Chapter 5

April 8, 2022

Outline

Outline

- In Chapter 3, we developed theoretical results for influence functions of parameters in finite-dimensional parametric models $\{f(X;\theta) : \theta \in \mathbb{R}^p\}$, p finite and where θ can be partitioned as

$$\theta = \left(\beta^T, \eta^T\right)^T : \beta \in \mathbb{R}^s, \eta \in \mathbb{R}^r, p = r + s$$

  with β the parameter of interest and η the nuisance parameter.
- In chapter 4 we considered influence functions of functionals $\psi(\theta) = \psi(F_\theta)$, where θ is an infinite dimensional parameter ranging in some functional space Θ, i.e. $\{f(X;\theta) : \theta \in \Theta\}$ is a nonparametric model
- In this chapter, we extend this theory to semiparametric modelds where the parameter $\theta = (\beta, \eta)$ can be partitioned into a finite parameter of interest β and an infinite dimensional nuisance parameter η.

- In this chapter we will consider semiparametric models that can be represented using the class of densities $\mathcal{M}_{F_{\theta^*}} = \{f(X;\theta) : \beta \in \mathbb{R}^s, \eta \in \Delta\}$, where β the parameter of interest and the infinite dimensional parameter η are variationally independent.
- That is, any choice of β and η in the neighborhood about the true β^* and η^* would result in a density $f(X;\theta)$ in the semiparametric model.
- This will allow us to define $\partial f(X;\theta)/\partial\beta|_{\theta^*}$.
- However keep in mind that some pbs lend themselves more naturally to models represented by the class of densities $f(X;\theta)$, where θ is infinite dimensional and $\psi(\theta)$ is an explicit functional of θ. We will make the distinction explicit when necessary.

- Example 1: Restricted moment model (See Chapter 1)

$$\begin{aligned} Y_i &= \mu(X_i,\beta) + \varepsilon_i, \\ \mathbb{E}(\varepsilon_i|X_i) &= 0 \end{aligned}$$

  $\mu(X_i,\beta)$ is known up to β, density of ε_i, η_1 has conditional mean zero given X_i and otherwise unrestricted and density η_2 of X_i is unrestricted. Thus $\theta = (\beta,\eta_1,\eta_2) \in \mathbb{R}^s \times \Gamma_1 \times \Gamma_2$.
- Example 2: Cox proportional hazards model (see Chapter 1)

$$\lambda(t|Z;\beta) = \lambda_0(t)\exp(\beta' Z)$$

  where $\lambda(t|Z;\beta)$ is the conditional hazard function of T at t given Z, let η_1 index the density of Z which is unrestricted, then:

$$\theta = (\beta,\lambda_0,\eta_1) \in \mathbb{R}^p \times \Gamma \times \Gamma_1,$$

  β is the regression parameter of interest (log hazards ratio), while $\lambda_0(t)$ is an unrestricted function of t.

Example 3: Semilinear or semiparametric regression model (see Chapter 1)

$$\mathbb{E}(Y|X,Z) = \beta^T X + \eta(Z),$$

β is the finite dimensional parameter of interest and the model is otherwise unrestricted,i.e. $\eta(Z)$ is unrestricted and joint density of X and Z is unrestricted.

- To compute influence functions in a semiparametric model, we define a regular parametric submodel

$$\mathcal{M}_{\text{sub},F_{\theta^*}} = \left\{ f\left(X;\theta = \left(\beta^T,\eta^T\right)^T\right) : \beta \in \mathbb{R}^s, \eta \in \mathbb{R}^r \right\}$$

  such that
  - $\mathcal{M}_{\text{sub},F_{\theta^*}} \subset \mathcal{M}_{F_{\theta^*}} = \{f(X;\theta = (\beta,\eta)) : \beta \in \mathbb{R}^s, \eta \in \Delta\}$ and
  - F_{θ^*} matches $f\left(X;\overline{\theta} = \left(\overline{\beta},\overline{\eta}\right)\right)$ for some $\left(\overline{\beta},\overline{\eta}\right)$ within $\mathbb{R}^s \times \mathbb{R}^r$
- As in Chapter 4, throughout we will assume the same regularity or smoothness conditions as where needed there for pathwise differentiation.

- One clarification: the term parametric *submodel* and parametric *model* can be confusing.
- A parametric model is a model whose probability densities are characterized through a finite number of parameters that the analyst believes will suffice to identify the unknown data generating mechanism
- In contrast, a parametric submodel is a conceptual tool used to develop theory for semiparametric models. Such parametric models cannot be used for data analysis as they require knowing what the true data generating mechanism is.

- Consider Example 2, a parametric submodel of the Cox proportional model may be taken to be: let $h_1(t)$ $...h_r$ (t) denote r different functions of t specified by the analyst

$$\mathcal{M}_{\text{sub},F_{\theta^*}} = \left\{ \begin{array}{c} \text{class of densities with hazard function:} \\ \lambda_0^*(t)\exp(\eta_1 h_1(t) + ... + \eta_r h_r(t))\exp(\beta' Z): \\ \beta \in \mathbb{R}^s, \eta \in \mathbb{R}^r \end{array} \right\}$$

- Because $\left(\beta^T, \eta^T\right)^T$ are unspecified, this moel is indeed finite dimensional
- For any choice of $\left(\beta^T, \eta^T\right)$, the resulting density follows a proportional hazards model and is therefore contained in the semiparametric model
- the truth is obtained at $\left(\beta^T, \eta^T\right) = \left(\beta^{*T}, 0^T\right)$
- The parametric submodel is "anchored" at the true model which is of course unknown and therefore such a parametric model is not useful for data analysis.

- In contrast, suppose one is willing to consider the parametric model

$$\lambda(t|Z;\beta) = \lambda_0 \exp(\beta' Z), \lambda_0 \text{ and } \beta \text{ unknown}$$

that is the baseline hazard is constant over time conditional on Z, i.e. the failure time is a conditional exponential model. This is a model which satisfies the proportional hazards assumption and may be used for data analysis, however if the model is incorrect, that is the true data generating model is not exponential, the resulting inferences may be completely incorrect and possibly meaningless.

