A note on semi-parametric estimators

Yen-Chi Chen
University of Washington
May 5, 2020

Contents

Semi-parametric model offers a set of flexible tools in data analysis that has a fast convergence rate. The high-level idea is that we want to estimate a parameter of interest that is a finite dimensional object but we do not place any parametric constraint on the full model.

To see why this is interesting, consider density estimation problem where we observe $X_1, \cdots, X_n \sim p$. To estimate the underlying PDF p, we may use kernel density estimator, histogram, basis approach, ...etc. However, we cannot estimate p in a fast rate; under usual assumption (known as the Hölder condition, which is similar to the condition that p is bounded twice-differentiable), the best rate (in terms of minimaxity) that we can achieve is $\sup_x |\widehat{p}(x) - p(x)| = O_P(n^{-\frac{2}{4+d}})$, which is clearly slower than the usual parametric rate $O_P(n^{-\frac{1}{2}})$. Note that d is the dimension of X. On the other hand, if we assume that the distribution is a normal distribution, we can estimate the mean and variance at rate $O_P(n^{-\frac{1}{2}})$ so we can achieve that convergence rate $\sup_x |\widehat{p}(x) - p(x)| = O_P(n^{-\frac{1}{2}})$.

The power of semi-parametric estimator is that there are cases where we can achieve a parametric rate without assuming that the entire model is a parametric model. For instance, consider estimating the population mean $\mu = \mathbb{E}(X)$ where $X \in \mathbb{R}^d$. Clearly, the sample mean

$$\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i$$

is a $\sqrt{n}$-consistent estimator. But note that the consistency of $\widehat{\mu}_n$ does not require any parametric assumption on the entire distribution. For another example, consider the linear regression problem where we model (assuming no intercept)

$$Y_i = \beta^T X_i + \varepsilon_i.$$

The least square estimator $\widehat{\beta}_n$ from minimizing

$$R_n(\beta) = \sum_{i=1}^n (Y_i - \beta^T X_i)^2$$

also converges to the true parameter (assuming that the linear model is correct) even if we do not specify the distribution of X and $\varepsilon|X$ to be parametric models.

In this note, we give a gentle introduction of semi-parametric estimators. We start with a properties of a parametric model that are relevant to semi-parametric models and then discuss how to construct a semi-parametric estimator. In particular, we will discuss three approaches–the first one is based on characterizing the tangent space, the second one is based on the geometry of influence function, and the third one is based on conditional expectation.

The first part of the note is mostly based on Chapter 3 and 4 of the following book:

> Tsiatis, A. (2007). Semiparametric theory and missing data. Springer Science & Business Media.

The later part of the note will be from Chapter 25 of the following book:

> Van der Vaart, A. W. (2000). Asymptotic statistics (Vol. 3). Cambridge university press.

1 Influence function and score vector

Consider a simple parametric model where we observe

$$X_1, \cdots, X_n \sim p(x; \theta_0),$$

where θ_0 is the underlying parameter. Assume that we can decompose the parameter $\theta_0 = (\beta_0, \eta_0)$ such that our primary interest is $\beta \in \mathbb{R}^q$. In this case, $\eta \in \mathbb{R}^r$ will be called a nuisance parameter.

An estimator $\widehat{\beta}_n$ of β_0 is called *asymptotically linear* with *influence function* $\psi(x)$ if

$$\sqrt{n}(\widehat{\beta}_n - \beta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(X_i) + o_P(1)$$

and $\mathbb{E}(\psi(X)) = 0$ and $\mathbb{E}(\psi(X)\psi(X)^T)$ is finite and non-singular. Note that here we did not specify how to construct $\widehat{\beta}_n$; there is often many asymptotically linear estimator of the same parameter of interest (each corresponds to a different influence function).

Since we are using a parametric model, the score vector is often of a lot of interest. Let

$$\begin{aligned} S_\theta(x;\theta) &= \frac{\partial}{\partial \theta} \log p(x;\theta) \in \mathbb{R}^{q+r} \\ S_\beta(x;\theta) &= \frac{\partial}{\partial \beta} \log p(x;\theta) \in \mathbb{R}^{q} \\ S_\eta(x;\theta) &= \frac{\partial}{\partial \eta} \log p(x;\theta) \in \mathbb{R}^{r} \end{aligned}$$

be the score vectors.

From the theory of MLE (maximum likelihood estimator), we know that the score vector often plays a key role in the construction of an estimator. So one may be wondering how will the score vector and the influence function related. The following theorem provides a powerful link between them.

Theorem 1 (Theorem 3.2 of Tsiatis 2007) *If $\widehat{\beta}_n$ is a regular[1] estimator, then*

$$\mathbb{E}(\psi(X) S_\beta^T(X;\theta_0)) = \mathbf{I}_{q\times q}, \quad \mathbb{E}(\psi(X) S_\eta^T(X;\theta_0)) = \mathbf{0}_{q\times r}, \tag{1}$$

where $\mathbf{I}_{q\times q}$ is the $q \times q$ identity matrix.

The power of Theorem 1 is that it works for any regular asymptotic normal (RAL) estimator! It does not require any specific construction of the estimator. For instance, if we use the method of moments to construct $\widehat{\beta}_n$, Theorem 1 will apply.

Equation (1) is not just a necessarily condition for an RAL estimator. If we have a function ψ that satisfies equation (1), we can construct an RAL estimator with influence function being ψ.

[1] in general, we are often using a regular estimator; the precise definition can be found in Definition 1 of Tsiatis 2007.

Theorem 2 (Converting influence function into an estimator) *Let $\psi(X)$ be a function satisfying equation (1). Assume that for each β, we have an estimator $\widehat{\eta}_n(\beta)$ such that $\sqrt{n}\|\widehat{\eta}_n(\beta)-\eta_0\|_{\max}$ is bounded in probabiiltiy. Define $m(X;\beta,\eta)=\psi(X)-\mathbb{E}_{X\sim p(\cdot;\beta,\eta)}(\psi(X))$ and let $\widehat{\beta}$ be the solution of*

$$\sum_{i=1}^{n} m(X_i;\beta,\widehat{\eta}_n(\beta))=0.$$

Then $\widehat{\beta}_n$ will be an RAL estimator with influence function $\psi(X)$.

The proof of this theorem can be found in page 39-40 of Tsiatis (2007).

Theorem 2 provides a procedure to convert a function satisfying equation (1) into an estimator with that function being an influence function. With Theorem 1, we can informally say that the set/space

$$\mathcal{G}=\left\{\psi:\mathbb{E}(\psi(X)S_\beta^T(X;\theta_0))=\mathbf{I}_{q\times q},\quad \mathbb{E}(\psi(X)S_\eta^T(X;\theta_0))=\mathbf{0}_{q\times r}\right\}$$

characterizes all RAL estimators. In particular, the second equality $\mathbb{E}(\psi(X)S_\eta^T(X;\theta_0))=\mathbf{0}_{q\times r}$ can be re-written as

$$\Pi(\psi(X)|\mathcal{F}_\eta)=0,$$

where $\mathcal{F}_\eta$ is the nuisance tangent space (will be introduced later).

2 Efficient influence function

Now we have learned that there can be many RAL estimators. One may be wondering: what is the optimal RAL estimator? how do we construct the optimal RAL estimator? Here the best is often referred to the estimator with the smallest variance (since an RAL estimator is asymptotically unbiased). To answer these questions, we go back to conditions in equation (1).

The condition

$$\mathbb{E}(\psi(X)S_\eta^T(X;\theta_0))=\mathbf{0}_{q\times r}$$

is particularly interesting since the conditional mean being 0 is often related to orthogonal projection. To further investigate this, we introduce the *nuisance tangent space*

$$\mathcal{F}_\eta=\{BS_\eta(x;\theta_0):B\in\mathbb{R}^{q\times r}\}.$$

Namely, $\mathcal{F}_\eta$ is a collection of functions that is along the direction of the score vector $S_\eta(x;\theta_0)$ and projected into the space of β (the effect of the coefficient matrix B). Similarly, we define

$$\mathcal{F}_\beta=\{BS_\beta(x;\theta_0):B\in\mathbb{R}^{q\times q}\}$$
$$\mathcal{F}=\{BS_\theta(x;\theta_0):B\in\mathbb{R}^{q\times(q+r)}\}$$

and we can write $\mathcal{F}=\mathcal{F}_\beta\oplus\mathcal{F}_\eta$ with the notation

$$\mathcal{F}_1\oplus\mathcal{F}_2=\{f_1+f_2:f_1\in\mathcal{F}_1,f_2\in\mathcal{F}_2\}.$$

$\mathcal{F}$ is called the tangent space and it has a very interesting property–it characterizes the space of all RAL estimators!

Theorem 3 (Theorem 3.4 of Tsiatis 2007) *Let ψ_1, ψ_2 be any two influence functions of RAL estimators of β_0. Then*

$$\psi_1 - \psi_2 \in \mathcal{F}^{\perp}.$$

Namely,

$$\mathbb{E}((\psi_1(X) - \psi_2(X))S_\theta^T(X;\theta_0)) = \mathbf{0}_{q\times(q+r)}.$$

In other words, Theorem 3 shows that if we can find an influence function ψ, then we can find all other influence functions by adding a term from the orthogonal space of the tangent space. Together with Theorem 1 and 2, this implies that we have a way to characterize all RAL estimators!

This is a powerful result for our purposes because the search of the optimal RAL estimator can be done by examining the tangent space if we have obtained an influence function. The influence function that corresponds to the RAL estimator with the smallest variance is called *efficient influence function*.

Theorem 3 implies that we will be working on estimator of the form:

$$\psi(x) = \psi_0(x) + g(x), \quad g \in \mathcal{F}_\eta,$$

where ψ_0 is a given function. Thus, the variance of the estimator constructed by ψ will be written as

$$\mathsf{Var}(\psi(X)) = \mathsf{Var}(\psi_0(X) + g(X)).$$

This quantity is not easy to analyze since it is a covariance matrix and $\psi_0(X)$ and $g(X)$ may be highly correlated. However, here is an interesting note from the Pythagorean theorem. For a function $h(x) \in \mathbb{R}^q$, denote its projection onto a subspace $\mathcal{G}$ as $\Pi(h(x)|\mathcal{G})$. Then

$$\mathsf{Var}(h(X)) = \mathsf{Var}(\Pi(h(X)|\mathcal{G})) + \mathsf{Var}(h(X) - \Pi(h(X)|\mathcal{G})) \geq \mathsf{Var}(\Pi(h(X)|\mathcal{G})).$$

Note that for matrices $A \geq B$, this means that $A - B$ is positive definite. From the above inequality, one can easily deduce the following result:

Theorem 4 (Theorem 3.5 of Tsiatis 2007) *Let ψ_0 be any influence function satisfying equation (1). The the efficient influence function can be constructed using*

$$\psi_{\mathsf{eff}}(X) = \psi_0(X) - \Pi(\psi_0(X)|\mathcal{F}^{\perp}) = \Pi(\psi_0(X)|\mathcal{F}).$$

In general, obtaining a projection is not easy. But here the space we are considering is a linear space so the projection can be easily defined. Take the tangent space as an example. For a function $h(x) \in \mathbb{R}^q$, its projection onto $\mathcal{F}$ is

$$\Pi(h(x)|\mathcal{F}) = \mathbb{E}(h(X)S_\theta^T(X;\theta_0))[\mathbb{E}(S_\theta(X;\theta_0)S_\theta^T(X;\theta_0))]^{-1}S_\theta(x;\theta_0).$$

Thus, for any influence function ψ_0 satisfying equation (1),

$$\psi_{\mathsf{eff}}(X) = \mathbb{E}(\psi_0(X)S_\theta^T(X;\theta_0))[\mathbb{E}(S_\theta(X;\theta_0)S_\theta^T(X;\theta_0))]^{-1}S_\theta(x;\theta_0).$$

2.1 Constructing efficient influence function using score vectors

The above procedure is useful when we have an influence function already. In practice, we may not have an influence function so how to construct an efficient influence function is unclear. Here is a simple approach that based on the score vectors.

We define the *efficient score vector* as

$$\begin{aligned} S_{\text{eff}}(X;\theta_0) &= \Pi(S_\beta(X;\theta_0)|\mathcal{F}_\eta^\perp) \\ &= S_\beta(X;\theta_0) - \Pi(S_\beta(X;\theta_0)|\mathcal{F}_\eta) \\ &= S_\beta(X;\theta_0) - \mathbb{E}(S_\beta(X;\theta_0)S_\eta^T(X;\theta_0))[\mathbb{E}(S_\eta(X;\theta_0)S_\eta^T(X;\theta_0))]^{-1}S_\eta(X;\theta_0) \end{aligned} \tag{2}$$

By construction, $S_{\text{eff}}(X;\theta_0)$ satisfies $\mathbb{E}(S_{\text{eff}}(X;\theta_0)S_\eta^T(X;\theta_0)) = \mathbf{0}_{q\times r}$ so it satisfies the second equality in equation (1).

To ensure the first equality in equation (1), using the fact that

$$\mathbb{E}(S_{\text{eff}}(X;\theta_0)S_\beta^T(X;\theta_0)) = \mathbb{E}(S_{\text{eff}}(X;\theta_0)S_{\text{eff}}^T(X;\theta_0)),$$

we construct the influence function as

$$\psi_{\text{score}}(X) = [\mathbb{E}(S_{\text{eff}}(X;\theta_0)S_{\text{eff}}^T(X;\theta_0))]^{-1}S_{\text{eff}}(X;\theta_0).$$

One can verify that $\psi_{\text{score}}(X)$ satisfies equation (1).

Because $S_{\text{eff}}(X;\theta_0) \in \mathcal{F}_\beta \oplus \mathcal{F}_\eta = \mathcal{F}$, it is easy to see that $\Pi(\psi_{\text{score}}(X)|\mathcal{F}) = \psi_{\text{score}}(X)$ so $\psi_{\text{score}}(X) = \psi_{\text{eff}}(X)$! Moreover, the minimal variance (known as the efficiency bound) will be

$$\text{Var}(\psi_{\text{eff}}(X)) = [\mathbb{E}(S_{\text{eff}}(X;\theta_0)S_{\text{eff}}^T(X;\theta_0))]^{-1}.$$

3 Semiparametric estimator

The above result is very powerful–we know how to construct the optimal (efficient) estimator. However, it is also restrictive in the sense that it requires a parametric model. In many scenarios, we are interested in $\beta \in \mathbb{R}^q$ that is finite dimensional but we do not want to limit ourselves to a parametric approach of η. Here we will generalize the techniques in the previous section so that they work even if the nuisance parameter η may be infinite dimensional. In this case, the problem is called the semiparametric estimation problem.

Here are two common examples.

- **Moment restricted model.** Consider a regression problem where

  $$Y_i = \mu(X_i;\beta) + \varepsilon_i, \quad \mathbb{E}(\varepsilon_i|X_i) = 0,$$

  where μ is a known function (for instance, linear regression requires $\mu(X_i;\beta) = \alpha + X_i^T\beta$). The parameter of interest is β. The nuisance parameter is the PDF of $\varepsilon|X = x$ and the PDF of X.

- **Proportional hazard model.** The proportional hazard model is a very popular approach in survival analysis. The data consists of $(X_1, T_1), \cdots, (X_n, T_n)$ and let $F(t|x)$ be the CDF of T given $X = x$. The survival function $S(t|x) = 1 - F(t|x)$. The hazard function is defined as $\lambda(t|x) = -\frac{d}{dt} \log S(t|x)$. The proportional hazard model assumes that

$$\lambda(t|x;\beta) = \lambda_0(t) \exp(\beta^T x).$$

  Again, β is the parameter of interest and λ_0, the baseline hazard function, is the nuisance parameter.

In the above problems, the nuisance parameters are functions so they are infinite dimensional objects. The amazing part of the semiparametric inference is that we can still obtain a $\sqrt{n}$-rate estimator for β!

3.1 Parametric submodels

In a semi-parametric model, the data $X_1, \cdots, X_n$ is generated as IID random variables from the density

$$p(x;\beta_0,\eta_0),$$

where $\beta_0 \in \mathbb{R}^q$ and η_0 is an infinite dimensional object (a function). The collection

$$\mathcal{P} = \{p(x;\beta,\eta) : \beta_0 \in \mathbb{R}^q, \eta \text{ is infinite dimensional}\}$$

is called a *semiparametric model*.

When the nuisance parameter η is infinite dimensional, it is hard to draw connections to the parametric model. A key insight from semiparametric inference is the use of *parametric submodels*. A parametric submodel is

$$\mathcal{P}_\gamma = \{p(x;\beta,\gamma) : \beta_0 \in \mathbb{R}^q, \gamma \in \mathbb{R}^r\}$$

where the nuisance parameter is represented by a finite dimensional vector γ. So clearly, $\mathcal{P}_\gamma \subset \mathcal{P}$. But we will add an additional constraint that there exists γ_0 such that

$$p(x;\beta_0,\gamma_0) = p(x;\beta_0,\eta_0),$$

i.e., the parametric submodel includes the model that generates our data.

Take the proportional hazard model as an example, one possible parametric submodel is

$$\lambda(t|x;\beta,\gamma) = \lambda_0(t) \exp(\gamma^T h(t) + \beta^T x),$$

where $h(t) \in \mathbb{R}^r$ is a given vector-valued function. Suppose that the proportional hazard model is correct, then $\lambda(t|x;\beta,\gamma = 0)$ reduces back to the proportional hazard model so $\gamma = \gamma_0 = 0$ leads to the correct model.

Under a parametric submodel, the problem reduces back to the parametric model problem. We define the score vectors

$$S_\gamma(X;\beta_0,\gamma_0) = \frac{\partial}{\partial \gamma_0} \log p(X;\beta_0,\gamma_0), \quad S_\beta(X;\beta_0,\gamma_0) = \frac{\partial}{\partial \beta_0} \log p(X;\beta_0,\gamma_0).$$

And we have the following results from the parametric models (previous sections):

- **Submodel nuisance tangent space:**

$$\mathcal{F}_\gamma = \{BS_\gamma(x;\beta_0,\gamma_0) : B \in \mathbb{R}^{q\times r}\}. \tag{3}$$

- **Efficient score vector:**

$$\begin{aligned} S_{\gamma,\mathsf{eff}}(X;\beta_0,\gamma_0) =& \ \Pi(S_\beta(X;\beta_0,\gamma_0)|\mathcal{F}_\gamma^\perp) \\ &= S_\beta(X;\beta_0,\gamma_0) - \Pi(S_\beta(X;\beta_0,\gamma_0)|\mathcal{F}_\gamma) \\ &= S_\beta(X;\beta_0,\gamma_0) - \mathbb{E}(S_\beta(X;\beta_0,\gamma_0)S_\gamma^T(X;\beta_0,\gamma_0))[\mathbb{E}(S_\gamma(X;\beta_0,\gamma_0)S_\gamma^T(X;\beta_0,\gamma_0))]^{-1}S_\gamma(X;\beta_0,\gamma_0). \end{aligned}$$

- **Efficient influence function:**

$$\psi_{\gamma,\mathsf{eff}}(X) = \psi_{\mathsf{score}}(X) = [\mathbb{E}(S_{\gamma,\mathsf{eff}}(X;\beta_0,\gamma_0)S_{\gamma,\mathsf{eff}}^T(X;\beta_0,\gamma_0))]^{-1}S_{\gamma,\mathsf{eff}}(X;\beta_0,\gamma_0).$$

- **Efficiency bound:**

$$V_{0,\gamma} = \mathsf{Var}(\psi_{\gamma,\mathsf{eff}}(X)) = [\mathbb{E}(S_{\gamma,\mathsf{eff}}(X;\beta_0,\gamma_0)S_{\gamma,\mathsf{eff}}^T(X;\beta_0,\gamma_0))]^{-1}.$$

Note that all these quantities depend on the particular submodel.

3.2 Generalizing RAL estimators to a semi-parametric model

The efficiency bound is useful if we are thinking of RAL estimators. So we have to generalize the RAL estimator to semiparametric models. An estimator $\widehat{\beta}_n$ is called an RAL estimator of β_0 for *a semi-parametric model* if it is an RAL estimator *for all parametric submodels*.

Note that this definition implicitly limits the estimators being considered but it provides many useful properties. For one example, originally $\mathcal{P} \supset \mathcal{P}_\gamma$ but now

$$\Psi \equiv \{\text{influence functions of RAL estimators in } \mathcal{P}\} \subset \{\text{influence functions of RAL estimators in } \mathcal{P}_\gamma\} \equiv \Psi_\gamma.$$

This is because an RAL estimator in $\mathcal{P}$ must be an RAL estimator for any $\mathcal{P}_\gamma$. For another example, the above inclusion implies that if ψ is the influence function of an RAL estimator in $\mathcal{P}$, then

$$\mathsf{Var}(\psi(X)) \geq V_{0,\gamma}$$

for any parametric submodel. Thus, we can try to maximize the lower bound and obtain

$$V_0 \equiv \sup_{\{\text{all parametric submodels}\}} V_{0,\gamma}. \tag{4}$$

By construction,

$$\mathsf{Var}(\psi(X)) \geq V_0$$

and V_0 is called the *semi-parametric efficiency bound*.

The semi-parametric efficiency bound gives us a concrete target in the construction of the influence function. If we can find an influence function that achieves the bound, then we know that it is the most efficient one among all RAL estimators. In the next section, we will provide more description about this bound.

3.3 Semi-paramteric nuisance tangent space

The efficiency bound given in equation (4) is a bit abstract–it is defined via taking the supremum over all parametric submodels. Although it has nice property, it is unclear how do we describe this bound. Here we will introduce a way to explicitly describe V_0 using the *nuisance tangent space* of a semi-parametric model.

In equation (3), we describe the submodel nuisance tangent space as

$$\mathcal{F}_\gamma = \{BS_\gamma(x;\beta_0,\eta_0) : B \in \mathbb{R}^{q\times r}\}.$$

Note that here we denote the true nuisance parameter as η_0; by the construction of submodels, there is always an element of γ that η_0 belongs to the submodel. In the semi-parametric efficiency bound, we are using all possible submodels. Thus, we define the semi-parametric nuisance tangent space as the mean-square closure of $\mathcal{F}_\gamma$ of all submodels. Specifically, the *semi-parametric nuisance tangent space* is

$$\mathcal{F}_{\mathsf{nuis}} = \left\{h(X) : \mathbb{E}(h(X)h(X)^T) < \infty, \lim_{m\to\infty} \mathbb{E}\|h(X) - B_m S_{\gamma_m}(X;\beta_0,\eta_0)\|^2 = 0\right\},$$

where B_m is a sequence of $q \times r_m$ matrices and $S_{\gamma_m}(X;\beta_0,\gamma_0)$ is a sequence of score vectors corresponding to a sequence of submodels. In the above expression, the dimension of each submodel (in the sequence) and submodels are allowed to change with respect to m. Note that $\mathcal{F}_{\mathsf{nuis}}$ is a subset of the Hilbert space.

With the nuisance tangent space, we define the *semi-parametric efficient score* for β as

$$S_{\mathsf{eff}}(X;\beta_0,\eta_0) = S_\beta(X;\beta_0,\eta_0) - \Pi(S_\beta(X;\beta_0,\eta_0)|\mathcal{F}_{\mathsf{nuis}}), \tag{5}$$

which is just a simple generalization from equation (2). As can be expected, the efficient score implies the efficiency bound.

Theorem 5 (Theorem 4.1 of Tsiatis (2007)) *The semi-parametric efficiency bound*

$$V_0 = [\mathbb{E}(S_{\mathsf{eff}}(X;\beta_0,\eta_0)S_{\mathsf{eff}}^T(X;\beta_0,\eta_0))]^{-1}.$$

Theorem 5 provides a concrete characterization of the efficiency bound using the semi-paramteric nuisance tangent space. The power of semi-paramteric nuisance tangent space is beyond the efficiency bound. It has a similar property as the nuisance tangent space in a parametric model. Specifically, Theorem 1 can be generalized as follows.

Theorem 6 (Theorem 4.2 of Tsiatis (2007)) *Let $\psi(X)$ be the influence function of a semi-parametric RAL estimator of β_0. Then*

$$\mathbb{E}(\psi(X)S_\beta^T(X;\beta_0,\eta_0)) = \mathbf{I}_{q\times q}, \quad \Pi(\psi(X)|\mathcal{F}_{\mathsf{nuis}}) = 0,$$

The power of Theorem 6 is that if we have a description about the space $\mathcal{F}_{\mathsf{nuis}}^{\perp}$ (an influence function has to be an element in this space), we can use this description and the Z-estimator construction in Theorem 2 to construct an estimator of β.

One may be wondering if we can generalize Theorem 3 and efficient influence function from parametric models into the semi-parametric submodels. The answer is yes! To do so, we need to define the semi-parametric tangent space (not limited to the nuisance parameter). Using a similar construction as for the nuisance parameter, we define the semi-parametric tangent space $\mathcal{F}_{\mathsf{semi}}$ to be the mean-square closure of all tangent spaces of all parametric submodels (including those who can may change β). Note that

$$\mathcal{F}_{\mathsf{semi}} = \overline{\mathcal{F}_{\mathsf{nuis}} \oplus \mathcal{F}_{\beta}},$$

where $\mathcal{F}_{\beta}$ is the tangent space of the parameter of interest. The notation $\overline{\mathcal{F}}$ is referred to the mean square closure of $\mathcal{F}$. With this, Theorem 3 can be generalized as follows:

Theorem 7 (Theorem 4.3 of Tsiatis 2007) *Let ψ_1, ψ_2 be any two influence functions of semi-parametric RAL estimators of β_0. Then*

$$\psi_1 - \psi_2 \in \mathcal{F}_{\mathsf{semi}}^{\perp}.$$

With this, the efficient influence function will be

$$\psi_{\mathsf{eff}}(X) = \psi(X) - \Pi(\psi(X)|\mathcal{F}_{\mathsf{semi}}^{\perp}) = \Pi(\psi(X)|\mathcal{F}_{\mathsf{semi}}).$$

To sum up, as long as we have a way to characterize the mean-square closure of the parametric submodels (for both nuistance tangent space and full tangent space), we are able to derive the efficiency bound and construct a semi-parametric efficient estimator. As a result, most literature in semi-parametric inference will be focusing on how to characterize the space $\mathcal{F}_{\mathsf{nuis}}$ and $\mathcal{F}_{\mathsf{semi}}$.

4 Finding efficient estimators: tangent space approach

In the previous section, we have seen that the semi-parametric estimator has several desirable properties. The key to constructing a semi-parametric estimator is the characterization of the nuisance tangent space. If we have the nuisance tangent space, we can construct the efficient scores and obtain the efficiency bound. Here we will discuss some common strategies for finding a semi-parametric estimator.

The main strategy can be roughly divided into the following steps.

- **Step 1 (nuisance tangent space): finding a characterization of the nuisance tangent space.** We first try to derive a form of any element in the nuisance tangent space $\mathcal{F}_{\mathsf{nuis}}$. A common way to this derivation is to analyze how a particular model condition (e.g. moment condition, Cox model condition) constrained the tangent space.

- **Step 2 (nonparametric tangent space): finding a characterization of the non-parametric tangent space.** The non-parametric tangent space is the mean-squared closure of all submodel tangent spaces $\mathcal{F}_{\mathsf{all}}$ without any constraint. In this, we will try to derive a general tangent space for the model. Sometimes, this is called unconstrained model analysis–we just work out a form of a general parametric submodel and gain insight on $\mathcal{F}_{\mathsf{all}}$. Note that this is not limited to the nuisance parameter–we will be considering *any* tangent space. Also note that the space $\mathcal{F}_{\mathsf{all}}$ is sometimes referred to as the Hilbert space (of the tangent space).

- **Step 3 (orthogonal complement of nuisance): finding an element in $\mathcal{F}_{\mathsf{nuis}}^{\perp}$.** This is generally done by considering any element $f \in \mathcal{F}_{\mathsf{all}}$ and then attempt to find the projection $\Pi(f|\mathcal{F}_{\mathsf{nuis}}^{\perp})$. A common strategy is to find $f^* \in \mathcal{F}_{\mathsf{all}}$ such that

  $$\mathbb{E}((f(X) - f^*(X))g^T(X)) = 0$$

  for all $g \in \mathcal{F}_{\mathsf{nuis}}$.

  Then we have $\Pi(f|\mathcal{F}_{\mathsf{nuis}}^{\perp}) = f - f^* \in \mathcal{F}_{\mathsf{nuis}}^{\perp}$.

- **Step 4 (RAL estimator and influence function): using $h \in \mathcal{F}_{\mathsf{nuis}}^{\perp}$ to construct an estimating equation.** One can easily verify that the function $h \in \mathcal{F}_{\mathsf{nuis}}^{\perp}$ satisfies the two conditions in Theorem 6 (after multiplying it by a normalizing matrix). The normalized version of $h \in \mathcal{F}_{\mathsf{nuis}}^{\perp}$ will be an influence function. Thus, we can use the construction in Theorem 2 to construct an estimator by solving the estimating equation.

- **Step 5 (efficient estimator): finding the efficient estimator by the tangent space S_β.** If we want to find the efficient estimator, we can work out S_β and then use the projection in equation (5) to construct the efficient score function and the efficient estimator.

We use the term *(full) model parameter* to describe the entire distribution function without any constraint, i.e., they are elements in $\mathcal{F}_{\mathsf{all}}$. And we use the term *nuisance parameter* to describe the distribution in the semi-parametric model, i.e., they are the distribution that satisfies constraints for a particular problem. In general, a model parameter corresponds uniquely to a nuisance parameter and vice versa.

4.1 Conditional factorization

A common method to achieve step 1 and 2 is the idea of conditional factorization. Here we illustrate the idea using an example of three variables and we focus on the step 2–finding the characterization of submodel for the entire tangent space (without any constraint).

Suppose that $X = (X_1, X_2, X_3)$. The joint PDF of X can be factorized into

$$p(x) = p(x_3|x_2, x_1)p(x_2|x_1)p(x_1).$$

We can then separately place a parametric submodel for each of the three (conditional) densities:

$$p(x;\theta) = p(x_3|x_2, x_1;\theta_3)p(x_2|x_1;\theta_2)p(x_1;\theta_1),$$

which implies

$$\log p(x;\theta) = \log p(x_3|x_2,x_1;\theta_3) + \log p(x_2|x_1;\theta_2) + \log p(x_1;\theta_1).$$

The score function will be

$$\begin{aligned} S_\theta(x;\theta) &= \frac{\partial}{\partial\theta}\log p(x;\theta) \\ &= \frac{\partial}{\partial\theta_1}\log p(x_1\theta_1) + \frac{\partial}{\partial\theta_2}\log p(x_2|x_1;\theta_2) + \frac{\partial}{\partial\theta_3}\log p(x_3|x_1,x_2;\theta_3) \\ &= S_{\theta_1}(x_1;\theta_1) + S_{\theta_2}(x_1,x_2;\theta_2) + S_{\theta_3}(x_1,x_2,x_3;\theta_3). \end{aligned}$$

By construction,

$$\mathbb{E}(S_{\theta_1}(X_1;\theta_1)) = 0, \quad \mathbb{E}(S_{\theta_2}(X_1,X_2;\theta_2)|X_1) = 0, \quad \mathbb{E}(S_{\theta_3}(X_1,X_2,X_3;\theta_3)|X_1,X_2) = 0.$$

As a result, one can immediately verify that

$$\mathcal{F}_{\theta_1} \perp \mathcal{F}_{\theta_2} \perp \mathcal{F}_{\theta_3}.$$

For instance, consider any two elements in $\mathcal{F}_{\theta_1}$ and $\mathcal{F}_{\theta_3}$, which can be represented as $B_1 S_{\theta_1}(X_1;\theta_1)$ and $B_3(X_1,X_2)S_{\theta_3}(X_1,X_2,X_3;\theta_3)$,

$$\begin{aligned} \mathbb{E}[S^T_{\theta_1}(X_1;\theta_1)B_1^T B_3(X_1,X_2)S_{\theta_3}(X_1,X_2,X_3;\theta_3)] &= \mathbb{E}\left\{\mathbb{E}[S^T_{\theta_1}(X_1;\theta_1)B_1^T B_3(X_1,X_2)S_{\theta_3}(X_1,X_2,X_3;\theta_3)|X_1,X_2]\right\} \\ &= \mathbb{E}\left\{ S^T_{\theta_1}(X_1;\theta_1)B_1^T B_3(X_1,X_2)\underbrace{\mathbb{E}[S_{\theta_3}(X_1,X_2,X_3;\theta_3)|X_1,X_2]}_{=0}\right\} \\ &= 0. \end{aligned}$$

Note that the coefficients of $\mathcal{F}_{\theta_3}$, $B_3 = B_3(X_1,X_2)$, can depend on X_1 and X_2 as well! This is because the constraint we need is $\int p(x_3|x_1,x_3)dx_3 = 0$ or equivalently, $\mathbb{E}(S_{\theta_3}(X_1,X_2,X_3;\theta_3)|X_1,X_2) = 0$.

With the above results, the (nonparametric) submodel tangent space can be decomposed into

$$\mathcal{F}_\theta = \mathcal{F}_{\theta_1} \oplus \mathcal{F}_{\theta_2} \oplus \mathcal{F}_{\theta_3}.$$

So we can work on characterizing each tangent space to obtain the full tangent space. Moreover, let $\mathcal{F}_{\mathsf{all,j}}$ be the mean-square closure of $\mathcal{F}_{\theta_j}$. One can also show that

$$\mathcal{F}_{\mathsf{all}} = \left\{\text{mean-square closure of } \mathcal{F}_{\mathsf{all},1} \oplus \mathcal{F}_{\mathsf{all},2} \oplus \mathcal{F}_{\mathsf{all},3}\right\}. \tag{6}$$

This gives us a way to characterize the nonparametric tangent space $\mathcal{F}_{\mathsf{all}}$.

Moreover, each of these spaces have a concrete characterization as the following theorem (assuming the dimension of β is $q = 1$; it can be easily generalized to any integer).

Theorem 8 *For each j, $\mathcal{F}_{\mathsf{all,j}}$ is the collection of all conditional mean-zero functions given $X_1,\cdots,X_{j-1}$ with finite variance, i.e,*

$$\mathcal{F}_{\mathsf{all,j}} = \{f(x) : \mathbb{E}(f(X)|X_1,\cdots,X_{j-1}) = 0, \quad \mathsf{Var}(f(X)) < \infty\}.$$

Note that the function f in the above expression is of any finite dimensions.

Proof. For simplicity, we will prove the case of $j = 2$, The same procedure applies for other cases. Let

$$\mathcal{F}^{\dagger}_{\mathsf{all},2} = \{f(x) : \mathbb{E}(f(X)|X_1) = 0, \quad \mathsf{Var}(f(X)) < \infty\}.$$

The goal is to show that $\mathcal{F}^{\dagger}_{\mathsf{all},2} = \mathcal{F}_{\mathsf{all},2}$.

Par I: $\mathcal{F}_{\mathsf{all},2} \subset \mathcal{F}^{\dagger}_{\mathsf{all},2}$. Let $g \in \mathcal{F}_{\mathsf{all},2}$, then g can be expressed as either a score vector of a paranetric submodel or the limit of a sequence of score vectors of submodels. Because every score vector $S_{\theta_2}(X_1, X_2; \theta_2)$ satisfies $\mathbb{E}(S_{\theta_2}(X_1, X_2; \theta_2)|X_1) = 0$, it is easy to see that g must also satisfy this condition, i.e.,

$$\mathbb{E}(g(X_1, X_2)|X_1) = 0$$

and the variance is finite (otherwise it cannot be the mean-square closure). Thus, we have proved this inclusion.

Part II: $\mathcal{F}^{\dagger}_{\mathsf{all},2} \subset \mathcal{F}_{\mathsf{all},2}$. Our strategy is simple. We will show that for any element $f \in \mathcal{F}^{\dagger}_{\mathsf{all},2}$, we can construct a parametric submodel with score vector being f.

Without lost of generality, suppose that $f \in \mathbb{R}^r$ for some integer r. Consider the following parametric submodel with $\theta_2 \in \mathbb{R}^r$:

$$p(x_2|x_1; \theta_2) = p(x_2|x_1; 0)(1 + \theta_2^T f(x_1, x_2)),$$

where $p(x_2|x_1; 0) = p(x_2|x_1; \eta_0)$ is the true PDF. Clearly, when $\theta_2 = 0$, this reduces to the true PDF.

So all we need is to show that under a proper choice of θ_2, $p(x_2|x_1; \theta_2)$ is a density function, which implies that it is an element of a particular parametric submodel. Now we check the if the conditional $f \in \mathcal{F}^{\dagger}_{\mathsf{all},2}$ implies this result. We first examine the score of this 'density':

$$\begin{aligned}
\widetilde{S}(x_1, x_2; \theta_2) &= \frac{\partial}{\partial \theta_2} \log p(x_2|x_1; \theta_2) \\
&= \frac{\partial}{\partial \theta_2} \left(\log p(x_2|x_1; 0) + \log(1 + \theta_2^T f(x_1, x_2)) \right) \\
&= \frac{f(x_1, x_2)}{1 + \theta_2^T f(x_1, x_2)}.
\end{aligned}$$

By taking $\theta_2 = 0$, this becomes

$$\widetilde{S}(x_1, x_2; \theta_2 = 0) = f(x_1, x_2).$$

The property of $\mathcal{F}^{\dagger}_{\mathsf{all},2}$ implies that

$$\mathbb{E}(f(X_1, X_2)|X_1) = 0 = \mathbb{E}(\widetilde{S}(X_1, X_2; \theta_2 = 0)|X_1)$$

so $f(x_1, x_2)$ is indeed a score vector (of the parametric submodel $p(x_2|x_1; \theta_2)$ when $\theta = 0$). This works for every element of $\mathcal{F}^{\dagger}_{\mathsf{all},2}$ so we conclude that $\mathcal{F}^{\dagger}_{\mathsf{all},2} \subset \mathcal{F}_{\mathsf{all},2}$, which completes the proof.

□

Theorem 8 is a very powerful result. It shows that without any additional constraints, the tangent space of a particular conditional density is the same as the entire conditional mean-zero space. In other words, if we

were given a function that has conditional mean-zero (and finite variance), we can always use it to construct a parametric submodel with that function being a tangent vector.

Although here we are using three variables as an example, this derivation works for any number of variables. Thus, Theorem 8 and equation (6) provide useful tools for characterizing a general semi-parametric tangent space. Note that this idea can also be applied to nuisance tangent space; all we need is to include the additional constraint from the specific model.

4.2 Example: restricted moment model

To give a concrete example, consider the restricted moment model such that we observe IID

$$(X_1, Y_1), \cdots, (X_n, Y_n)$$

from an unknown distribution. We assume that

$$Y_i = \mu(X_i; \beta) + \varepsilon_i, \quad \mathbb{E}(\varepsilon_i | X_i) = 0,$$

where μ is a known function. The parameter of interest is $\beta \in \mathbb{R}^q$.

Step 1 and 2. Both the nonparametric and semi-parametric model are characterized by the two model parameters: the PDF of $\varepsilon|X = x$ and the PDF of X, i.e.,

$$p(x) = \eta_1(x), \quad p(\varepsilon|x) = \eta_2(x, \varepsilon).$$

Also, we know that each of them forms a nonparametric tangent space $\mathcal{F}_{\mathsf{all},1}$ and $\mathcal{F}_{\mathsf{all},2}$ such that

$$\mathcal{F}_{\mathsf{all},1} = \{f(x) : \mathbb{E}(f(X)) = 0\}, \quad \mathcal{F}_{\mathsf{all},2} = \{g(x,\varepsilon) : \mathbb{E}(g(X,\varepsilon)|X = x) = 0\}.$$

In the semi-parametric model, we have the moment constraint

$$\mathbb{E}(\varepsilon|X) = 0,$$

so $\mathcal{F}_{\mathsf{all},1} = \mathcal{F}_{\mathsf{nuis},1}$ but $\mathcal{F}_{\mathsf{all},2} \neq \mathcal{F}_{\mathsf{nuis},2}$.

To characterize $\mathcal{F}_{\mathsf{nuis},2}$, consider any parametric submodel $p(\varepsilon|x;\gamma_2)$. This constraint implies

$$\begin{aligned}
0 &= \frac{\partial}{\partial \gamma_2} \mathbb{E}(\varepsilon_i | X_i) \\
&= \frac{\partial}{\partial \gamma_2} \int \varepsilon p(\varepsilon|x;\beta_0,\gamma_2) d\varepsilon \\
&= \int \varepsilon p(\varepsilon|x;\beta_0,\gamma_2) \frac{\partial}{\partial \gamma_2} \log p(\varepsilon|x;\beta_0,\gamma_2) d\varepsilon \\
&= \mathbb{E}\left(\varepsilon S_{\gamma_2}(\varepsilon, X; \beta_0, \gamma_2) \middle| X = x\right).
\end{aligned}$$

Thus, the nuisance tangent space being used is

$$\mathcal{F}_{\mathsf{nuis},2} = \{g(x,\varepsilon) : \mathbb{E}(g(X,\varepsilon)|X = x) = 0, \mathbb{E}\left(\varepsilon g(X,\varepsilon)\middle|X = x\right)\}.$$

Remark. Note that here the nuisance parameter is independent of β since the parametric submodel we use is for $\varepsilon|x$. If we consider a parametric submodel of $y|x$, then β will also be perturbed in this case. Then the resulting tangent space is not the nuisance tangent space but the entire semi-parametric tangent space.

It is more instructive to write it as

$$\begin{aligned}\mathcal{F}_{\mathsf{nuis},2} &= \mathcal{F}_{\mathsf{all},2} \cap \mathcal{F}_{\mathsf{nuis},2,1} \\ \mathcal{F}_{\mathsf{all},2} &= \{g(x,\varepsilon) : \mathbb{E}(g(X,\varepsilon)|X=x) = 0\} \\ \mathcal{F}_{\mathsf{nuis},2,1} &= \{g(x,\varepsilon) : \mathbb{E}\left(\varepsilon g(X,\varepsilon)\middle|X=x\right)\}.\end{aligned}$$

Finally, the semi-parametric nuisance tangent space will be

$$\mathcal{F}_{\mathsf{nuis}} = \mathcal{F}_{\mathsf{nuis},1} \oplus \mathcal{F}_{\mathsf{nuis},2} = \mathcal{F}_{\mathsf{nuis},1} \oplus (\mathcal{F}_{\mathsf{all},2} \cap \mathcal{F}_{\mathsf{nuis},2,1}).$$

By some algebra, we have the following lemma about the relationship of these spaces.

Lemma 9 (Lemma 4.3-4.5 of Tsiatis (2007)) *The nuisance tangent spaces have the following relationship:*

1. $\mathcal{F}_{\mathsf{all},2} = \mathcal{F}_{\mathsf{nuis},1}^{\perp}$.
2. $\mathcal{F}_{\mathsf{nuis},1} \subset \mathcal{F}_{\mathsf{nuis},2,1}$.
3. $\mathcal{F}_{\mathsf{nuis}} = \mathcal{F}_{\mathsf{nuis},2,1}$.

In particular, the third property of Lemma 9 is a powerful result. It directly describes the semi-parametric nuisance tangent space:

$$\mathcal{F}_{\mathsf{nuis}} = \mathcal{F}_{\mathsf{nuis},2,1} = \{g(x,\varepsilon) : \mathbb{E}\left(\varepsilon g(X,\varepsilon)\middle|X=x\right)\}.$$

Step 3: orthogonal complement of the nuisance. Moreover, the orthogonal complement of $\mathcal{F}_{\mathsf{nuis}}$ has the following form:

Theorem 10 (Theorem 4.8 of Tsiatis (2007)) *The orthogonal complement of $\mathcal{F}_{\mathsf{nuis}}$ is*

$$\mathcal{F}_{\mathsf{nuis}}^{\perp} = \{A(X)\varepsilon : \textit{for all } A(x) \in \mathbb{R}^q\}$$

One can easily verify that for any $A(x)\varepsilon$ and $g(x,\varepsilon) \in \mathcal{F}_{\mathsf{nuis}}$,

$$\mathbb{E}(A(X)\varepsilon g(X,\varepsilon)|X=x) = A(X)\mathbb{E}(\varepsilon g(X,\varepsilon)|X=x) = 0$$

by the definition of $\mathcal{F}_{\mathsf{nuis}}$. So the above form makes sense.

Step 4: RAL estimator and influence function. There are two powerful implications from Theorem 10. First, note that $A(X)$ in the space $\mathcal{F}_{\mathsf{nuis}}^{\perp}$ behaves like coefficients. So for any function $\omega(x,\varepsilon)$, the projection

$$\Pi(\omega(x,\varepsilon)|\mathcal{F}_{\mathsf{nuis}}^{\perp}) = \mathbb{E}(\varepsilon\omega(x,\varepsilon))\mathbb{E}(\varepsilon^2|x)^{-1}\varepsilon = \mathbb{E}(\varepsilon\omega(x,\varepsilon))V^{-1}(x)\varepsilon,$$

where $V(x) = \mathbb{E}(\varepsilon^2|x)$.

The second powerful implication is based on the fact that $\varepsilon = (Y - \mu(X;\beta))$. So any elements in $\mathcal{F}_{\mathsf{nuis}}^{\perp}$ can be written as

$$A(X)(Y - \mu(X;\beta)).$$

By Theorem 6, any influence function is an element in $\mathcal{F}_{\mathsf{nuis}}^{\perp}$, which implies that any β that solves the following equation

$$0 = \sum_{i=1}^{n} C_0 A(X_i)(Y_i - \mu(X_i;\widehat{\beta})) \tag{7}$$

leads to an RAL estimator! This is essentially a general form of a Z-estimator. Note that $C_0 = [\mathbb{E}(\varepsilon A(X) S_{\beta}^T(\varepsilon, X; \beta_0, \eta_0))]^{-1}$ is a normalizing matrix to ensure the first condition (versus S_β) in Theorem 6. When C_0 is invertible, the we can multiply both sides by C_0^{-1} and this will not change the estimator so it can be replaced by the identity matrix.

Step 5: efficient esitmator. With the knowledge of the nuisance tangent space, we are able to find the efficient influence function. To start with, we note that the efficient score vector is

$$S_{\mathsf{eff}}(\varepsilon, X) = \Pi(S_\beta(\varepsilon, X; \beta_0, \eta_0)|\mathcal{F}_{\mathsf{nuis}}^{\perp}) = \mathbb{E}(S_\beta(\varepsilon, X; \beta_0, \eta_0)\varepsilon)V^{-1}(X)\varepsilon.$$

Thus, we need to compute $\mathbb{E}(S_\beta(\varepsilon, X; \beta_0, \eta_0)\varepsilon)$.

It turns out that there is a simple closed-form of $\mathbb{E}(S_\beta(\varepsilon, X; \beta_0, \eta_0)\varepsilon)$. Recall the condition that $\mathbb{E}(\varepsilon|X) = 0$. Taking derivative with respect to β leads to

$$\begin{aligned}
0 &= \frac{\partial}{\partial\beta}\mathbb{E}(\varepsilon|X) \\
&= \frac{\partial}{\partial\beta}\int \varepsilon p(\varepsilon|X;\beta,\gamma)d\varepsilon \\
&= \frac{\partial}{\partial\beta}\int (y-\mu(X;\beta))p(y-\mu(X;\beta)|X;\beta,\gamma)dy \\
&= -\nabla_\beta\mu(X;\beta)\int p(y-\mu(X;\beta)|X;\beta,\gamma)dy + \int (y-\mu(X;\beta))\frac{\partial}{\partial\beta}p(y-\mu(X;\beta)|X;\beta,\gamma)dy \\
&= -\nabla_\beta\mu(X;\beta) + \int \varepsilon\frac{1}{p(\varepsilon|X;\beta,\gamma)}\frac{\partial}{\partial\beta}\log p(\varepsilon|X;\beta,\gamma)dy \\
&= -\nabla_\beta\mu(X;\beta) + \mathbb{E}(S_\beta(\varepsilon,X;\beta,\eta)\varepsilon).
\end{aligned}$$

Therefore,

$$\mathbb{E}(S_\beta(\varepsilon, X; \beta_0, \eta_0)\varepsilon) = \nabla_\beta\mu(X;\beta_0).$$

So the efficient score vector is

$$S_{\mathsf{eff}}(\varepsilon, X) = \nabla_\beta\mu(X;\beta_0)V^{-1}(X)\varepsilon$$

and the corresponding estimator is from solving the estimating equation

$$0 = \sum_{i=1}^{n} \nabla_\beta \mu(X_i; \widehat{\beta}) V^{-1}(X_i)(Y_i - \mu(X_i; \widehat{\beta})). \tag{8}$$

The above estimating equation is known to be the optimal generalized estimating equation (GEE). So the efficiency theory explains why the optimal GEE will take this form.

4.3 Example: Cox Model

Consider a simple survival problem where for each individual, we observe a time-to-event variable T and a covariate $X \in \mathbb{R}^q$. Our data will be IID

$$(X_1, T_1), \cdots, (X_n, T_n) \sim p(t,x).$$

In survival analysis, instead of using $p(t,x)$, we often express it as a conditional hazard function $\lambda(t|x)$ and a marginal density $p(x)$ as

$$p(t,x) = \lambda(t|x)\exp(-\Lambda(t|x))p(x),$$

where $\Lambda(t|x) = \int_0^t \lambda(u|x)du$ is the cumulative hazard. The Cox (proporitional hazard) model assume that

$$\lambda(t|x) = \lambda_0(t)\exp(\beta^T X)$$

and the goal is to estimate β.

In Cox's seminal work, he proposed the famous profile likelihood method (also called conditional likelihood or partial likelihood) such that we can estimate β by solving the following estimating equation[2]:

$$0 = \sum_{i=1}^{n} \left(X_i - \frac{\sum_{j=1}^n X_i \exp(X_i^T \widehat{\beta}) I(T_i \le T_j)}{\sum_{j=1}^n \exp(X_i^T \widehat{\beta}) I(T_i \le T_j)} \right).$$

This estimator $\widehat{\beta}$ is asymptotically normal and converges at rate $\sqrt{n}$. An interesting fact is that we do not even need to estimate λ_0 in this case! We will show that the profile likelihood method (also called partial likelihood) will be an efficient estimator for estimating β.

Step 1: nuisance tangent space. We start with writing the PDF as usual the conditional survival:

$$p(t,x) = \lambda(t|x)\exp(-\Lambda(t|x))p(x), \quad \lambda(t|x) = \lambda_0(t)\exp(\beta^T x).$$

Thus, the nuisance/model parameters are $\lambda_0(t)$ and $p(x)$. We can ignore $p(x)$ since it does not have any interaction with β so we will focus on $p(t|x)$.

We construct a parametric submodel of $\lambda(t)$ via

$$\lambda(t;\gamma) = \lambda_0(t)\exp(\gamma^T b(t)), \gamma \in \mathbb{R}^r,$$

[2] see *Cox, D. R. (1972). Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2), 187-202.*

where r is any integer and $b(t)$ behaves like a 'direction'. This is a submodel for the nuisance parameter only since it does not affect the parameter of interest β.

Although this is a particular form of a parametric submodel, its mean square closure (across different dimensions and different score vectors) covers all possible submodels. This leads to $p(t|x;\gamma) = \lambda(t|x;\gamma)\exp(-\int_0^t \lambda(s|x;\gamma))$ and it is easy to see that when $\gamma = 0$, we obtain the true model. Let $S_\gamma(T,X;0) = \frac{\partial}{\partial\gamma}\log p(t|x;\gamma)\big|_{\gamma=0}$ be the score vector. The nuisance tangent space of this submodel is

$$\{BS_\gamma(T,X;0) : B \in \mathbb{R}^{q\times r}\}.$$

To understand the nuisance tangent space, we now analyze the score vector. A direct computation shows that

$$\begin{aligned}
S_\gamma(t,x;\gamma) &= \frac{\partial}{\partial\gamma}\log p(t|x;\gamma) \\
&= \frac{\partial}{\partial\gamma}\log\left(\lambda_0(t)\exp\left(\gamma^T b(t) + \beta^T x - \int_0^t \lambda_0(s)\exp\left(\gamma^T b(s) + \beta^T x\right) ds\right)\right) \\
&= b(t) - \frac{\partial}{\partial\gamma}\int_0^t \lambda_0(s;\gamma) b(s)\exp(\beta^T x + \gamma^T b(s)) ds
\end{aligned}$$

At $\gamma = 0$, this will become

$$\begin{aligned}
S_\gamma(t,x;0) &= b(t) - \int_0^t \lambda_0(s)\exp(\beta^T x) b(s) ds \\
&= \int b(s) dN_t(s) - \int \lambda_0(s)\exp(\beta^T x) b(s) Y_t(s) ds \\
&= \int b(s) dM_{t,x}(s), \qquad (9) \\
N_t(s) &= I(s \geq t) \\
Y_t(s) &= I(s \leq t) \\
dM_{t,x}(s) &= dN_t(s) - \lambda_0(s)\exp(\beta^T x) Y_t(s) ds.
\end{aligned}$$

Thus, the score vector of a parametric submodel of dimension r with direction $b(s)$ can be written as

$$S_\gamma(T,X;0) = \int b(s) dM_{T,X}(s).$$

The stochastic process $M_{T,X}(s)$ is also known as a *counting process*. Using the fact that

$$BS_\gamma(T,X;0) = \int Bb(s) dM_{T,X}(s) = \int \widetilde{b}(s) dM_{T,X}(s)$$

for some function $\widetilde{b}(s) \in \mathbb{R}^q$, the nuisance tangent space under the Cox model is

$$\mathcal{F}_{\mathsf{nuis}} = \left\{\int \widetilde{b}(s) dM_{X,T}(s) : \widetilde{b} \in \mathbb{R}^q \text{ is any function}\right\}.$$

Step 2: nonparametric tangent space. Now we study the nonparametric model. Without any particular restriction, a parametric submodel can be generically written as

$$\lambda(t|x;\gamma) = \lambda_0(t|x)\exp(\gamma^T a(t,x)), \gamma \in \mathbb{R}^r,$$

and $a(t,x)$ is any function of both augments. Thus, the difference to the Cox model (specific model) is that we allow the submodel direction $a(t,x)$ to depend on both t and x. Using the same derivation, one can show that the tangent score will be

$$S_\gamma^\dagger(t,x;0) = \int a(s,x) dM_t(s) \tag{10}$$

and the corresponding tangent space is

$$\mathcal{F}_{\mathsf{all}} = \left\{ \int \widetilde{a}(s,X) dM_{X,T}(s) : \widetilde{a} \in \mathbb{R}^q \text{ is any function} \right\}.$$

Step 3: orthogonal complement of the nuisance. With the above analysis, we now are in a good position to understand $\mathcal{F}_{\mathsf{nuis}}^\perp$. One way to derive an element in $\mathcal{F}_{\mathsf{nuis}}^\perp$ is to consider any element in $\mathcal{F}_{\mathsf{all}}$ that is perpendicular to any function $h \in \mathcal{F}_{\mathsf{nuis}}$. It is hard to directly construct such an element so we try to use the idea of 'projection'.

For any function $a(t,x) \in \mathbb{R}^q$, consider

$$g_a(X,T) = \int a(s,X) dM_{X,T}(s) - \int a^*(s) dM_{X,T}(s) = \int [a(s,X) - a^*(s)] dM_{X,T}(s) \in \mathcal{F}_{\mathsf{all}}$$

where $a^*(s)$ is some function that we want to find. The first part is the original element in $\mathcal{F}_{\mathsf{all}}$ computed from $a(t,x)$ and the second part is an element in $\mathcal{F}_{\mathsf{nuis}}$. If we choose $a^*(s)$ nicely, we can make $g_a(X,T)$ to be perpendicular to every element in $\mathcal{F}_{\mathsf{nuis}}$. Namely, we choose $a^*(s)$ such that

$$\begin{aligned} 0 &= \mathbb{E}(g_a(X,T) \int b(s) dM_{X,T} ds) \\ &= \mathbb{E}\left(\int [a(u,X) - a^*(u)] dM_{X,T}(u) \int b(s) dM_{X,T}(s) \right) \end{aligned}$$

for any function $b(s)$. This is the covariance of the integral of two counting processes (martingales in this case) so it turns out that[3] it is equivalent to the requirement that

$$\begin{aligned} 0 &= \mathbb{E}\left(\int [a(s,X) - a^*(s)] b(s) \lambda_0(s) \exp(\beta^T X) Y_T(s) ds \right) \\ &= \int \mathbb{E}\left[[a(s,X) - a^*(s)] \exp(\beta^T X) Y_T(s) \right] b(s) \lambda_0(s) ds \end{aligned}$$

for all $b(s)$. Thus, we need to choose $a^*(s)$ such that

$$0 = \mathbb{E}\left[[a(s,X) - a^*(s)] \exp(\beta^T X) Y_T(s) \right],$$

which leads to

$$a^*(s) = \frac{\mathbb{E}(a(s,X) \exp(\beta^T X) Y_T(s))}{\mathbb{E}(\exp(\beta^T X) Y_T(s))}. \tag{11}$$

Thus, the space $\mathcal{F}_{\mathsf{nuis}}^\perp$ can be characterized by

$$\mathcal{F}_{\mathsf{nuis}}^\perp = \left\{ \int (a(s,X) - a^*(s)) dM_{X,T}(s) : a(s,x) \in \mathbb{R}^q \text{ is any function} \right\}.$$

[3] see this book for more details: *Fleming, T. R., & Harrington, D. P. (2011). Counting processes and survival analysis (Vol. 169). John Wiley & Sons.*

Step 4: RAL estimator and influence function. Using the fact that $Y_T(s) = I(s \le T)$, we can rewrite equation (11) as

$$a^*(s) = \frac{\mathbb{E}(a(s,X)\exp(\beta^T X)I(s \le T))}{\mathbb{E}(\exp(\beta^T X)I(s \le T))}.$$

With the observed data, we can estimate it using

$$\widehat{a}^*(s) = \frac{\sum_{j=1}^n a(s,X_j)\exp(\beta^T X_j)I(s \le T_j)}{\sum_{j=1}^n \exp(\beta^T X_j)I(s \le T_j)}.$$

Thus, given any function $a(t,x)$, we can construct an RAL estimator $\widehat{\beta}_a$ by solving the estimating equation:

$$\begin{aligned}
0 &= \sum_{i=1}^n \int (a(s,X_i) - \widehat{a}^*(s))dM_{X_i,T_i}(s) \\
&= \sum_{i=1}^n \int \left(a(s,X_i) - \frac{\sum_{j=1}^n a(s,X_j)\exp(\beta^T X_j)I(s \le T_j)}{\sum_{j=1}^n \exp(\beta^T X_j)I(s \le T_j)} \right) (dN_{T_i}(s) - \lambda_0(s)\exp(\beta^T X_i)I(s \le T_i)(s)ds) \\
&= \sum_{i=1}^n \int \left(a(s,X_i) - \frac{\sum_{j=1}^n a(s,X_j)\exp(\beta^T X_j)I(s \le T_j)}{\sum_{j=1}^n \exp(\beta^T X_j)I(s \le T_j)} \right) dN_{T_i}(s) \\
&= \sum_{i=1}^n \left(a(T_i,X_i) - \frac{\sum_{j=1}^n a(T_i,X_j)\exp(\beta^T X_j)I(T_i \le T_j)}{\sum_{j=1}^n \exp(\beta^T X_j)I(T_i \le T_j)} \right).
\end{aligned}$$

Note that in the last equality, we use the fact that $\frac{1}{n}\sum_{i=1}^n N_{T_i}(s) = \widehat{F}(t)$ is the empirical distribution of T. This construction is based on Theorem 2 and 6. Note that similar to equation (7), there will be a normalizing matrix

$$C_0 = \left[\mathbb{E}(\int (a(s,X) - \widehat{a}^*(s))dM_{X,T}(s))S_\beta(X,T;\beta_0,\eta_0) \right]^{-1}$$

in front of the estimating equation to ensure the first condition in Theorem 6 holds. Luckily, this matrix is often invertible so in practice we can replace it by the identity matrix.

Step 5: efficient estimator. The above procedure shows a generic way of constructing an RAL estimator. Among all RAL estimators, we are particularly interested in the efficient estimator. To construct the efficient estimator, we need to derive the tangent vector $S_\beta(X,T;\beta_0)$. By a direct differentiation of $\log p(t|x;\beta)$ with respect to β, it is easy to see that

$$S_\beta(X,T;\beta_0) = \int X dM_{X,T}(s)$$

so it turns out that the choice $a(s,X) = X$ will be the estimating equation leading to the most efficient estimator, i.e, the efficient estimator of β is obtained by solving

$$0 = \sum_{i=1}^n \left(X_i - \frac{\sum_{j=1}^n X_i \exp(\beta^T X_j)I(T_i \le T_j)}{\sum_{j=1}^n \exp(\beta^T X_j)I(T_i \le T_j)} \right),$$

which is the Cox's profile likelihood method!

Remark on censoring problem. Note that the above derivation can be easily generalized to the case of censoring. The main derivation will be very similar but there will be an additional nuisance parameter due to the censoring variable. The quantity $N_T(s) = I(s \ge T)$ in counting process $dM_{X,T}(s)$ will be replaced by $N_{T,\Delta}(s) = I(s \ge T, \Delta = 1)$, where Δ is the binary indicator such that $\Delta = 1$ represents observing the event (not censored). See Section 5.2 of Tsiatis 2007 for more details.

5 Finding efficient estimators: geometric approach

Here we describe a geometric perspective of the influence function and this leads to an alternative way of finding an efficient estimator. This part is based on Chapter 25.3 of the following book

> Van der Vaart, A. W. (2000). Asymptotic statistics (Vol. 3). Cambridge university press.

5.1 Differentiation in quadratic mean

A more general way of writing a parametric submodel is based on the idea of differentiation in quadratic mean (DQM). We characterize a parametric submodel using the DQM at the true model with a differentiation (direction) $g(x)$. Let $p(x;t)$ be a parametric model with a parameter $t \in \mathbb{R}$ and $P(x;t)$ be its CDF with $t \in \mathbb{R}$. This parametric model is DQM at $t = 0$ with a differentiation $g(x)$ if

$$\int \left(\frac{\sqrt{dP(x;t)} - \sqrt{dP(x;0)}}{t} - \frac{1}{2} g(x) \sqrt{dP(x;0)} \right)^2 \to 0. \tag{12}$$

This formulation is rather abstract so here we give some expressions that show more connections to the previous problem. Equation (12) implies

$$\int \left(\frac{\sqrt{\frac{dP(x;t)}{dP(x;0)}} - 1}{t} - \frac{1}{2} g(x) \right)^2 dP(x;0) \to 0 \tag{13}$$

and using the fact that $p(x;t) = \frac{dP(x;t)}{dx}$,

$$\mathbb{E} \left(\frac{\sqrt{\frac{p(X;t)}{p(X;0)}} - 1}{t} - \frac{1}{2} g(X) \right)^2 \to 0 \tag{14}$$

To see how equation (14) links to our problem, let $p(x;\beta_0,\gamma)$ is a parametric submodel and $\gamma = 0$ is the true model. Consider the specific submodel

$$p(x;\beta_0,\gamma) = p(x;\beta_0,\eta_0)(1+\gamma^T S_\gamma(x;\beta_0,0)).$$

One can easily verify that the above model has a score vector $S_\gamma(x;\beta_0,0)$ at $\gamma = 0$. For any choice of $\gamma \in \mathbb{R}^r$, we can always reparametrize it as $\gamma = t\rho_0$ such that $\rho_0 \in \mathbb{R}^r$ and $t = \|\gamma\| \in \mathbb{R}$. The entire r-parameter submodel can be expressed in terms of two parameters (t,ρ_0). For each fixed ρ_0, the submodel has only one parameter t so we can associate it with equation (14) by setting

$$p(x;t) = p(x;\beta_0,t\rho_0).$$

Clearly, the density ratio becomes

$$\frac{p(X;t)}{p(X;0)} = (1+t\rho_0^T S_\gamma(x;\beta_0,0))$$

So the RHS of equation (14) is

$$\begin{aligned}\mathbb{E}\left(\frac{\sqrt{\frac{p(X;t)}{p(X;0)}}-1}{t}-\frac{1}{2}g(X)\right)^2 &= \mathbb{E}\left(\frac{\sqrt{(1+t\rho_0^T S_\gamma(x;\beta_0,0))}-1}{t}-\frac{1}{2}g(X)\right)^2\\ &= \mathbb{E}\left(\frac{\frac{1}{2}t\rho_0^T S_\gamma(x;\beta_0,0))+O(t^2)}{t}-\frac{1}{2}g(X)\right)^2\\ &\to 0,\end{aligned}$$

which implies

$$g(x)=\rho_0^T S_\gamma(x;\beta_0,0)).$$

Thus, any parametric submodel can be expressed in terms of DQM with differentiation $g(x)=\rho_0^T S_\gamma(x;\beta_0,0))$. Therefore, some literature will introduce the concept of parametric submodels using DQM. This concepts is a general way of constructing a parametric submodel.

5.2 Geometry of influence functions

From equation (12), the DQM

$$\int\left(\frac{\sqrt{dP(x;t)}-\sqrt{dP(x;0)}}{t}-\frac{1}{2}g(x)\sqrt{dP(x;0)}\right)^2\to 0$$

defines a general description of a parametric submodel. In Section 3, the parametric submodel is defined via perturbing the nuisance parameter. Using the DQM, we can consider submodels that also change the parameter of interest β. In many places, people often write

$$P_t(x)=P(x;t),\quad P_0(x)=P(x;0)$$

and use P_0 to denote the true distribution (i.e., the CDF formed by $p(x;\beta_0,\eta_0)$). With this notation, the parameter of interest β can be written as

$$\beta_t=\Omega(P_t),\quad \beta_0=\Omega(P_0),$$

where $\Omega:\mathcal{P}\mapsto\mathbb{R}^q$ is a statistical functional.

Moreover, the DQM provides a general way of defining a semi-parametric tangent space. Let

$$\mathcal{F}_{0,\text{semi}}=\{g\in L_2(P_0): g \text{ can be used in equation (12)}\}$$

and then we have

$$\mathcal{F}_{\text{semi}}=\overline{\mathcal{F}_{0,\text{semi}}},$$

where $\overline{\mathcal{F}_{0,\text{semi}}}$ is the mean squared closure of $\mathcal{F}_{0,\text{semi}}$. Note that there are two common forms of g in equation (12). If we write the submodel (of PDF) as $p_t(x)=p_0(x)+tg(x)$, then equation (12) requires $\int g(x)dx=0$. If we write the submodel as $p_t(x)=p_0(x)(1+tg(x))$ or $p_t(x)=p_0(x)\exp(1+tg(x))$, then equation (12) requires $\int g(x)p_0(x)dx=0$.

The DQM informs us a derivative of P_t along the direction g. How does this affects the changes in β_t? Since $\beta_t = \Omega(P_t)$ is defined via a statistical functional, informally we can write

$$\lim_{t\to 0} \frac{\beta_t - \beta_0}{t} = \lim_{t\to 0} \frac{\Omega(P_t) - \Omega(P_0)}{t} = \mathcal{L}(g),$$

where $\mathcal{L} : L_2(P_0) \to \mathbb{R}^q$ is a functional. Formally speaking, $\mathcal{L}$ is the Hadamard differential of the functional Ω at P_0. By the Riesz representation theorem, there exists a function $L \in L_2(P_0)$ and $L : \mathbb{R}^d \to \mathbb{R}^q$ such that

$$\mathcal{L}(g) = \langle L, g \rangle_{L_2(P_0)} = \int L(x) g(x) dP_0(x).$$

Namely, $L(x)$ is the basis function that measures the influence of moving the model along direction $g(x)$ on the parameter of interest β_t. Note that L depends on P_0 but is independent of g. This function $L(x)$ is the formal definition of *influence function*! Namely, $L(x) = \psi(x)$ using the notation from the previous sections.

To see why $L(x)$ is an influence function, consider $g(x) = g_\eta(x)$ that only changes the nuisance parameter. Using the derivation in Section 5.1, we know that

$$g(x) = \gamma^T S_\gamma(x; \beta_0, \eta_0)$$

when we use a particular submodel $\gamma \in \mathbb{R}^r$. In this case, $\beta_t = \beta_0$ so

$$\begin{aligned}
0 &= \lim_{t\to 0} \frac{\beta_t - \beta_0}{t} \\
&= \int L(x) g(x) dP_0(x) \\
&= \int L(x) \gamma^T S_\gamma(x; \beta_0, \eta_0) dP_0(x) \\
&= \gamma \int L(x) S_\gamma^T(x; \beta_0, \eta_0) dP_0(x) \\
&= \gamma \mathbb{E}(L(X) S_\gamma^T(X; \beta_0, \eta_0))
\end{aligned}$$

for any $\gamma \in \mathbb{R}^r$. Thus, we conclude that $\mathbb{E}(L(X) S_\gamma^T(X; \beta_0, \eta_0)) = 0$, which is a requirement for L being an influence function in Theorem 5.

To verify that $L(x)$ is an influence function, we also need to consider the case when $g(x) = g_\beta(x)$ that changes β_t entirely. For simplicity, consider $g(x) = a_0^T S_\beta(x; \beta_0, \eta_0)$. In this case, $\beta_t = \beta_0 + a_0 t$. A direct computation shows that

$$\begin{aligned}
a_0 &= \lim_{t\to 0} \frac{\beta_t - \beta_0}{t} \\
&= \int L(x) g(x) dP_0(x) \\
&= \int L(x) a_0^T S_\beta(x; \beta_0, \eta_0) dP_0(x) \\
&= a_0 \int L(x) S_\beta^T(x; \beta_0, \eta_0) dP_0(x) \\
&= a_0 \mathbb{E}(L(X) S_\beta^T(X; \beta_0, \eta_0))
\end{aligned}$$

for all a_0. Thus, we conclude that $\mathbb{E}(L(X)S_\beta^T(X;\beta_0,\eta_0)) = \mathbf{I}_{q\times q}$, which is the other requirement for L being an influence function in Theorem 5. Note that although in Theorem 5, the two conditions are stated as necessarily conditions for influence functions, using the same argument as Theorem 2, we can construct an estimator of $\beta_0 = \Omega(P_0)$ with $L(x)$ be the influence function. So verifying the two conditions is enough to argue that $L(x)$ is indeed an influence function.

In general, L is non-unique, just like in the usual case that we have many influence functions for the same parameter of interest. However, if we constraint ourselves to $L \in \mathcal{F}_{\mathsf{semi}}$, then there is only one unique element $L^* \in \mathcal{F}_{\mathsf{semi}}$ such that $\mathcal{L}(g) = \langle L^*, g\rangle_{L_2(P_0)}$. And it can be characterized by the projection

$$L^*(x) = \Pi(L(x)|\mathcal{F}_{\mathsf{semi}}) = \Pi(\psi(x)|\mathcal{F}_{\mathsf{semi}}) = \psi_{\mathsf{eff}}(x),$$

which means that $L^*(x) = \psi_{\mathsf{eff}}(x)$ is the *efficient influence function*! Thus, the efficient influence function can be defined as the unique element in $L^* = \mathcal{F}_{\mathsf{semi}}$ such that g, the differentiation of P_t, is associated with the differentiation of Ω via

$$\lim_{t\to 0}\frac{\Omega(P_t)-\Omega(P_0)}{t} = \langle L^*, g\rangle_{L_2(P_0)}.$$

With the above notation, we have another way to characterize efficiency. To compare different influence functions, we consider any direction $g \in \mathcal{F}_{\mathsf{semi}}$ so the quantity

$$\lim_{t\to 0}\frac{\beta_t-\beta_0}{t} = \langle L, g\rangle_{L_2(P_0)} = \rho_0$$

is fixed. The variance (covariance matrix) of an influence function describes the uncertainty of the corresponding estimator and can be succinctly written as

$$\mathsf{Var}(L(X)) = \mathsf{Cov}(L(X)) = \int L(x)L^T(x)dP_0(x) = \langle L, L^T\rangle_{L_2(P_0)}.$$

Thus, we want to find the influence function $L_\dagger$ such that $\langle L_\dagger, L_\dagger^T\rangle_{L_2(P_0)}$ is minimized subject to the constraint $\langle L_\dagger, g\rangle_{L_2(P_0)} = \rho_0$. Because $g \in \mathcal{F}_{\mathsf{semi}}$, we can decompose $L = L_1 + L_2$ such that $L_1 = \Pi(L_1|\mathcal{F}_{\mathsf{semi}}) \in \mathcal{F}_{\mathsf{semi}}$ and $L_2 \in \mathcal{F}_{\mathsf{semi}}^\perp$. It follows immediately that

$$\begin{aligned}\mathsf{Var}(L(X)) &= \langle L_1, L_1^T\rangle_{L_2(P_0)} + \langle L_2, L_2^T\rangle_{L_2(P_0)}\\ \langle L_1, g\rangle_{L_2(P_0)} &= \rho_0\\ \langle L_2, g\rangle_{L_2(P_0)} &= 0.\end{aligned}$$

As a result, the optimal choice $L_\dagger$ will be

$$L_\dagger(x) = L_1(x) = \Pi(L(x)|\mathcal{F}_{\mathsf{semi}}) = L^*(x),$$

the efficient influence function.

5.3 Example: simple missing at random problem

Now we illustrate how the above procedure can be used to construct an efficient estimator using a simple missing at random problem. Consider two random variables problem: X, Y. X is the covariate that is always

observed. Y is the response of interest but sometimes it may be missing. Let R be a binary response indicator where $R = 1$ indicates that Y is observed and 0 is the case where Y is missing. The parameter of interest is $\beta = \mathbb{E}(Y)$. In this case, the underlying distribution is $P(x,y,r)$ or its PDF form $p(x,y,r)$.

In general, this problem is unidentified since when $R = 0$, we will never observe Y. A popular assumption that identifies $R = 0$ is

$$\text{(MAR)} \quad Y \perp R|X.$$

This is known as missing at random (MAR).

Under (MAR), there are two ways we can rewrite the parameter of interest β. The first one is

$$\beta = \mathbb{E}(Y) = \mathbb{E}\left(\frac{YR}{P(R=1|X)}\right) = \int \frac{yr}{p(R=1|x)} p(x,y,r)dxdydr.$$

This is related to the inverse probability weighting (IPW) estimator since we can estimate β via

$$\widehat{\beta}_{IPW} = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i R_i}{P(R=1|X_i)} = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i R_i}{\pi(X_i)}.$$

The quantity $\pi(x) = P(R=1|X=x)$ is known as the propensity score. If we know $\pi(x)$, we can immediately use the above estimator. If we do not know it, we can estimate by regressing R with X.

The other way is to rewrite β as

$$\beta = \mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(\mathbb{E}(Y|X,R=1)) = \mathbb{E}(m_1(X)) = \int\int yp(y|x,R=1)dyp(x)dx,$$

which leads to the regression adjustment (RA) estimator

$$\widehat{\beta}_{RA} = \frac{1}{n}\sum_{i=1}^{n} m_1(X_i),$$

where $m_1(x) = \mathbb{E}(Y|X=x,R=1)$ is the observed outcome regression. When $m_1(x)$ is unknown, we can estimate it using the observation with the outcome Y.

The above two ways of explicitly express how the parameter of interest β is written in terms of the underlying PDF $p(x,y,r)$. Then we can study the underlying efficient influence function.

The strategy can be summarized as follows.

1. **Step 1: finding a parametric submodel.** We start with a parametric submodel:

   $$p_t(x,y,r) = p_0(x,y,r)(1+tg(x,y,r)),$$

   where $p_0(x,y,r)$ is the true model and $g(x,y,r)$ is any submodel such that $\int p_0(x,y,r)g(x,y,r)dxdydr = 0$. Note that $\int f(r)dr = \sum_r f(r)$.

2. **Step 2: deriving the form of the statistical functional.** We derive a closed-form of the parameter interest and write it as a statistical functional. Namely, we have a closed-form of Ω such that $\beta_t = \Omega(P_t)$ and $\beta_0 = \Omega(P_0)$. Note that P_t is the CDF induced by p_t.

3. **Step 3: finding the EIF using the L_2 projection.** We attempt to find L^* such that

$$\frac{\beta_t - \beta_0}{t} \to \langle L^*, g \rangle_{L_2(P_0)} = \int L^*(x,y,r) g(x,y,r) p_0(x,y,r) dxdydr.$$

The function $L^*(x,y,r)$ will be the efficient influence function.

We illustrate this process using both the IPW and the RA approaches. And we will show that they both lead to the same efficient estimator, which is known as the doubly-robust estimator. Since the first two steps have been derived in the above discussion, in what follows we will focus on Step 3.

5.3.1 Method 1: inverse probability weighting

The IPW estimator uses the expression

$$\beta = \mathbb{E}\left(\frac{YR}{P(R=1|X)}\right) = \int \frac{yr}{p(1|x)} p(x,y,r) dxdydr,$$

where we use $p(1|x) = p(R=1|x)$ as abbreviataion. Thus,

$$\beta_t = \int \frac{yr}{p_t(1|x)} p_t(x,y,r) dxdydr = \int \frac{yI(r=1)}{p_t(1|x)} p_t(x,y,r) dxdydr$$

The difference

$$\begin{aligned}
\beta_t - \beta_0 &= \int \frac{yI(r=1)}{p_t(1|x)} p_t(x,y,r) dxdydr - \int \frac{yI(r=1)}{p_0(1|x)} p_0(x,y,r) dxdydr \\
&= \underbrace{\int \frac{yI(r=1)}{p_t(1|x)} p_t(x,y,r) dxdydr - \int \frac{yI(r=1)}{p_0(1|x)} p_t(x,y,r) dxdydr}_{(I)} \\
&\quad + \underbrace{\int \frac{yI(r=1)}{p_0(1|x)} p_t(x,y,r) dxdydr - \int \frac{yI(r=1)}{p_0(1|x)} p_0(x,y,r) dxdydr}_{(II)}.
\end{aligned}$$

Analysis of (I). Since we will only focus on the leading term, the first quantity can be written as

$$\begin{aligned}
(I) &= \int yI(r=1) \left(\frac{1}{p_t(1|x)} - \frac{1}{p_0(1|x)}\right) p_t(x,y,r) dxdydr \\
&= \int yI(r=1) \left(\frac{1}{p_t(1|x)} - \frac{1}{p_0(1|x)}\right) p_0(x,y,r) dxdydr (1 + O(t))
\end{aligned}$$

and we will ignore the $O(t)$ term since it is a smaller order one.

The quantity $\frac{1}{p_t(1|x)} = \frac{p_t(x)}{p_t(1,x)}$. By the expression $p_t(x,y,r) = p_0(x,y,r)(1+tg(x,y,r))$, it can be written as

$$\begin{aligned}
\frac{1}{p_t(1|x)} &= \frac{p_t(x)}{p_t(1,x)} \\
&= \frac{p_0(x)+t\int p_0(x,y',r')g(x,y',r')dy'dr'}{p_0(1,x)+t\int p_0(x,y',1)g(x,y',1)dy'} \\
&= \frac{p_0(x)+t\int p_0(x,y',r')g(x,y',r')dy'dr'}{p_0(1,x)} + t\frac{p_0(x)}{p_0^2(1,x)}\int p_0(x,y',1)g(x,y',1)dy' \\
&= \frac{1}{p_0(1|x)} + t\frac{1}{p_0(1,x)}\int p_0(x,y',r')g(x,y',r')dy'dr' - t\frac{1}{p_0(1|x)p_0(1,x)}\int p_0(x,y',1)g(x,y',1)dy' + o(t).
\end{aligned}$$

Using this, we can rewrite (I) as

$$\begin{aligned}
(I) &\approx \int yI(r=1)\left(\frac{1}{p_t(1|x)} - \frac{1}{p_0(1|x)}\right)p_0(x,y,r)dxdydr \\
&= t\int \frac{yI(r=1)}{p_0(1,x)}\int p_0(x,y',r')g(x,y',r')dy'dr'p_0(x,y,r)dxdydr \\
&\quad - t\int \frac{yI(r=1)}{p_0(1|x)p_0(1,x)}\int p_0(x,y',1)g(x,y',1)dy'p_0(x,y,r)dxdydr \\
&= t\int y\underbrace{\frac{I(r=1)p_0(x,y,r)}{p_0(1,x)}dydr}_{=p_0(y|x,1)}\int p_0(x,y',r')g(x,y',r')dy'dr'dx \\
&\quad - t\int y\underbrace{\frac{I(r=1)p_0(x,y,r)}{p_0(1|x)p_0(1,x)}dydr}_{=p_0(y|x,1)/p_0(1|x)}\int p_0(x,y',1)g(x,y',1)dy'dx \\
&= t\int m_0(x,1)\int p_0(x,y',r')g(x,y',r')dy'dxdr' \\
&\quad - t\int \frac{m_0(x,1)}{p(1|x)}\int p_0(x,y',1)g(x,y',1)dy'dx \\
&= t\int m_0(x,1)\int p_0(x,y',r')g(x,y',r')dy'dxdr' \\
&\quad - t\int \frac{m_0(x,1)}{p(1|x)}\int I(r'=1)p_0(x,y',r')g(x,y',r')dy'dxdr' \\
&= t\int \left(m_0(x,1) - \frac{m_0(x,1)I(r'=1)}{p_0(1|x)}\right)p_0(x,y',r')g(x,y',r')dy'dxdr' \\
&= t\left\langle m_0(x,1) - \frac{m_0(x,1)I(r=1)}{p_0(1|x)}, g(x,y,r)\right\rangle_{L_2(P_0)}.
\end{aligned}$$

Analysis of (II). This part is very straight forward since

$$p_t(x,y,r) - p_0(x,y,r) = tp_0(x,y,r)g(x,y,r).$$

Thus,

$$\begin{aligned}(II) &= \int \frac{yI(r=1)}{p_0(1|x)} p_t(x,y,r)dxdydr - \int \frac{yI(r=1)}{p_0(1|x)} p_0(x,y,r)dxdydr \\ &= t\int \frac{yI(r=1)}{p_0(1|x)} p_0(x,y,r)g(x,y,r)dxdydr \\ &= t\left\langle \frac{yI(r=1)}{p_0(1|x)}, g(x,y,r)\right\rangle_{L_2(P_0)}.\end{aligned}$$

Putting both terms together, we conclude that

$$\frac{\beta_t - \beta_0}{t} \to \left\langle m_0(x,1) + \frac{y - m_0(x,1)I(r=1)}{p_0(1|x)}, g(x,y,r)\right\rangle_{L_2(P_0)}.$$

Note that for any constant c_0,

$$\langle c_0, g(x,y,r)\rangle_{L_2(P_0)} = 0$$

due to the normalizing constraint $\int p_0(x,y,r)g(x,y,r)dxdydr = 0$. Thus,

$$\frac{\beta_t - \beta_0}{t} \to \left\langle m_0(x,1) + \frac{y - m_0(x,1)I(r=1)}{p_0(1|x)} + c_0, g(x,y,r)\right\rangle_{L_2(P_0)}$$

for any constant c_0. As a result, we know that the efficient influence function $L^*(x,y,r) = m_0(x,1) + \frac{y-m_0(x,1)I(r=1)}{p_0(1|x)} + c_0$ for some constant c_0. To find c_0, we use the constraint from the influence function that

$$\mathbb{E}(L^*(X,Y,R)) = 0 = \mathbb{E}\left(m_0(X,1) + \frac{Y - m_0(X,1)I(R=1)}{p_0(1|X)} + c_0\right)$$

which implies that $c_0 = -\beta_0$. Thus,

$$L^*(x,y,r) = \frac{(y - m_0(x,1))I(r=1)}{p_0(1|x)} + m_0(x,1) - \beta_0$$

and this implies a closed-form of the efficient estimator:

$$\widehat{\beta}_{\text{eff}} = \frac{1}{n}\sum_{i=1}^{n} \frac{(Y_i - m(X_i,1))I(R_i=1)}{P(R=1|X_i)} + m(X_i,1),$$

which is known as the doubly-robust estimator.

5.3.2 Method 2: regression adjustment (g-computation)

In the RA estimator, note that we express β using

$$\beta = \Omega(p) = \int\int yp(y|x,R=1)dyp(x)dx.$$

Thus,

$$\beta_t = \Omega(p_t) = \int\int yp_t(y|x,R=1)dyp_t(x)dx.$$

The difference

$$\begin{aligned}\beta_t - \beta_0 &= \int\int y p_t(y|x,R=1)dy p_t(x)dx. - \int\int y p_0(y|x,R=1)dy p_0(x)dx \\ &= \underbrace{\int\int y p_t(y|x,R=1)dy p_t(x)dx. - \int\int y p_0(y|x,R=1)dy p_t(x)dx}_{(I)} \\ &\quad + \underbrace{\int\int y p_0(y|x,R=1)dy p_t(x)dx - \int\int y p_0(y|x,R=1)dy p_0(x)dx}_{(II)}.\end{aligned}$$

Analysis of (I). We can write it as

$$\begin{aligned}(I) &= \int\int y[p_t(y|x,R=1) - p_0(y|x,R=1)]dy p_t(x)dx \\ &= \int\int y[p_t(y|x,R=1) - p_0(y|x,R=1)]dy p_0(x)dx(1+O(t)).\end{aligned}$$

We will ignore $O(t)$ term since it will be of the next order.

Using the perturbation that $p_t(x,y,r) = p_0(x,y,r)(1+tg(x,y,r))$,

$$\begin{aligned}p_t(y|x,1) &= \frac{p_t(x,y,1)}{p(x,1)} \\ &= \frac{p_0(x,y,1)+tp_0(x,y,1)g(x,y,1)}{p_0(x,1)+t\int p_0(x,y',r)g(x,y',r))dy'} \\ &= \frac{p_0(x,y,1)+tp_0(x,y,1)g(x,y,1)}{p_0(x,1)} - t\frac{\int p_0(x,y',r)g(x,y',r))dy'}{p_0^2(x,1)}p_0(x,y,1) \\ &= p_0(y|x,1)+tp_0(y|x,1)g(x,y,1) - t\frac{\int p_0(x,y',1)g(x,y',1))dy'}{p_0(x,1)}p_0(y|x,1)\end{aligned}$$

Thus,

$$p_t(y|x,1) - p_0(y|x,1) = t\left(p_0(y|x,1)g(x,y,1) - \frac{\int p_0(x,y',1)g(x,y',1)dy'}{p_0(x,1)}p_0(y|x,1)\right) \tag{15}$$

so (I) will contains two terms at the order of $O(t)$. The first term is

$$\begin{aligned}t\int y p_0(y|x,1)g(x,y,1)p_0(x)dxdy &= t\int y\frac{p_0(x,y,r)}{p_0(x,1)}I(r=1)g(x,y,r)p_0(x)dxdydr \\ &= t\int y\frac{yI(r=1)}{p(1|x)}g(x,y,r)p_0(x,y,r)dxdydr \\ &= t\left\langle \frac{yI(r=1)}{p_0(1|x)}, g(x,y,r)\right\rangle_{L_2(P_0)}.\end{aligned}$$

For the second term,

$$
\begin{aligned}
t\int y\frac{\int p_0(x,y',1)g(x,y',1)dy'}{p_0(x,1)}p_0(y|x,1)dyp_0(x)dx &= t\int m_0(x,1)\frac{\int p_0(x,y',1)g(x,y',1)dy'}{p_0(x,1)}p_0(x)dx \\
&= t\int \frac{m_0(x,1)}{p_0(1|x)}p_0(x,y',1)g(x,y',1)dy'dx \\
&= t\int \frac{m_0(x,r)I(r=1)}{p_0(1|x)}g(x,y',r)p_0(x,y',r)dy'dxdr \\
&= t\left\langle \frac{m_0(x,r)I(r=1)}{p_0(1|x)}, g(x,y,r)\right\rangle_{L_2(P_0)}.
\end{aligned}
$$

As a result, the leading term of (I) is

$$
(I) = t\left\langle \frac{(y-m_0(x,r))I(r=1)}{p_0(1|x)}, g(x,y,r)\right\rangle_{L_2(P_0)} + o(t).
$$

Analysis of (II). Again, using the perturbation that $p_t(x,y,r) = p_0(x,y,r)(1+tg(x,y,r))$,

$$
\begin{aligned}
p_t(x) &= \int p_t(x,y,r)dydr \\
&= \int p_0(x,y,r)dydr + t\int p_0(x,y,r)g(x,y,r)dydr \\
&= p_0(x) + t\int p_0(x,y,r)g(x,y,r)dydr.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
(II) &= t\int yp_0(y|x,R=1)\int p_0(x,y',r)g(x,y',r)dy'drdydx \\
&= t\int m_0(x,1)\int p_0(x,y',r)g(x,y',r)dy'drdx \\
&= t\int m_0(x,1)p_0(x,y',r)g(x,y',r)dy'drdx \\
&= t\langle m_0(x,1), g(x,y,r)\rangle_{L_2(P)}.
\end{aligned}
$$

As a result, we conclude that

$$
\beta_t - \beta_0 = t\left\langle \frac{(y-m_0(x,r))I(r=1)}{p_0(1|x)} + m_0(x,1), g(x,y,r)\right\rangle_{L_2(P_0)} + o(t)
$$

so

$$
\frac{\beta_t-\beta_0}{t} \to \left\langle \frac{(y-m_0(x,r))I(r=1)}{p_0(1|x)} + m_0(x,1), g(x,y,r)\right\rangle_{L_2(P_0)},
$$

which leads to the same efficient influence function as the IPW case.

Remark. Although this procedure gives the efficient influence function, it requires the knowledge of having a simple form of expressing the parameter of interest in terms of the underlying distribution, i.e., we need to know how the statistical functional Ω looks like. In some problems (for instance, the Cox model), this may not be an easy task so sometimes we will still need to start with characterizing the nuisance tangent space.

5.4 More about DQM

The concept of DQM appears in many analysis of a parametric model. In particular, when we are dealing with the problem of *local alternatives*–a sequence of model that is very close to the true model but we can still derive meaningful results. This gives us some hints on why we may use it as a way to construct parametric submodels. The key part of a parametric submodel is its behavior close to the true model.

To see how DQM is useful in local alternatives, consider IID observations $X_1, \cdots, X_n$ from an unknown parametric density $p(x;\theta_0)$. The collection $\{p(x;\theta) : \theta \in \Theta\}$ is the parametric model we are considering. Consider a sequence $h_n \to 0$ and a sequence of models $p(x;\theta_0 + h_n)$ and the hypothesis testing problem:

$$H_0 : \theta = \theta_0, \quad H_{a,n} : \theta = \theta_0 + h_n.$$

Note that the alternative hypothesis changes when $n \to 0$.

When h_n converges too fast, we will not be able to distinguish the two hypothesis. When h_n converges too slow, the problem is trivial in the sense that we can easily distinguish the two hypothesis. However, there is an interesting regime where when h_n converges to 0 at a particular rate, we will see interesting results.

Specifically, consider the likelihood ratio test (since it is the uniformly most powerful test in the simple versus simple tests) and the test statistic

$$\lambda_n = \sum_{i=1}^{n} \log \frac{p(X_i;\theta_0 + h_n)}{p(X_i;\theta_0)}.$$

To obtain a regular asymptotic results in a likelihood model, we often assume that the likelihood function is twice-differentiable to obtain a reasonable Fisher's information matrix.

But here is an interesting fact: we do not need twice-differentiation in this local alternative case. What we need is that $p(x;\theta)$ is DQM in a neighborhood of θ_0.

To see this, note that the DQM in a multivariate case can be written as (using the form of equation (14))

$$\int \left(\sqrt{p(x;\theta_0 + h_n)} - \sqrt{p(x;\theta_0)} - \frac{1}{2} h_n^T \ell'(\theta_0|x) \sqrt{p(x;\theta_0)} \right)^2 dx = o(\|h_n\|^2),$$

where $\ell'(\theta|x) = \frac{\partial}{\partial\theta}\ell(\theta|x) = \frac{\partial}{\partial\theta}p(x;\theta)$ is the score of the likelihood model. Let

$$\delta_n(x) = \sqrt{p(x;\theta_0 + h_n)} - \sqrt{p(x;\theta_0)} - \frac{1}{2} h_n^T \ell'(\theta_0|x) \sqrt{p(x;\theta_0)}.$$

The DQM requires $\|\delta_n\|^2_{L_2(P)} = o(\|h_n\|^2)$.

Using the DQM, we can rewrite the test statistic as

$$\begin{aligned}
\lambda_n &= \sum_{i=1}^{n} \log \frac{p(X_i;\theta_0+h_n)}{p(X_i;\theta_0)} \\
&= 2\sum_{i=1}^{n} \log \sqrt{\frac{p(X_i;\theta_0+h_n)}{p(X_i;\theta_0)}} \\
&= 2\sum_{i=1}^{n} \log \frac{\sqrt{p(X_i;\theta_0)}+\frac{1}{2}h_n^T\ell'(\theta_0|X_i)\sqrt{p(X_i;\theta_0)}+\delta_n(X_i)}{\sqrt{p(X_i;\theta_0)}} \\
&= 2\sum_{i=1}^{n} \log \left(1+\frac{1}{2}h_n^T\ell'(\theta_0|X_i)+o_P(\|h_n\|)\right) \\
&\approx 2\sum_{i=1}^{n} \left(\frac{1}{2}h_n^T\ell'(\theta_0|X_i)-\frac{1}{4}h_n^T\ell'(\theta_0|X_i)\ell'^T(\theta_0|X_i)h_n\right) \\
&= \sqrt{n}h_n\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\ell'(\theta_0|X_i)-\frac{1}{2}\sqrt{n}h_n^T\left[\frac{1}{n}\sum_{i=1}^{n}\ell'(\theta_0|X_i)\ell'^T(\theta_0|X_i)\right]\sqrt{n}h_n.
\end{aligned}$$

Therefore, if we choose $h_n = n^{-1/2}c_0$ for some constant c_0, we obtain

$$\lambda_n = c_0\frac{1}{\sqrt{n}}Z_n - \frac{1}{2}c_0^T I_n(\theta_0)c_0,$$

where $Z_n \xrightarrow{d} N(0, I(\theta_0))$ and

$$I(\theta_0) = \mathbb{E}(\ell'(\theta_0|X_1)\ell'^T(\theta_0|X_1))$$

is the information matrix and

$$I_n(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\ell'(\theta_0|X_i)\ell'^T(\theta_0|X_i) \xrightarrow{p} I(\theta_0).$$

By the Slutsky's theorem,

$$\lambda_n \xrightarrow{d} N\left(-\frac{1}{2}c_0^T I(\theta_0)c_0, c_0^T I(\theta_0)c_0\right),$$

a regular distribution!

This gives us two conclusions:

- **Convergence rate of local alternatives.** The rate $h_n = n^{-1/2}c_0$ is the critical rate such that if a local alternative converges faster than this, then there is no way to distinguish them (H_0 and H_a) and if the local alternative converges slower than this, then we have an asymptotic probability 1 to distinguish them.
- **Differentiation in quadratic mean.** In our analysis, we did not assume that the likelihood function is twice differentiable. All we need is the DQM. Thus, DQM can be viewed as a weak condition that allows us to work on the neighborhood of the true model. This explains why many literature use DQM as a way to specify a parametric submodel.

6 Finding efficient estimators: conditional expectation

Let X be a random variable and let $Z = m(X)$ for some known fixed function m. Let P be the distribution of X and Q be the distribution of Z. The tangent space of X and the tangent space of Z are associated as follows.

Theorem 11 (Lemma 25.34 of van der Vaart 2000) *Assume the above notations. Suppose that P is DQM with respect to direction $g(x)$. Then Q is DQM with respect to direction $w(z)$*

$$w(z) = \mathbb{E}(g(X)|Z = z).$$

If we view $w(z) = Ag(z)$, where $A : L_2(P) \to L_2(Q)$ is an operator, then its adjoint $A^* : L_2(Q) \to L_2(P)$ can be expressed as

$$A^*h(x) = \mathbb{E}(h(Z)|X = x).$$

The power of Theorem 11 is that it converts any parametric submodel of X into a parametric submodel of $Z = m(X)$. This is useful because Z is often something that we observe (so it contains less information than X) but we want to consider a submodel over the full variable X.

Example: censoring. Consider a simple censoring problem where we observe (Y, Δ) such that $Y = \min\{T, C\}$ and $\Delta = I(Y = T)$ and T is the outcome of interest and C is the censoring time. We assume that both $T, C \geq 0$ and $T \perp C$. Let $p(y, \delta)$ be the PDF of the observed variables and $p_T(t)$ and $p_C(t)$ be the PDF of T and C. We use the subscript to denote the true model, i.e., $p_0(y, \delta)$ is the PDF that generates our data and $p_{T,0}$, $p_{C,0}$ are the true PDF of T and C.

