0.34040   0.201  0.84059
SexM:AgeF1    0.02257    0.47360   0.048  0.96198
SexM:AgeF2    1.55330    0.51325   3.026  0.00247 **
SexM:AgeF3    1.25227    0.45539   2.750  0.00596 **
SexM:LrnSL    0.07187    0.40805   0.176  0.86019
AgeF1:LrnSL -0.43101    0.47948  -0.899  0.36870
AgeF2:LrnSL  0.52074    0.48567   1.072  0.28363
AgeF3:LrnSL       NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.6036) family taken to be 1)

    Null deviance: 235.23  on 145  degrees of freedom
Residual deviance: 167.53  on 128  degrees of freedom
AIC: 1100.5

Number of Fisher Scoring iterations: 1


              Theta:  1.604
          Std. Err.:  0.214

 2 x log-likelihood:  -1062.546
```

# 5. Summary and discussion

- As usual …

# Computational details

> If necessary or useful, information about certain computational details such as version numbers, operating systems, or compilers could be included in an unnumbered section. Also, auxiliary packages (say, for visualizations, maps, tables, ...) that are not cited in the main text can be credited here.

The results in this paper were obtained using R 3.4.1 with the **MASS** 7.3.47 package. R itself and all packages used are available from the Comprehensive R Archive Network (CRAN) at https://CRAN.R-project.org/.

## Acknowledgments

> All acknowledgments (note the AE spelling) should be collected in this unnumbered section before the references. It may contain the usual information about funding and feedback from colleagues/reviewers/etc. Furthermore, information such as relative contributions of the authors may be added here (if any).

# A. More technical details

Appendices can be included after the bibliography (with a page break). Each section within the appendix should have a proper section title (rather than just *Appendix*).

For more technical style details, please check out JSS's style FAQ at `https://www.jstatsoft.org/pages/view/style#frequently-asked-questions` which includes the following topics:

- Title vs. sentence case.
- Graphics formatting.
- Naming conventions.
- Turning JSS manuscripts into R package vignettes.
- Trouble shooting.
- Many other potentially helpful details...

# B. Using BibTEX

References need to be provided in a BIBTEX file (`.bib`). All references should be made with `\cite`, `\citet`, `\citep`, `\citealp` etc. (and never hard-coded). This commands yield different formats of author-year citations and allow to include additional details (e.g., pages, chapters, ...) in brackets. In case you are not familiar with these commands see the JSS style FAQ for details.

Cleaning up BIBTEX files is a somewhat tedious task – especially when acquiring the entries automatically from mixed online sources. However, it is important that informations are complete and presented in a consistent style to avoid confusions. JSS requires the following format.

- JSS-specific markup (`\proglang`, `\pkg`, `\code`) should be used in the references.
- Titles should be in title case.
- Journal titles should not be abbreviated and in title case.
- DOIs should be included where available.
- Software should be properly cited as well. For R packages `citation("pkgname")` typically provides a good starting point.

**Affiliation:**
Xiaohua Zhou
Professor
Beijing International Center for Mathematical Research
Peking University
5 Yiheyuanlu Road
100871 Beijing, China
E-mail: azhou@bicmr.pku.edu.cn
URL: http://faculty.washington.edu/azhou

*Journal of Statistical Software* http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd