A note on marginal structural models with continuous treatment

Yen-Chi Chen
University of Washington
July 23, 2024

Useful references:

> Richardson, B.D., Blette, B.S., Gilbert, P.B., & Hudgens, M.G. (2024). Addressing Confounding and Continuous Exposure Measurement Error Using Corrected Score Functions. Arxiv: 2407.09443

Marginal structural models (MSMs) are methods in the causal inference and are popular for handling a large size of potential outcomes. Let A be a treatment variable that could be either continuous or even multivariate. Let $Y \in \mathbb{R}$ be the outcome variable of interest.

The treatment variable leads to a potential outcome $Y(a)$. In the simplest binary treatment case, $a \in \{0,1\}$, so there are only two potential outcomes $Y(1), Y(0)$. However, if there are multiple treatments, say $a \in \{0,1\}^K$, there will be 2^K potential outcomes, a large number of potential outcomes. In the case of continuous treatment, $a \in \mathbb{R}$, and there will be infinite number of potential outcomes $Y(a)$.

Thus, in multiple treatment problems (often occur in dynamic treatment scenarios) or continuous treatment scenarios, the potential outcomes are not easy to work with.

The MSM attempts to solve this problem by imposing a parametric model on the *marginal expectation*:

$$\mathbb{E}(Y(a)) = \theta(a;\gamma), \tag{1}$$

where γ is the parameter of the model.

Examples:

1. *Linear Model.* $\mathbb{E}(Y(a)) = \theta(a;\gamma) = \gamma_0 + \gamma_1^T a$.
2. *Log-linear model.* $\log \mathbb{E}(Y(a)) = \gamma_0 + \gamma_1^T a$ or equivalently,
$$\theta(a;\gamma) = \exp(\gamma_0 + \gamma_1^T a).$$
3. *Logistic model.* For $Y \in [0,1]$, or a binary Y, we may use $\log\left\{\frac{\mathbb{E}(Y(a))}{1-\mathbb{E}(Y(a))}\right\} = \gamma_0 + \gamma_1^T a$ or equivalently,
$$\theta(a;\gamma) = \frac{\exp(\gamma_0 + \gamma_1^T a)}{1 - \exp(\gamma_0 + \gamma_1^T a)}.$$

In this note, we will consider the case of having K continuous treatments, i.e., $a \in \mathbb{R}^K$. For simplicity, we will assume that the parameter in the MSM $\gamma \in \mathbb{R}^d$.

1 Data and assumptions

We assume that we observe all the confounders $L \in \mathbb{R}^p$. Thus, our data are the triplets:

$$(Y_1, A_1, L_1), \cdots, (Y_n, A_n, L_n),$$

where $Y_i \in \mathbb{R}$, $A_i \in \mathbb{R}^K$, and $L_i \in \mathbb{R}^p$.

For identifications, we will assume the following conditions:

- **A1: Consistency.** Given $A = a$, the observed outcome Y is the same as $Y(a)$.
- **A2: Ignorability.** For any a, we assume that

$$Y(a) \perp A | L.$$

2 Inverse probability weighting

Under equation (1), an estimator $\widehat{\gamma}$ implies $\widehat{\theta}(a) = \theta(a; \widehat{\gamma})$. So we only need to estimate parameter γ.

A popular approach of estimating γ is the inverse probability weighting (IPW). In the MSM, this is related to the following Z-estimator equation:

$$\begin{aligned} \Psi(\gamma; Y, A, L) &= \frac{Y - \theta(A; \gamma)}{p(A|L)} \cdot h(A), \\ \mathbb{E}(\Psi(\gamma; Y, A, L)) &= 0, \end{aligned} \tag{2}$$

where $h(a) \in \mathbb{R}^d$ is any measurable function of A. Note that $h(a)$ has to be a vector of d functions to ensure the Z-equation leads to a feasible solution of γ.

We will show that if the MSM is correct, then the true parameter γ^* solves equation (2).

Theorem 1 *Under (A1-2), and assume that* $\mathbb{E}(Y(a)) = \theta(a; \gamma^*)$ *and* $\frac{h(a)}{p(a)} \neq 0$ *for all* a *almost surely for* P_A*, the probability measure of* A*. Then* γ^* *solves equation* (2)*.*

Proof. The proof is based on the following trick–if we can show that $\mathbb{E}(\Psi(\gamma^*; Y, A, L)|A = a) = 0$ for all a, then based on the law of total expectation, we have $\mathbb{E}(\Psi(\gamma^*; Y, A, L)) = 0$.

A direct derivation shows

$$\begin{aligned}
\mathbb{E}(\Psi(\gamma;Y,A,L)|A=a) &= \mathbb{E}\left\{\frac{Y-\theta(a;\gamma)}{p(a|L)}|A=a\right\}h(a) \\
&= \mathbb{E}\left\{\frac{\mathbb{E}[Y|A=a,L]-\theta(a;\gamma)}{p(a|L)}|A=a\right\}h(a) \\
&\overset{(A1)}{=} \mathbb{E}\left\{\frac{\mathbb{E}[Y(a)|A=a,L]-\theta(a;\gamma)}{p(a|L)}|A=a\right\}h(a) \\
&\overset{(A2)}{=} \mathbb{E}\left\{\frac{\mathbb{E}[Y(a)|L]-\theta(a;\gamma)}{p(a|L)}|A=a\right\}h(a) \\
&= h(a)\times\int\frac{\mathbb{E}(Y(a)|L=\ell)-\theta(a;\gamma)}{p(a|\ell)}p(\ell|a)d\ell \\
&= h(a)\times\int[\mathbb{E}(Y(a)|L=\ell)-\theta(a;\gamma)]p(\ell)d\ell/p(a) \\
&= \frac{h(a)}{p(a)}\times[\mathbb{E}(Y(a))-\theta(a;\gamma)].
\end{aligned}$$

Thus, if $\mathbb{E}(Y(a))=\theta(a;\gamma^*)$, then the above equation is 0 as long as $\frac{h(a)}{p(a)}$ is non-zero.

□

Based on Theorem 1, we will want to include $p(a)$ into $h(a)$, i.e., $h(a)=p(a)\omega(a)$ so that the ratio $\frac{h(a)}{p(a)}=\omega(a)$ is non-zero.

2.1 Stabilized estimator

In the above derivation, we still need to choose $\omega(a)\in\mathbb{R}^d$, a d-dimensional vector of functions. A popular choice is

$$\omega(a)=\frac{\partial}{\partial\gamma}\theta(a;\gamma^*)\equiv s(a;\gamma^*),$$

or equivalently,

$$h(a)=p(a)\cdot s(a;\gamma^*). \tag{3}$$

This is called the stabilized weight.

Why do we want to choose $\omega(a)=s(a;\gamma^*)$? First, the partial derivative (like a score function) satisfies $s(a;\gamma^*)=\frac{\partial}{\partial\gamma}\theta(a;\gamma^*)\in\mathbb{R}^d$. Moreover, this choice leads to an elegant form when we analyze the variance of the resulting estimator.

To see this, let $\widehat{\gamma}_n$ be the solution of the IPW estimating equations:

$$0=\frac{1}{n}\sum_{i=1}^{n}\Psi(\widehat{\gamma}_n;Y_i,A_i,L_i)\equiv L_n(\widehat{\gamma}_n),\qquad \Psi(\gamma;Y,A,L)=\frac{Y-\theta(A;\gamma)}{p(A|L)}\cdot h(A). \tag{4}$$

$\widehat{\gamma}_n$ is called *stabilized estimator*.

Note that the true parameter γ^* satisfies

$$0 = \mathbb{E}(\Psi(\gamma^*; Y, A, L)) \equiv L(\gamma^*)$$

by assumption. Using the Taylor expansion, we have

$$\begin{aligned}
L_n(\gamma^*) - \underbrace{L(\gamma^*)}_{=0} &= L_n(\gamma^*) - \underbrace{L_n(\widehat{\gamma}_n)}_{=0} \\
&\approx (\gamma^* - \widehat{\gamma}_n)^T \frac{\partial}{\partial \gamma} L_n(\gamma^*) \\
&\approx (\gamma^* - \widehat{\gamma}_n)^T \frac{\partial}{\partial \gamma} L(\gamma^*) \\
&= -(\gamma^* - \widehat{\gamma}_n)^T \frac{\partial}{\partial \gamma} \mathbb{E}\left[\frac{\theta(A;\gamma^*)}{p(A|L)} h(A) \right] \\
&= -(\gamma^* - \widehat{\gamma}_n)^T \mathbb{E}\left[\underbrace{\frac{\partial}{\partial \gamma} \theta(A;\gamma^*)}_{= s(A;\gamma^*)} \frac{h(A)}{p(A|L)} \right] \\
&= \mathbb{E}\left\{ \frac{h(A)}{p(A|L)} s(A;\gamma^*)^T \right\} (\widehat{\gamma}_n - \gamma^*).
\end{aligned}$$

Thus,

$$\widehat{\gamma}_n - \gamma^* = \mathbb{E}\left\{ \frac{h(A)}{p(A|L)} s(A;\gamma^*)^T \right\}^{-1} L_n(\gamma^*). \tag{5}$$

Now we investigate the covariance matrix of L_n. Recall that

$$L_n(\gamma) = \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i - \theta(A_i;\gamma)}{p(A_i|L_i)} \cdot h(A_i).$$

Using the fact that it has mean 0, so $\mathsf{Cov}(L_n(\gamma)) = \mathbb{E}(L_n(\gamma) L_n(\gamma)^T)$. Thus, its covariance matrix will be

$$\mathsf{Cov}(L_n(\gamma)) = \frac{1}{n} \mathbb{E}\left\{ \left[\frac{Y - \theta(A;\gamma)}{p(A|L)} \right]^2 h(A) h(A)^T \right\}. \tag{6}$$

By comparing equations (5) and (6), one can see that the choice

$$h(a) = p(a) \frac{\partial}{\partial \gamma} \theta(a;\gamma^*) = p(a) s(a;\gamma^*)$$

leads to

$$\begin{aligned}
\widehat{\gamma}_n - \gamma^* &= \mathbb{E}\left\{ \frac{p(A)}{p(A|L)} s(A;\gamma^*) s(A;\gamma^*)^T \right\}^{-1} L_n^*(\gamma^*), \\
\mathsf{Cov}(L_n^*(\gamma^*)) &= \frac{1}{n} \mathbb{E}\left\{ (Y - \theta(A;\gamma))^2 \frac{p^2(A)}{p^2(A|L)} s(A;\gamma^*) s(A;\gamma^*)^T \right\},
\end{aligned}$$

which is an elegant form of the covariance matrix.

2.2 Nuisance parameters

To use the IPW estimator under equation (4), we need to estimate two nuisance functions: $p(a)$ and $p(a|\ell)$. The first nuisance parameter $p(a)$ is the marginal density of the treatment A, which can be done by either a parametric model (when the number of treatment K is large) or a nonparametric model (when the number of treatment K is small).

The second parameter $p(a|\ell)$ is the conditional density of treatment given the confounders. It is a function of a total of $K+p$ variables, which could be hard to estimate nonparametrically. Thus, researchers often use a parametric model for this density. For instance, if $A \in \mathbb{R}^K$ are all continuous, one may assume that

$$A|L \sim N(g(L;\tau), \Sigma(L;\rho)),$$

where $g(L;\tau) \in \mathbb{R}^K$ and $\Sigma(L;\rho) \in \mathbb{R}^{K\times K}$ are parametric functions indexed by the parameters τ and ρ.

Fortunately, the two nuisances $p(a), p(a|\ell)$ are identifiable from the data. So we can directly estimate them.

To summarize, the IPW approach is the following procedure.

1. Estimate $\widehat{p}(a), \widehat{p}(a|\ell)$ from $(A_1, L_1), \cdots, (A_n, L_n)$.
2. Construct the estimating equation
$$\widehat{\Psi}(\gamma; Y, A, L) = \frac{\widehat{p}(A)}{\widehat{p}(A|L)}(Y - \theta(A;\gamma))s(A;\gamma), \qquad s(a;\gamma) = \frac{\partial}{\partial\gamma}\theta(a;\gamma).$$
3. Find $\widehat{\gamma}_n$ by solving
$$0 = \frac{1}{n}\sum_{i=1}^{n} \widehat{\Psi}(\widehat{\gamma}_n; Y_i, A_i, L_i).$$
4. Obtain the estimator $\widehat{\theta}(a) = \theta(a;\widehat{\gamma}_n)$.

Note that solving the equation in the third step could be non-trivial since γ appears in two terms: $\theta(a;\gamma)$ and $s(a;\gamma)$. One may use an iterative procedure to solve find a solution by solving the following equation

$$\widehat{\gamma}^{(t+1)} : 0 = \frac{1}{n}\sum_{i=1}^{n} \frac{\widehat{p}(A_i)}{\widehat{p}(A_i|L_i)}(Y_i - \theta(A_i;\gamma))s(A_i;\widehat{\gamma}^{(t)}).$$

Namely, we use the previous iteration's value $\widehat{\gamma}^{(t)}$ in the 'score' $s(A_i;\widehat{\gamma}^{(t)})$, so only the $\theta(a;\gamma)$ will need to be solved.

3 Regression adjustment (g-computation)

In causal inference, an alternative approach to the IPW is the regression adjustment (also known as the g-computation). In the MSM, the regression adjustment does not directly use equation (1). Instead, the regression adjustment models the marginal conditional mean:

$$\mathbb{E}(Y(a)|A = a, L = \ell) = \mu(a,\ell) = \mu(a,\ell;\beta), \tag{7}$$

where β is the parameter in the model. Under assumptions (A1-2), this is equivalent to

$$\mathbb{E}(Y|A,L) = \mu(A,L;\beta).$$

Note that the model in equation (7) and the model in equation (1) are variationally dependent.

Assumption(A1-2) implies the following connection:

$$\begin{aligned}\theta(a) &\equiv \mathbb{E}(Y(a)) \\ &= \mathbb{E}(\mathbb{E}(Y(a)|L)) \\ &= \mathbb{E}(\mathbb{E}(Y(a)|L,A=a)) \\ &= \mathbb{E}(\mu(a,L)).\end{aligned}$$

Therefore, the model $\mu(a,\ell;\beta)$ implies a model on equation (1) that

$$\mathbb{E}(Y(a)) = \mathbb{E}[\mu(a,L;\beta)].$$

The regression adjustment is very easy to use because the regression function

$$\mathbb{E}(Y|A,L) = \mu(A,L;\beta)$$

is the canonical regression problem.

3.1 Estimator of the MSM

In general, we obtain an estimator $\widehat{\beta}_n$ by a least square approach of $\mathbb{E}(Y|A,L) = \mu(A,L;\beta)$ or a risk minimization method. For the least square,

$$\widehat{\beta}_n = \mathsf{argmin}_\beta \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu(A_i,L_i;\beta))^2.$$

Then we construct the estimator of the MSM via

$$\widehat{\theta}(a) = \frac{1}{n}\sum_{i=1}^{n}\mu(a,L_i;\widehat{\beta}_n).$$

Note that one may combine the above two equations into a chained estimating equation:

$$\begin{aligned}0 &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu(A_i,L_i;\beta))\frac{\partial}{\partial\beta}\mu(A_i,L_i;\beta), \\ 0 &= \frac{1}{n}\sum_{i=1}^{n}[\theta(a) - \mu(a,L_i;\beta)]\end{aligned}$$

and obtain $\widehat{\beta}_n$ and $\widehat{\theta}(a)$ by solving the above equations.

3.2 Model congeniality

The model we impose on the regression adjustment may not imply a nice MSM. Thus, there could be model congeniality issues here. However, there are two special cases where the regression adjustment model and the corresponding MSM are congenial.

Linear model. Under the linear model,

$$\mathbb{E}(Y|A,L) = \mu(A,L;\beta) = \beta_A^T A + \beta_L^T L + \beta_0$$

implies that

$$\mathbb{E}(Y(a)) = \underbrace{\beta_A^T}_{\gamma_1} a + \underbrace{\beta_L^T \mathbb{E}(L) + \beta_0}_{\gamma_0},$$

so the resulting MSM is also linear.

Log-Linear model without interactions. Suppose that

$$\log \mathbb{E}(Y|A,L) = \beta_A^T A + \beta_L^T L + \beta_0.$$

This implies that

$$\mathbb{E}(Y|A,L) = \exp(\beta_A^T A) \cdot \exp(\beta_L^T L + \beta_0).$$

Thus,

$$\mathbb{E}(Y(a)) = \mathbb{E}(\mathbb{E}(Y|a,L)) = \exp(\underbrace{\beta_A^T}_{=\gamma_1} a) \cdot \underbrace{\mathbb{E}[\exp(\beta_L^T L)] \cdot \exp(\beta_0)}_{=\exp(\gamma_0)}$$

is still log-linear. Note that we can allow within A interactions and within L interactions–the only interactions that will be ruled out is the interaction between A and L.