Because T and C are independent, any parametric submodel of $p_{T,C}$ that preserves the independence can be decomposed into the a submodel of p_T and a submodel of p_C. Theorem 11 shows that if we can convert any parametric submodel of $p_T(t)$ into a parametric submodel of $p(y, \delta)$. Consider a parametric submodel

$$p_{T,\varepsilon}(t) = p_{T,0}(t)(1 + \varepsilon g(t)).$$

Then by Theorem 11, this submodel will imply the follow parametric submodel of $p_\varepsilon(y, \delta)$:

$$p_\varepsilon(y, \delta) = p_0(y, \delta)(1 + \varepsilon w(y, \delta)),$$

where

$$w(y, \delta) = \mathbb{E}(g(T)|Y = y, \Delta = \delta) = \delta g(y) + (1 - \delta)\frac{\int_y^\infty g(t)p_{T,0}(t)dt}{1 - F_{T,0}(y)}. \tag{16}$$

For the censoring variable, a similar analysis shows that if $p_C(c)$ is DQM with direction $q(c)$, then the $p(y, \delta)$ is DQM with direction $\xi(y, \delta)$

$$\begin{aligned}\xi(y, \delta) &= \mathbb{E}(q(C)|Y = y, \Delta = \delta) \\ &= (1 - \delta)q(y) + \delta\frac{\int_y^\infty q(c)p_{C,0}(c)dc}{1 - F_{C,0}(y)}.\end{aligned}$$

6.1 Finding a computable influence function and efficient estimator

In general, the parameter of interest β will be a statistical functional of the complete data X. We often have an RAL estimator with an influence function $\psi(x)$ for β. Theorem 11 immediately implies that for the observed data (Z), the corresponding influence function will be

$$\psi_*(z) = \mathbb{E}(\psi(X)|Z = z).$$

This is because by definition, an influence function has mean 0 so it can always be used as a submodel direction. Thus, this immediately gives us an estimator using the observed data. Alternatively, using the second statement of Theorem 11, another strategy of finding an influence function using the observed data is to find $\psi^\dagger(z)$ such that

$$\mathbb{E}(\psi^\dagger(Z)|X = x) = \psi(x).$$

Sometimes, the first strategy is useful and sometimes the second one is better. To illustrate the idea, consider the following current status model.

6.2 Example: current status model

In the current status model, there are two positive random variables T and Y. T is the outcome variable of interest but is unobserved. Y is a measurement time point that is always observed. The observed data is

$$(Y_1, \Delta_1), \cdots, (Y_n, \Delta_n),$$

where $\Delta = I(T \leq Y)$ is a status variable denoting if the event occurs before the measurement time (i.e., $T \leq Y$). We assume that T and Y are independent.

Let p_T, p_Y be the PDF of T and Y and P_T, P_Y be the corresponding CDFs. The observed-data density (PDF corresponding to the observed data) is

$$p(y, \delta) = p_Y(y) P_T(y)^\delta (1 - P_T(y))^{1-\delta},$$

The conditional densities

$$p(y, \delta | T = t) = \delta \frac{p_Y(y) I(y \geq t)}{1 - P_Y(t)} + (1 - \delta) \frac{p_Y(y) I(y < t)}{P_Y(t)},$$
$$p(t | Y = y, \Delta = \delta) = \delta \frac{p_T(t)}{P_T(y)} I(t \leq y) + (1 - \delta) \frac{p_T(t)}{1 - P_T(y)} I(t > y).$$

In this case, the complete-data distribution consists of both (T, Y) and the observed-data distribution consists of (Y, Δ). The independence assumption of $T \perp Y$ implies that

$$p_{T,Y}(t, y) = p_T(t) p_Y(y)$$

so any submodel direction $g(t, y)$ maintaining the independence relation must be decomposed as $g(t, y) = g_1(t) + g_2(y)$. Thus, we can analyze the submodel along p_T and p_Y individually.

Any submodel of $p_T(t)$ along direction $g(t)$ will be a submodel of $p(y,\delta)$ along direction

$$A_T g(y,\delta) = w_g(y,\delta) = \mathbb{E}(g(T)|Y=y,\Delta=\delta) = \delta\frac{\int_{t=0}^{t=y} g(t)dP_T(t)}{P_T(y)} + (1-\delta)\frac{\int_{t=y}^{t=\infty} g(t)dP_T(t)}{1-P_T(y)},$$

where A_T is a mapping from $L_2(P_T)$ to $L_2(P_{Y,\Delta})$. Because Y is observed, the submodel of $p_Y(y)$ along direction $g(y)$ will be a submodel of $p(y,\delta)$ along direction $g(y)$; this will corresponds to another mapping A_Y from $L_2(P_Y)$ to $L_2(P_{Y,\Delta})$. Note that

$$L_2(P_X) = \{g(x) : \mathbb{E}(g(X)) = 0, \mathbb{E}(g^2(X)) < \infty\}$$

is the collection of all possible submodel directions with respect to the distribution of X. Also note that since an influence function $\psi(Y,T)$ has mean 0, $\psi \in L_2(P_{Y,T})$ so it can always be used as a submodel direction.

Finding an influence function under the observed data. Now suppose that we want to estimate a statistical functional $\beta = \Omega(F_T) = \int \xi(t)dP_T(t)$ and let β_0 be the true parameter. Under the complete-data case (knowing T and Y), a natural estimator is

$$\widehat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n}\xi(T_i)$$

and this estimator has an influence function

$$\psi_0(t) = \xi(t) - \beta_0.$$

Because $\psi_0(t) \in L_2(P_T)$, we can use this function as the direction of a parametric submodel of P_T, which leads to the direction in the observed-data distribution $P_{Y,\Delta}$

$$\begin{aligned} A_T\psi_0(y,\delta) = w_{\psi_0}(y,\delta) &= \mathbb{E}(\psi_0(T)|Y=y,\Delta=\delta) \\ &= \mathbb{E}(\xi(T)|Y=y,\Delta=\delta) - \beta_0 \\ &= \delta\frac{\int_{t=0}^{t=y}\xi(t)dP_T(t)}{P_T(y)} + (1-\delta)\frac{\int_{t=y}^{t=\infty}\xi(t)dP_T(t)}{1-P_T(y)} - \beta_0, \end{aligned}$$

which leads to an RAL estimator

$$\widehat{\beta}_1 = \frac{1}{n}\sum_{i=1}^{n}\Delta_i\frac{\int_{t=0}^{t=Y_i}\xi(t)dP_T(t)}{P_T(Y_i)} + (1-\Delta_i)\frac{\int_{t=Y_i}^{t=\infty}\xi(t)dP_T(t)}{1-P_T(Y_i)}.$$

Clearly, if we know $p_T(t)$, then $\widehat{\beta}_1$ is indeed a consistent estimator under very mild assumptions so this approach indeed gives us an estimator. However, in general we do not know p_T so we cannot compute $\widehat{\beta}_1$.

To resolve this issue, we consider the adjoint mapping of A_T, i.e., a mapping from $L_2(P_{Y,\Delta})$ to $L_2(P_T)$. By Theorem 11, this corresponds to the conditional expectation $\mathbb{E}(\cdot|T=t)$. Specifically, we want to find the function $h(y,\delta)$ such that

$$\mathbb{E}(h(Y,\Delta)|T=t) = \xi(t).$$

Note that here we ignore β_0 in the influence function since it will not contribute to the calculation. A direct computation shows that

$$\xi(t) = \int h(y,\delta)p(y,\delta|T=t)dyd\delta = \int_0^t h(y,1)p_Y(y)dy + \int_t^\infty h(y,0)p_Y(y)dy.$$

Taking a derivative with respect to t in both sides, we obtain

$$\xi'(t) = h(t,1)p_Y(t) - h(t,0)p_Y(t).$$

There are many solutions to this equation (each of them corresponds to an influence function) and one simple choice is

$$h_c(t,\delta) = \delta\frac{\xi'(t)}{p_Y(t)}I(t \le c) - (1-\delta)\frac{\xi'(t)}{p_Y(t)}I(t > c) + \ell(c) \tag{17}$$

for any constant c and $\ell(c)$ is a normalizing constant. To ensure that $h_c(t,\delta)$ has mean β_0, we need to ensure that it has mean 0. This gives

$$\begin{aligned}
\beta_0 = \mathbb{E}(h_c(Y,\Delta)) &= \mathbb{E}\left(\Delta\frac{\xi'(Y)}{p_Y(Y)}I(Y \le c) - (1-\Delta)\frac{\xi'(Y)}{p_Y(Y)}I(Y > c) + \ell(c)\right) \\
&= \int_0^c \xi'(y)F_T(y)dy - \int_c^\infty \xi'(y)(1-F(y))dy + \ell(c) \\
&= \int_0^c \xi'(y)dy - \int_0^c \xi'(y)(1-F_T(y))dy - \int_c^\infty \xi'(y)(1-F(y))dy + \ell(c) \\
&= \int_0^c \xi'(y)dy - \int_0^\infty \xi'(y)(1-F(y))dy + \ell(c).
\end{aligned}$$

So we will choose $\ell(c) = -\int_0^c \xi'(y)dy - \xi(0) = -\xi(c)$.

Thus, an RAL estimator of β is

$$\widehat{\beta}_c = \frac{1}{n}\sum_{i=1}^n h_c(Y_i,\Delta_i) = \frac{1}{n}\sum_{i=1}^n \Delta_i\frac{\xi'(Y_i)}{p_Y(Y_i)}I(Y_i \le c) - (1-\Delta_i)\frac{\xi'(Y_i)}{p_Y(Y_i)}I(Y_i > c) - \xi(c).$$

Compared with $\widehat{\beta}_1$, the estimator $\widehat{\beta}_c$ is better because we can estimate p_Y easily using the observed data.

Geometry induced by independence. To find the efficient estimator, we need to find the characterization of $\mathcal{F}_{\mathsf{semi,obs}}$ and make the projection of $h_c(t,\delta)$ onto it. Let $\mathcal{F}_{\mathsf{semi,full}}$ be the semi-parametric tangent space of the complete data ($P_{Y,T}$). The fact that Y and T are independent implies that

$$\mathcal{F}_{\mathsf{semi,full}} = \mathcal{F}_{\mathsf{semi,T}} \oplus \mathcal{F}_{\mathsf{semi,Y}}.$$

The mappings A_T and A_Y are

$$A_T : \mathcal{F}_{\mathsf{semi,T}} \to \mathcal{F}_{\mathsf{semi,obs}}, \quad A_Y : \mathcal{F}_{\mathsf{semi,Y}} \to \mathcal{F}_{\mathsf{semi,obs}}.$$

Because the parameter of interest $\beta = \Omega(P_T)$ only depends on T, every influence function $g(t,y) = g(t) \in \mathcal{F}_{\mathsf{semi,T}}$ so the observed-data tangent space will be

$$\mathcal{F}_{\mathsf{semi,obs}} = \mathsf{Range}(A_T).$$

Here is an interesting note. For any element $w \in \mathsf{Range}(A_T)$, it can be written as $w(y,\delta) = \mathbb{E}(g(T)|Y = y,\Delta = \delta)$ for some function g. Now if we consider the adjoint map A_Y^*, we have

$$A_Y^* w(y) = \mathbb{E}(w(Y,\Delta)|Y = y) = \mathbb{E}(\mathbb{E}(g(T)|Y,\Delta)|Y = y) = \mathbb{E}(g(T)|Y = y) \overset{Y \perp T}{=} \mathbb{E}(g(T)) = 0.$$

The last equality follows from the fact that g is a submodel direction ($E(g(T)) = 0$). As a result,

$$\mathcal{F}_{\mathsf{semi,obs}} = \mathsf{Range}(A_T) = \mathsf{Kernel}(A_Y^*).$$

Finding the efficient influence function. Now we are in a good position to derive the efficient influence function. Recall from equation (17) that

$$\psi_c(y,\delta) = h_c(y,\delta) - \beta_0 = \delta\frac{\xi'(y)}{p_Y(y)}I(y \le c) - (1-\delta)\frac{\xi'(y)}{p_Y(y)}I(y > c) - \beta_0 - \xi(c)$$

is an influence function. We want to find its projection $\Pi(\psi_c|\mathcal{F}_{\mathsf{semi,obs}})$.

One trick (that we have used before) is to find $\omega_c \in \mathcal{F}_{\mathsf{semi,obs}}^{\perp}$ such that

$$\psi_c(y,\delta) - \omega_c(y,\delta) \in \mathcal{F}_{\mathsf{semi,obs}},$$

i.e.,

$$A_Y^*(\psi_c - \omega_c)(y) = \mathbb{E}(\psi_c(Y,\Delta) - \omega_c(Y,\Delta)|Y = y) = 0. \tag{18}$$

Instead of considering a generic function $\omega_c(y,\delta)$, we consider $\omega_c(y,\delta) = \omega_c(y)$ because for any function $w_g(y,\delta) = \mathbb{E}(g(T)|Y = y, \Delta = \delta)$ with $\mathbb{E}(g(T)) = 0$, we have

$$\mathbb{E}(w_g(Y,\Delta)\omega_c(Y)) = \mathbb{E}(\mathbb{E}(g(T)|Y,\Delta)\omega_c(Y)) = \mathbb{E}(\mathbb{E}(g(T)\omega_c(Y)|Y,T)) = \mathbb{E}(g(T)\omega_c(Y)) = \mathbb{E}(g(T))\mathbb{E}(\omega_c(Y)) = 0$$

so $\omega_c(y) \in \mathcal{F}_{\mathsf{semi,obs}}^{\perp}$. Thus, equation (18) becomes

$$0 = \mathbb{E}(\psi_c(Y,\Delta) - \omega_c(Y)|Y) = \mathbb{E}(\psi_c(Y,\Delta)|Y) - \omega_c(Y).$$

So we will choose

$$\omega_c(y) = \mathbb{E}(\psi_c(Y,\Delta)|Y = y) = \frac{F_T(y)\xi'(y)}{p_Y(y)} - \frac{\xi'(y)}{p_Y(y)}I(y > c) - \beta_0 - \xi(c)$$

and the projection is the efficient influence function

$$\psi_{\mathsf{eff}}(y,\delta) = \psi_c(y,\delta) - \omega_c(y,\delta) = (\delta - F_T(y))\frac{\xi'(y)}{p_Y(y)} = -\delta(1-F_T(y))\frac{\xi'(y)}{p_Y(y)} + (1-\delta)F_T(y)\frac{\xi'(y)}{p_Y(y)}.$$

With this, an efficient estimator can be constructed using the procedure described in Theorem 2.