Influence Functions for Semiparametric Estimators

- The nuisance tangent space for a semiparametric moel denoted by Λ_{nuis} is defined as the mean-square closure of parametric submodel nuisance tangent spaces, where a parametric submodel nuisance tangent space is the set of elements

$$\Lambda_{\text{sub,nuis}} = \left\{A^{s\times r} S_\eta(X;\theta^*)\right\}$$

$S_\eta(X;\theta^*) = \partial \log f(X;\beta^*,\eta)/\partial\eta|_{\eta^*}$ is the score vector for the nuisance parameter η for some parametric submodel assuming β^* is known, and $A^{s\times r}$ is a conformable matrix with s rows.

- Specifically the mean-square closure of the space above is defined as the space of functions $\Lambda_{\text{nuis}} = \{h(X) \text{ such that } \mathbb{E}\left(h^T(X)h(X)\right) < \infty\}$ and there exists a sequence of parametric submodels indexed by j such that

$$||h(X) - A_j^{s\times r} S_{\eta,j}(X;\theta^*)||^2 \overset{j\to\infty}{\longrightarrow} 0.$$

- Recall that the efficient influence function of β for a parametric submodel $\mathcal{M}_{\text{sub},F_{\theta^*}}$ with finite dimensional nuisance parameter η is given by

$$\varphi_{\beta}^{\text{eff}}(X) = \left\{\mathbb{E}S_{\beta}^{\text{eff}}S_{\beta}^{\text{eff}T}\right\}^{-1} S_{\beta}^{\text{eff}}(X;\beta^*,\eta^*),$$

where

$$S_{\beta}^{\text{eff}}(X;\beta^*,\eta^*) = S_{\beta}(X;\beta^*,\eta^*) - \Pi\left(S_{\beta}|\Lambda_{\text{sub,nuis}}\right),$$

the CR efficiency bound for estimating β in the submodel is

$$\left\{\mathbb{E}S_{\beta}^{\text{eff}}S_{\beta}^{\text{eff}T}\right\}^{-1}.$$

- The variance of any RAL semiparametric influence function must be greater or equal to

$$\left\{\mathbb{E}S_{\beta,\eta}^{\text{eff}}S_{\beta,\eta}^{\text{eff}\ T}\right\},^{-1}$$

for all parametric submodels.

- Hence the variance of any semiparametric influence function must be greater than or equal to

$$\mathcal{V} = \sup_{\mathcal{M}_{\text{sub},F_{\theta^*}}} \left\{\mathbb{E}S_{\beta,\eta}^{\text{eff}}S_{\beta,\eta}^{\text{eff}\ T}\right\}^{-1}.$$

- The supremum $\mathcal{V}$ is defined to be semiparametric efficiency bound. Any semiparametric estimator $\widehat{\beta}_n$ with asymptotic variance achieving this bound for $f^*(X) = f(X;\beta^*,\eta^*)$ is said to be semiparametric locally efficient at f^*. If the same estimator $\widehat{\beta}_n$ if semiparametric efficient regardless of f^*, then we say such estimator is globally semiparametric efficient.

Theorem 5.1. The Semiparametric efficiency bound $\mathcal{V}$ is equal to

$$\mathcal{V} = \left\{\mathbb{E}S_{\text{eff}}S_{\text{eff}}{}^{T}\right\}^{-1}$$

where

$$S_{\text{eff}}(X;\beta^*,\eta^*) = S_{\beta}(X;\beta^*,\eta^*) - \Pi\left(S_{\beta}|\Lambda_{\text{nuis}}\right)$$

is defined as the semiparametric efficient score for β.

Proof: For simplicity, take β to be a scalar parameter although this can easily be extended to $s > 1$. Since $\mathcal{M}_{\text{sub},F_{\theta^*}} \subset \mathcal{M}_{F_{\theta^*}}$, this implies that

$$\begin{aligned} & ||S_{\beta}(X;\beta^*,\eta^*) - \Pi\left(S_{\beta}|\Lambda_{\text{nuis}}\right)|| \\ \leq\ & ||S_{\beta}(X;\beta^*,\eta^*) - \Pi\left(S_{\beta}|\Lambda_{\text{sub,nuis}}\right)|| \end{aligned}$$

or

$$||S_{\text{eff}}(X;\beta^*,\eta^*)|| \leq ||S_{\beta,\eta}^{\text{eff}}||$$

for all $\mathcal{M}_{\text{nuis},F_{\theta^*}}$. Hence

$$||S_{\text{eff}}(X;\beta^*,\eta^*)||^{-2} \geq \sup_{\mathcal{M}_{\text{nuis},F_{\theta^*}}} ||S_{\beta,\eta}^{\text{eff}}||^{-2} = \mathcal{V}$$

To complete the proof requires also showing that

$$||S_{\text{eff}}(X;\beta^*,\eta^*)||^{-2} \leq \mathcal{V}$$

Because $\Pi\left(S_\beta\left(X\right)|\Lambda_{\text{nuis}}\right) \in \Lambda_{\text{nuis}}$, this means that there exists a sequence of parametric submodels with nuisance scores S_{η_j} and vectors A_j such that

$$||\Pi\left(S_\beta\left(X\right)|\Lambda_{\text{nuis}}\right) - A_j^T S_{\eta,j}\left(X;\theta^*\right)||^2 \overset{j\to\infty}{\longrightarrow} 0.$$

Therefore

$$\begin{aligned}
\mathcal{V}^{-1} &\leq ||S^{\text{eff}}_{\beta,\eta_j}||^2 \\
&= ||S_\beta\left(X;\beta^*,\eta^*\right) - \Pi\left(S_\beta|\Lambda_{\text{nuis},sub,j}\right)||^2 \\
&\leq ||S_\beta\left(X;\beta^*,\eta^*\right) - A_j^T S_{\eta,j}\left(X;\theta^*\right)||^2 \\
&= ||S_\beta\left(X;\beta^*,\eta^*\right) - \Pi\left(S_\beta\left(X\right)|\Lambda_{\text{nuis}}\right)||^2 \\
&\quad + ||\Pi\left(S_\beta\left(X\right)|\Lambda_{\text{nuis}}\right) - A_j^T S_{\eta,j}\left(X;\theta^*\right)||^2
\end{aligned}$$

taking $j \to \infty$ implies

$$\begin{aligned}
\mathcal{V}^{-1} &\leq ||S_\beta\left(X;\beta^*,\eta^*\right) - \Pi\left(S_\beta\left(X\right)|\Lambda_{\text{nuis}}\right)||^2 \\
&= ||S_{\text{eff}}\left(X;\beta^*,\eta^*\right)||^2
\end{aligned}$$

proving the result.

Definition 5.1. The efficient influence function is defined as the influence function of a semiparametric RAL estimator, if it exists, that achieves the semiparametric efficiency bound.

Theorem 5.2. Any RAL estimator for β must have an influence function $\varphi\left(X\right)$ that satisfies

$$(i)\ \mathbb{E}\left\{\varphi\left(X\right)S_\beta^T\left(X\right)\right\} = \mathbb{E}\left\{\varphi\left(X\right)S_{\text{eff}}\left(X;\beta^*,\eta^*\right)^T\right\} = I$$

and $(ii)\ \Pi\left(\varphi\left(X\right)|\Lambda_{\text{nuis}}\right) = 0$, i.e. $\varphi\left(X\right)$ is orthogonal to the nuisance tangent space. The efficient influence function is equal to

$$\begin{aligned}
\varphi_{\text{eff}}\left(X\right) &= \mathbb{E}\left\{S_{\text{eff}}S_{\text{eff}}^T\right\}^{-1} S_{\text{eff}}\left(X;\beta^*,\eta^*\right) \\
S_{\text{eff}}\left(X;\beta^*,\eta^*\right) &= S_\beta\left(X;\beta^*,\eta^*\right) - \Pi\left(S_\beta|\Lambda_{\text{nuis}}\right) \\
&= \Pi\left(S_\beta|\Lambda_{\text{nuis}}^\perp\right)
\end{aligned}$$

Proof of (ii) Note that from previous chapter under any parametric submodel

$$\frac{\partial\beta\left(F_t\right)}{\partial t} = \mathbb{E}\left\{\varphi\left(X\right)S_t\left(X\right)^T\right\}$$

taking t in the direction of η implies that

$$0 = \mathbb{E}\left\{\varphi\left(X\right)S_{\eta_t}\left(X\right)^T\right\}$$

for any nuisance submodel, which implies that for any sequence of submodels $S_{\eta,j}\left(X;\theta^*\right)$ and matrices A_j,

$$||h(X) - A_j^T S_{\eta,j}\left(X;\theta^*\right)||^2 \overset{j\to\infty}{\longrightarrow} 0 \text{ for } h \in \Lambda_{\text{nuis}}$$

$$0 = \mathbb{E}\left\{\varphi\left(X\right)\left[h\left(X\right) - A_j^T S_{\eta_t}\left(X\right)^T\right]\right\} + \mathbb{E}\left\{\varphi\left(X\right)h^T\left(X\right)\right\}$$

$$|\mathbb{E}\left\{\varphi\left(X\right)\left[h\left(X\right) - A_j^T S_{\eta_t}\left(X\right)^T\right]\right\}| \leq ||\varphi||||h - A_j^T S_{\eta_t}||$$

the result follows by letting $j \to \infty$.

Part (i) of the theorem follows from the fact that φ is the IF of a RAL estimator and (ii).

Theorem 5.3. If a semiparametric RAL estimator for β exiss, then the influence function of this estimator must belong to the space of influence functions, the linear variety $\left\{\varphi\left(X\right) + \Lambda^\perp\right\}$ where $\varphi\left(X\right)$ is the influence function of any semiparametric RAL estimator of β and Λ is the semiparametric tangent space, i.e. the closed linear span of all scores for the model, including score of β. Futhermore, if a RAL estimator for β exists that achieves the semiparametric efficiency bound, then the IF of this estimator must be the unique element $\varphi_{\text{eff}}\left(X\right) = \varphi\left(X\right) - \Pi\left(\varphi\left(X\right)|\Lambda^\perp\right) = \Pi\left(\varphi\left(X\right)|\Lambda\right)$ and the space of IF may be written $\varphi_{\text{eff}}\left(X\right) \oplus \Lambda^\perp \subseteq \Lambda_{\text{nuis}}^\perp$.

In the rest of the course, we will use the space of influence functions or even the space orthogonal to the nuisance tangent space for semiparametric models to construct "good" estimating equations of functionals. "good" in the sense that they have small bias and in many instances will given RAL estimators that are locally or globally efficient in the semiparametric model. We illustrate this fact with the familiar restricted mean model next, linking it to the class of so called generalized estimating equations (GEE).

- The semiparametric restricted mean model:

$$\begin{aligned} Y &= m(Z;\beta)+\varepsilon \\ \mathbb{E}\{\varepsilon|Z\} &= 0. \end{aligned}$$

- Note that the likelihood for one observation can be written

$$\begin{aligned} f_{Y,Z}(y,z;\beta,\eta) &= f_{\varepsilon,Z}(y-m(Z;\beta),z;\beta,\eta) \\ &= \eta_1(\varepsilon,z)\,\eta_2(z) \end{aligned}$$

where $\eta_1(\varepsilon,z)$ is the conditional density of ε given z is only restricted by conditions

$$\begin{aligned} \int \eta_1(\varepsilon,z)\,d\mu(\varepsilon) &= 1 \\ \int \varepsilon\eta_1(\varepsilon,z)\,d\mu(\varepsilon) &= 0 \\ \int \eta_2(z)\,d\mu(z) &= 1 \end{aligned}$$

- Result 5.1: As we will show below, the nuisance tangent space of RMM is given by:

$$\begin{aligned} \Lambda_{\text{nuis}} &= \Lambda_{\text{nuis},1}\oplus\Lambda_{\text{nuis},2}\subset L_2^0 \\ \Lambda_{\text{nuis},1} &= \left\{ \begin{array}{c} a_1(\varepsilon,Z):\mathbb{E}\{a_1(\varepsilon,Z)|Z\}=0 \\ \mathbb{E}\{\varepsilon a_1^T(\varepsilon,Z)|Z\}=0 \end{array} \right\}\cap L_2^0 \\ \Lambda_{\text{nuis},2} &= \{a_2(Z):\mathbb{E}\{a_2(Z)\}=0\}\cap L_2^0 \end{aligned}$$

- Result 5.2: The ortho-complement to the nuisance tangent space

$$\begin{aligned} \Lambda_{\text{nuis}}^{\perp} &= \Lambda_{\text{nuis},1}^{\perp}\cap\Lambda_{\text{nuis},2}^{\perp}\cap L_2^0 \\ &= \{h(Z)\,\varepsilon(\beta^*):h\}\cap L_2^0 \end{aligned}$$

- Result 5.3: The set of influence functions is therefore given by

$$\left\{\mathbb{E}\left\{h(Z)\,\varepsilon(\beta^*)\,S_\beta^T\right\}^{-1}h(Z)\,\varepsilon(\beta^*):h\right\}\cap L_2^0$$

where S_β^T is the score of β under the RMM.

- Result 5.4: The efficient score for β is given by

$$\begin{aligned} S_{\text{eff}}(Z;\beta^*,\eta^*) &= \Pi\left(S_\beta|\Lambda_{\text{nuis}}^\perp\right) \\ &= h^{opt}(Z)\,\varepsilon(\beta) \\ h^{opt}(Z) &= \mathbb{E}\left\{S_\beta\varepsilon^T(\beta)|Z\right\}\mathbb{E}\left\{\varepsilon(\beta)\,\varepsilon^T(\beta)|Z\right\}^{-1} \end{aligned}$$

- Proof Result 5.1. Consider a regular parametric submodel $\eta_t(\varepsilon,z) = \eta_{1,t}(\varepsilon,z)\,\eta_{2,t}(z)$ with unknown parameter $t \in (-\epsilon,\epsilon)$ such that

$$\eta_{t=0}(\varepsilon,z) = f(\varepsilon,z).$$

It is straightforward to verify that corresponding score equations may be written

$$\frac{\partial \log \eta_t(\varepsilon,z)}{\partial t} = a_1(\varepsilon,z) + a_2(z)$$

where $\mathbb{E}\{a_1(\varepsilon,Z)|Z\} = \mathbb{E}\{a_2(Z)\} = 0$ and thus, $\mathbb{E}\{a_1(\varepsilon,Z)\,a_2(Z)\} = 0$. The first term a_1 is a score for the conditional density of $\varepsilon|Z$, while a_2 is a score for the density of Z.

Furthermore, we also have that for all $t \in (-\epsilon,\epsilon)$:

$$\int \eta_{1,t}(\varepsilon,z)\,\varepsilon^T d\mu(\varepsilon) = 0,$$

therefore

$$\int \frac{\partial \eta_{1,t}(\varepsilon,z)}{\partial t}|_{t=0}\varepsilon^T d\mu(\varepsilon) = 0,$$

which in turn implies that $\mathbb{E}\left\{a_1(\varepsilon,Z)\,\varepsilon^T|Z\right\} = 0$ must hold for all scores $a_1(\varepsilon,Z)$ in the semiparametric model.

- From these results, we have that:

$$\begin{aligned} \Lambda_{\text{nuis}} &= \Lambda_{\text{nuis},1} \oplus \Lambda_{\text{nuis},2} \subset L_2^0 \\ \Lambda_{\text{nuis},1} &= \left\{ \begin{array}{c} a_1(\varepsilon,Z) : \mathbb{E}\{a_1(\varepsilon,Z)|Z\} = 0 \\ \mathbb{E}\left\{\varepsilon a_1^T(\varepsilon,Z)|Z\right\} = 0 \end{array} \right\} \cap L_2^0 \\ \Lambda_{\text{nuis},2} &= \{a_2(Z) : \mathbb{E}\{a_2(Z)\} = 0\} \cap L_2^0 \end{aligned}$$

- Technically, for each element of Λ_{nuis} we must also exhibit parametric submodels with corresponding score equal to the given element. I will leave it as hwk to work this out. See Tsiatis Section 4.5.

- Proof of Result 5.2. The ortho-complement to the nuisance tangent space

$$\begin{aligned}\Lambda_{\text{nuis}}^{\perp} &= (\Lambda_{\text{nuis},1} \oplus \Lambda_{\text{nuis},2})^{\perp} \\ &= \Lambda_{\text{nuis},1}^{\perp} \cap \Lambda_{\text{nuis},2}^{\perp} \cap L_2^0 \text{ (why?)}\end{aligned}$$

Note that any function $b(\varepsilon, Z) \in \Lambda_{\text{nuis},2}^{\perp} \cap L_2^0$ must satisfy

$$\mathbb{E}\left(b(Z,\varepsilon)\, a_2(Z)\right) = 0$$

for all $a_2(Z) \in \Lambda_{\text{nuis},2} \cap L_2^0$, that is $\mathbb{E}(b(Z,\varepsilon)|Z) = 0$. (why?)

- Therefore, to characterize $\Lambda_{\text{nuis}}^{\perp}$ entails finding all $b(\varepsilon, Z) \in \Lambda_{\text{nuis},2}^{\perp}$ such that $\mathbb{E}\left(b(X,\varepsilon)\, a_1(\varepsilon, Z)\right) = 0$ for all $a_1(\varepsilon, Z) \in \Lambda_{\text{nuis},1}^{\perp}$.
- Now recall that $a_1(\varepsilon, Z) \in \Lambda_{\text{nuis},1}$ implies $\mathbb{E}\{a_1(\varepsilon, Z)\,\varepsilon | Z = 0\}$; therefore, a natural candidate for $\Lambda_{\text{nuis},2}^{\perp}$ is given by

$$\{b(Z,\varepsilon) = h(Z)\,\varepsilon, h(Z) \text{ unrestricted}\} \cap L_2^0.$$

- To confirm that this is in fact correct, note that the subspace of $\Lambda_{\text{nuis},2}^{\perp}$ orthogonal to ε given Z is given by

$$\left\{C(\varepsilon, Z) - \mathbb{E}\left(C\varepsilon^T | Z\right)\mathbb{E}\left(\varepsilon\varepsilon^T | Z\right)\varepsilon : \mathbb{E}(C|Z) = 0\right\} \cap L_2^0.$$

Each element of this set is the residual of a projection of an element of $\{C(\varepsilon, Z) : \mathbb{E}(C|Z) = 0\} \cap L_2^0$ on the space of functions given by $\{b(X,\varepsilon) = h(Z)\,\varepsilon, h(Z) \text{ unrestricted}\} \cap L_2^0$ proving the result. In other words,

$$\Lambda_{\text{nuis},1} = \left\{C(\varepsilon, Z) - \mathbb{E}\left(C\varepsilon^T | Z\right)\mathbb{E}\left(\varepsilon\varepsilon^T | Z\right)\varepsilon : \mathbb{E}(C|Z) = 0\right\} \cap L_2^0$$

is an alternative characterization of $\Lambda_{\text{nuis},1}$.

- Proof of Result 5.4: Follows directly from definition of orthogonal projection. i.e.

$$\begin{aligned}S_{\text{eff}}(\varepsilon, Z; \beta^*, \eta^*) &= \Pi\left(S_\beta | \Lambda_{\text{nuis}}^{\perp}\right) \\ &= \Pi\left(S_\beta | \{b(Z,\varepsilon) = h(Z)\,\varepsilon, h(Z) \text{ unrestricted}\} \cap L_2^0\right) \\ &= \mathbb{E}\left\{S_\beta \varepsilon^T(\beta^*) | Z\right\}\mathbb{E}\left\{\varepsilon(\beta^*)\,\varepsilon^T(\beta^*) | Z\right\}^{-1}\varepsilon(\beta^*).\end{aligned}$$

Note also that for all submodels

$$\int \eta_1(\varepsilon(\beta), z)\,\varepsilon(\beta)^T\, d\mu(\varepsilon) = 0 \text{ for all } \beta$$

therefore

$$\begin{aligned}0 &= \int \frac{\partial \eta_1(\varepsilon(\beta), z)}{\partial \beta}|_{\beta^*}\varepsilon(\beta^*)^T\, d\mu(\varepsilon) \\ &\quad - \int \eta_1(\varepsilon(\beta^*), z)\,\frac{\partial m(z;\beta)^T}{\partial \beta}|_{\beta^*} d\mu(\varepsilon)\end{aligned}$$

which implies that

$$\mathbb{E}\left\{S_\beta \varepsilon^T(\beta^*) | Z\right\} = \frac{\partial m(Z;\beta)^T}{\partial \beta}|_{\beta^*}.$$

We conclude that

$$\begin{aligned}&S_{\text{eff}}(X; \beta^*, \eta^*) \\ &= \mathbb{E}\left\{S_\beta \varepsilon^T(\beta^*) | Z\right\}\mathbb{E}\left\{\varepsilon(\beta^*)\,\varepsilon^T(\beta^*) | Z\right\}^{-1}\varepsilon(\beta^*) \\ &= D^T(Z)\, V(Z)^{-1}\varepsilon(\beta^*)\end{aligned}$$

where

$$\begin{aligned}D(Z) &= \frac{\partial m(Z;\beta)}{\partial \beta^T}|_{\beta^*} \\ V(Z) &= \mathbb{E}\left\{\varepsilon(\beta)\,\varepsilon^T(\beta) | Z\right\}^{-1}\end{aligned}$$

Note also that by Result 5.3: The set of influence functions is given by

$$\left\{\mathbb{E}\left\{h\left(Z\right)\varepsilon\left(\beta^{*}\right)S_{\beta}^{T}\right\}^{-1}h\left(Z\right)\varepsilon\left(\beta^{*}\right):h\right\}\cap L_{2}^{0}$$

Using a similar argument as above one can show that

$$\mathbb{E}\left\{h\left(Z\right)\varepsilon\left(\beta\right)S_{\beta}^{T}\right\}=\mathbb{E}\left\{h\left(Z\right)D\left(Z\right)\right\}$$

and therefore the set of influence functions is given by

$$\left\{\mathbb{E}\left\{h\left(Z\right)D\left(Z\right)\right\}^{-1}h\left(Z\right)\varepsilon:h\right\}\cap L_{2}^{0}$$

and the efficient influence function is given by

$$\mathbb{E}\left\{D^{T}\left(Z\right)V\left(Z\right)^{-1}D\left(Z\right)\right\}^{-1}D^{T}\left(Z\right)V\left(Z\right)^{-1}\varepsilon$$

and the semiparametric efficiency bound is given by

$$\mathbb{E}\left\{D^{T}\left(Z\right)V\left(Z\right)^{-1}D\left(Z\right)\right\}^{-1}$$

Adaptive Semiparametric Estimators for the Restricted Mean Model

- The class of influence functions is particularly useful as it provides very good candidate estimating equations.
- Specifically, let $u_{h}\left(X;\beta\right)=U_{h}\left(\beta\right)=h\left(Z\right)\left(Y-m(Z;\beta)\right)=h\left(Z\right)\varepsilon\left(\beta\right)$ with h of dimensions $\dim\left(\beta\right)\times\dim\left(\varepsilon\right)$. Note that $\left\{U_{h}\left(\beta\right);h\right\}\cap L_{2}^{0}=\Lambda_{\text{nuis}}^{\perp}$. As we have previously shown, under fairly weak regularity conditions, the solution to equation

$$\sum_{i}U_{h,i}\left(\widehat{\beta}_{h}\right)=0 \tag{1}$$

for user specified h is a RAL estimator w influence function given by $\mathbb{E}\left\{h\left(Z\right)D\left(Z\right)\right\}^{-1}h\left(Z\right)\varepsilon\left(\beta\right)$. The estimating equations (1) is an example of so-called generalized estimating equations. Using semiparametric theory, we have shown that any RAL estimator must have influence function (asymptotically) equivalent to one in our class!!!

Adaptive Semiparametric Estimators for the Restricted Mean Model

- It is reasonable to restrict attention to the semiparametric efficient estimator which is given by $\widehat{\beta}_{h_{opt}}$ where

$$h_{opt}\left(Z\right)=D^{T}\left(Z\right)V\left(Z\right)^{-1}\varepsilon\left(\beta\right)$$

$$\begin{aligned}D(Z) &= \frac{\partial m(Z;\beta)}{\partial\beta^{T}}|_{\beta^{*}}\\ V\left(Z\right) &= \mathbb{E}\left\{\varepsilon\left(\beta^{*}\right)\varepsilon^{T}\left(\beta^{*}\right)|Z\right\}^{-1}\end{aligned}$$

Thus this estimator is not feasible as it requires $D(Z)$ which depends on β^{*} and on $V\left(Z\right)=var\left(Y|Z\right)$.

Adaptive Semiparametric Estimators for the Restricted Mean Model

- The dependence on β^{*} is not much of an issue, as one can substitute in an (inefficient) preliminary estimator of β say by solving the GEE with user specified choice of $h_{I}(Z)$. For example when $m(Z;\beta)=g^{-1}\left((1,Z')\beta\right)$ where g is a link function, then it is standard to take $h_{I}(Z)=(1,Z)$. We denote such an initial estimator $\widehat{\beta}_{I}$.
- Dependence of the efficient estimator on $var\left(Y|Z\right)=var\left(\varepsilon|Z\right)$ is potentially more challenging, since our semiparetric model allows the distribution of $\varepsilon|Z$ to be unrestricted. One may wish to proceed by estimating $var\left(Y|Z\right)$ nonparametrically using saying kernel smoothing or similar techniques, however in finite sample the resulting estimator is likely to be very unstable particularly if Z is multi dimensional.

Adaptive Semiparametric Estimators for the Restricted Mean Model

- An alternative approach is to specify a working model for $V(Z)$, say $V(Z;\beta,\gamma)$ with finite dimensional parameter γ. which in the case of scalar Y may be estimated by solving the estimating equation:

$$0=\sum_i T\left(Z_i,\widehat{\beta}_I,\gamma\right)\left\{(Y_i-m(Z;\widehat{\beta}_I)^2-V\left(Z;\widehat{\beta}_I,\gamma\right)\right\}$$

where $T\left(Z_i,\widehat{\beta}_I\right)$ is user-specified, e.g. $\partial V\left(Z;\widehat{\beta}_I,\gamma\right)/\partial\gamma$, and of the same dimension as γ. Denote the resulting estimator $\widehat{\gamma}$.

Adaptive Semiparametric Estimators for the Restricted Mean Model

- Replacing $V\left(Z;\widehat{\beta}_I,\widehat{\gamma}\right)$ for $V(Z)$ into $h_{opt}(Z)$ will result under mild regularity conditions, in a RAL estimator for β with influence function

$$\mathbb{E}\left\{D^T(Z)V(Z;\beta^*,\gamma^*)^{-1}D(Z)\right\}^{-1}D^T(Z)V(Z;\beta^*,\gamma^*)^{-1}\varepsilon$$

where γ^* is the probability limit of $\widehat{\gamma}$.

- Consequently, solutions to such estimating equations where one uses a working variance model is what is called a locally efficient estimator because the resulting estimator is RAL under the semiparametric model irrespective of whether the variance model is correct or not, but it only achieves the semiparametric efficiency bound if the working variance model is correctly specified.

Adaptive Semiparametric Estimators for the Restricted Mean Model

- One must take care when estimating the asymptotic variance of locally efficient estimator, note that its asymptotic variance is given by the variance of its influence function

$$\mathbb{E}\left\{D^T(Z)V(Z;\beta^*,\gamma^*)^{-1}D(Z)\right\}^{-1}$$
$$\mathbb{E}\left\{D^T(Z)V(Z;\beta^*,\gamma^*)^{-1}V(Z)V(Z;\beta^*,\gamma^*)^{-1}D(Z)\right\}$$
$$\mathbb{E}\left\{D^T(Z)V(Z;\beta^*,\gamma^*)^{-1}D(Z)\right\}^{-1}$$

the so-called sandwich variance estimator.

- This only simplifies to the SEB if the variance model is correct.

- Alternatively, one may use the nonparametric bootstrap.

Semiparametric location-shift model

- Suppose we observe iid data (Y,Z) Y is continuous that follows a semiparametric location-shift model

$$Y=m(Z;\beta^*)+\varepsilon,$$

where $m(Z,\cdot)$ is known, $\beta^*\in\mathbb{R}^s$ and $\varepsilon_i\perp Z_i$,

$$\eta(\varepsilon,Z)=\eta_1(\varepsilon)\eta_2(Z),$$

η_1 and η_2 unrestricted.

- We aim to make inferences about the finite dimensional parameter β^*.

- Note that for identification, we assume $m(0;\beta^*)=m(Z;0)=0$

Semiparametric location-shift model

- Result 5.5. As we will show below, the nuisance tangent space of LSM is given by:

$$\begin{aligned}\Lambda_{\text{nuis}} &= \Lambda_{\text{nuis},1} \oplus \Lambda_{\text{nuis},2} \subset L_2^0 \\ \Lambda_{\text{nuis},1} &= \{a_1(\varepsilon) : \mathbb{E}\{a_1(\varepsilon)\} = 0\} \cap L_2^0 \\ \Lambda_{\text{nuis},2} &= \{a_2(Z) : \mathbb{E}\{a_2(Z)\} = 0\} \cap L_2^0\end{aligned}$$

- Result 5.6: The ortho-complement to the nuisance tangent space

$$\begin{aligned}\Lambda_{\text{nuis}}^{\perp} &= \Lambda_{\text{nuis},1}^{\perp} \cap \Lambda_{\text{nuis},2}^{\perp} \cap L_2^0 \\ &= \left\{ \begin{array}{c} h(Z,\varepsilon) - \mathbb{E}\{h(Z,\varepsilon)|Z\} \\ -\mathbb{E}\{h(Z,\varepsilon)|\varepsilon\} + \mathbb{E}\{h(Z,\varepsilon)\} \end{array} : h \right\} \cap L_2^0\end{aligned}$$

- Result 5.7: The set of influence functions is therefore given by

$$\left\{ \begin{array}{c} \mathbb{E}\left\{\widetilde{h}(Z,\varepsilon) S_\beta^T\right\}^{-1} \widetilde{h}(Z,\varepsilon) \\ \widetilde{h}(Z,\varepsilon) = \left(\begin{array}{c} h(Z,\varepsilon) - \mathbb{E}\{h(Z,\varepsilon)|Z\} \\ -\mathbb{E}\{h(Z,\varepsilon)|\varepsilon\} + \mathbb{E}\{h(Z,\varepsilon)\} \end{array} \right) \end{array} : h \right\} \cap L_2^0$$

where S_β^T is the score of β under the LSM

- Result 5.8: The efficient score for β is given by

$$\begin{aligned}S_{\text{eff}}(X;\beta^*,\eta^*) &= \Pi\left(S_\beta|\Lambda_{\text{nuis}}^{\perp}\right) \\ &= \widetilde{h}_{opt}(Z,\varepsilon) \\ h_{opt}(Z,\varepsilon) &= \dot{m}(Z;\beta^*)\kappa(\varepsilon) \\ \dot{m}(Z;\beta^*) &= \frac{\partial m(Z;\beta^*)}{\partial\beta}|_{\beta^*} \\ \kappa(\varepsilon) &= -\frac{\partial \log \eta_1(\varepsilon)}{\partial\varepsilon}\end{aligned}$$

Because $\mathbb{E}\kappa(\varepsilon) = 0$ (show this), we have that

$$S_{\text{eff}}(X;\beta^*,\eta^*) = \left\{\dot{m}(Z;\beta^*) - \mathbb{E}\left(\dot{m}(Z;\beta^*)\right)\right\}\kappa(\varepsilon)$$

- Proof Result 5.5. Consider a regular parametric submodel $\eta_t(\varepsilon,z) = \eta_{1,t}(\varepsilon)\eta_{2,t}(z)$ with unknown parameter $t \in (-\epsilon,\epsilon)$ such that $\eta_{t=0}(\varepsilon,z) = f(\varepsilon,z)$. It is straightforward to verify that corresponding score equations may be written

$$\frac{\partial \log \eta_t(\varepsilon,z)}{\partial t} = a_1(\varepsilon) + a_2(z)$$

where $\mathbb{E}\{a_1(\varepsilon)\} = \mathbb{E}\{a_2(Z)\} = 0$ and $\mathbb{E}\{a_1(\varepsilon)a_2(Z)\} = 0$ by the independence assumption. The first term a_1 is a score for the conditional density of ε , while a_2 is a score for the density of Z.

- From these results, we have that:

$$\begin{aligned}\Lambda_{\text{nuis}} &= \Lambda_{\text{nuis},1} \oplus \Lambda_{\text{nuis},2} \subset L_2^0 \\ \Lambda_{\text{nuis},1} &= \{a_1(\varepsilon) : \mathbb{E}\{a_1(\varepsilon)\} = 0\} \cap L_2^0 \\ \Lambda_{\text{nuis},2} &= \{a_2(Z) : \mathbb{E}\{a_2(Z)\} = 0\} \cap L_2^0\end{aligned}$$

- Technically, for each element of Λ_{nuis} we must also exhibit parametric submodels with corresponding score equal to the given element. I will leave it as hwk to work this out.

- Proof of Result 5.6. The ortho-complement to the nuisance tangent space

$$\begin{aligned}\Lambda_{\text{nuis}}^{\perp} &= (\Lambda_{\text{nuis},1} \oplus \Lambda_{\text{nuis},2})^{\perp} \\ &= \Lambda_{\text{nuis},1}^{\perp} \cap \Lambda_{\text{nuis},2}^{\perp} \cap L_2^0 \text{ (why?)}\end{aligned}$$

Note that any function $b(\varepsilon, Z) \in \Lambda_{\text{nuis},2}^{\perp} \cap L_2^0$ must satisfy

$$\mathbb{E}(b(Z,\varepsilon)\, a_2(Z)) = 0$$

for all $a_2(Z) \in \Lambda_{\text{nuis},2} \cap L_2^0$, that is $\mathbb{E}(b(Z,\varepsilon)|Z) = 0$. (why?)

- Therefore, to characterize $\Lambda_{\text{nuis}}^{\perp}$ entails finding all $b(\varepsilon, Z) \in \Lambda_{\text{nuis},2}^{\perp}$ such that $\mathbb{E}(b(X,\varepsilon)\, a_1(\varepsilon)) = 0$ for all $a_1(\varepsilon) \in \Lambda_{\text{nuis},1}^{\perp}$; that is $\mathbb{E}(b(X,\varepsilon)|\varepsilon)) = 0$.
- Therefore, $\Lambda_{\text{nuis}}^{\perp} = \{b(\varepsilon, Z) : \mathbb{E}(b(X,\varepsilon)|\varepsilon) = \mathbb{E}(b(Z,\varepsilon)|Z) = 0\} \cap L_2^0$.

- We conjecture that

$$\Lambda_{\text{nuis}}^{\perp} = \left\{ \begin{pmatrix} h(Z,\varepsilon) - \mathbb{E}\{h(Z,\varepsilon)|Z\} \\ -\mathbb{E}\{h(Z,\varepsilon)|\varepsilon\} + \mathbb{E}\{h(Z,\varepsilon)\} \end{pmatrix} \right\} \cap L_2^0$$

- Proving the result entails showing that

$$\Lambda_{\text{nuis}}^{\perp} \supseteq \left\{ \begin{pmatrix} h(Z,\varepsilon) - \mathbb{E}\{h(Z,\varepsilon)|Z\} \\ -\mathbb{E}\{h(Z,\varepsilon)|\varepsilon\} + \mathbb{E}\{h(Z,\varepsilon)\} \end{pmatrix} \right\} \cap L_2^0$$

and

$$\Lambda_{\text{nuis}}^{\perp} \subseteq \left\{ \begin{pmatrix} h(Z,\varepsilon) - \mathbb{E}\{h(Z,\varepsilon)|Z\} \\ -\mathbb{E}\{h(Z,\varepsilon)|\varepsilon\} + \mathbb{E}\{h(Z,\varepsilon)\} \end{pmatrix} \right\} \cap L_2^0$$

- The first inclusion is easy to verify while the second follows from the fact that any $h(Z,\varepsilon)$ in $\Lambda_{\text{nuis}}^{\perp}$ must satisfy $\mathbb{E}\{h(Z,\varepsilon)|Z\} = \mathbb{E}\{h(Z,\varepsilon)|\varepsilon\} = 0$.

- Proof of Result 5.8: Follows directly from previous result on how to project onto $\Lambda_{\text{nuis}}^{\perp}$

$$\begin{aligned}& S_{\text{eff}}(X; \beta^*, \eta^*) \\ =\ & \Pi\left(S_\beta | \Lambda_{\text{nuis}}^{\perp}\right) \\ =\ & S_\beta - \mathbb{E}\{S_\beta|Z\} - \mathbb{E}\{S_\beta|\varepsilon\} + \mathbb{E}\{S_\beta\}\end{aligned}$$

- To show this, it suffices to note that $S_\beta - \mathbb{E}\{S_\beta|Z\} - \mathbb{E}\{S_\beta|\varepsilon\} + \mathbb{E}\{S_\beta\} \in \Lambda_{\text{nuis}}^{\perp}$ and

$$\begin{aligned}& S_\beta - \Pi\left(S_\beta | \Lambda_{\text{nuis}}^{\perp}\right) \\ =\ & \mathbb{E}\{S_\beta|Z\} + \mathbb{E}\{S_\beta|\varepsilon\} - \mathbb{E}\{S_\beta\}\end{aligned}$$

is orthogonal to $\Lambda_{\text{nuis}}^{\perp}$.

- Next, note that the score for β is given by

$$\begin{aligned} S_\beta &= \frac{\partial \log \eta_1(\varepsilon(\beta))}{\partial \beta} \\ &= -\frac{\partial \log \eta_1(\varepsilon(\beta))}{\partial \beta} \dot{m}(Z;\beta^*) \\ &= \dot{m}(Z;\beta^*)\kappa(\varepsilon) \end{aligned}$$

We conclude that

$$\begin{aligned} S_{\text{eff}}(X;\beta^*,\eta^*) &= \dot{m}(Z;\beta^*)\kappa(\varepsilon) - \mathbb{E}\left\{\dot{m}(Z;\beta^*)\kappa(\varepsilon)|Z\right\} \\ &\quad -\mathbb{E}\left\{\dot{m}(Z;\beta^*)\kappa(\varepsilon)|\varepsilon\right\} + \mathbb{E}\left\{\dot{m}(Z;\beta^*)\kappa(\varepsilon)\right\} \\ &= \left\{\dot{m}(Z;\beta^*) - \mathbb{E}\left\{\dot{m}(Z;\beta^*)\right\}\right\}\kappa(\varepsilon) \end{aligned}$$

by independence and the fact that $\kappa(\varepsilon)$ has mean zero (why is that?)

- Note also that the set of influence functions is given by

$$\left\{ \begin{array}{c} \mathbb{E}\left\{\widetilde{h}(Z,\varepsilon) S_\beta^T\right\}^{-1} \widetilde{h}(Z,\varepsilon) \\ \widetilde{h}(Z,\varepsilon) = \begin{pmatrix} h(Z,\varepsilon) - \mathbb{E}\{h(Z,\varepsilon)|Z\} \\ -\mathbb{E}\{h(Z,\varepsilon)|\varepsilon\} + \mathbb{E}\{h(Z,\varepsilon)\} \end{pmatrix} : h \end{array} \right\} \cap L_2^0$$

Note that because $\mathbb{E}_\beta\left\{\widetilde{h}(Z,\varepsilon(\beta))\right\} = 0$ for all β, then

$$\mathbb{E}\left\{\widetilde{h}(Z,\varepsilon) S_\beta^T\right\} = -\mathbb{E}\left\{\frac{\partial \widetilde{h}(Z,\varepsilon)}{\partial \beta^T}\right\}$$

and therefore influence functions are of the form

$$-\mathbb{E}\left\{\frac{\partial \widetilde{h}(Z,\varepsilon)}{\partial \beta^T}\right\}\widetilde{h}(Z,\varepsilon)$$

- Therefore the efficient influence function is given by

$$\begin{aligned} &\mathbb{E}\left\{\left\{\dot{m}(Z;\beta^*) - \mathbb{E}\left\{\dot{m}(Z;\beta^*)\right\}\right\}^{\otimes 2}\kappa(\varepsilon)^2\right\}^{-1} \\ &\times\left\{\dot{m}(Z;\beta^*) - \mathbb{E}\left\{\dot{m}(Z;\beta^*)\right\}\right\}\kappa(\varepsilon) \end{aligned}$$

and the semiparametric efficiency bound is given by

$$\mathbb{E}\left\{\kappa(\varepsilon)^2\right\}^{-1}\left[\mathbb{E}\left\{\left\{\dot{m}(Z;\beta^*) - \mathbb{E}\left\{\dot{m}(Z;\beta^*)\right\}\right\}^{\otimes 2}\right\}\right]^{-1}$$

Adaptive Semiparametric Estimators for the Location Shift Model

- Note that as in the case of the restricted mean model, constructing a semiparametric estimator based on the efficient score requires knowing the density of ε or more precisely knowing $\kappa(\varepsilon)$ .Consider a feasible semiparametric efficient estimator which is given by $\widehat{\beta}_{\widehat{h}_{opt}}$ which solves

$$\mathbb{P}_n\left\{\dot{m}(Z;\beta) - \mathbb{P}_n\left\{\dot{m}(Z;\beta)\right\}\right\}\widehat{\kappa}\left(Y - m(Z;\widehat{\beta}_{\widehat{h}_{opt}})\right) = 0$$

where we have replace $\mathbb{E}\left\{\dot{m}(Z;\beta)\right\}$ with its sample version $\mathbb{P}_n\left\{\dot{m}(Z;\beta)\right\}$ and κ with an estimator $\widehat{\kappa}$.

Adaptive Semiparametric Estimators for the Location Shift Model

- Ideally, one may wish to proceed by estimating $\kappa(\varepsilon) = -\partial \log \eta_1(\varepsilon) / \partial \varepsilon$ in order to remain in the semiparametric model which does not restrict the latter. however in finite sample the resulting estimator is likely to be very unstable unless sample size is very large, particularly because it's the ratio of a derivative of density divided by a density , i.e. density appears in the denominator.

Adaptive Semiparametric Estimators for the for the Location Shift Model

- An alternative approach is to specify a working model for $\eta_1(\varepsilon)$, say $\eta_1(\varepsilon;\gamma)$ with finite dimensional parameter γ. which may be estimated by standard maximum likelihood, i.e. by solving :

$$\widehat{\gamma} = \arg\max_{\gamma} \mathbb{P}_n \log \eta_1\left(\varepsilon\left(\widehat{\beta}_I\right);\gamma\right)$$

where $\widehat{\beta}_I$ is an inefficient initial estimator of β solving

$$\mathbb{P}_n\left[\widetilde{h}\left(\varepsilon\left(\widehat{\beta}_I\right), Z\right)\right] = 0$$

for user-specified h, e.g. $h\left(\varepsilon\left(\widehat{\beta}_I\right), Z\right) = \varepsilon\left(\widehat{\beta}_I\right)\dot{m}(Z;\beta)$

Adaptive Semiparametric Estimators for the for the Location Shift Model

- Replacing $\eta_1(\varepsilon;\widehat{\gamma})$ for $\eta_1(\varepsilon;\widehat{\gamma})$ into $\widetilde{h}_{opt}$ will result under mild regularity conditions, in a RAL estimator for β with influence function

$$\mathbb{E}\left\{\left\{\dot{m}(Z;\beta^*) - \mathbb{E}\left\{\dot{m}(Z;\beta^*)\right\}\right\}\frac{\partial\kappa(\varepsilon;\gamma^*)}{\partial\beta^T}|_{\beta^*}\right\}^{-1}$$
$$\times\left\{\dot{m}(Z;\beta^*) - \mathbb{E}\left\{\dot{m}(Z;\beta^*)\right\}\right\}\{\kappa(\varepsilon;\gamma^*) - \mathbb{E}\{\kappa(\varepsilon;\gamma^*)\}\}$$

where γ^* is the probability limit of $\widehat{\gamma}$, You will show this as homework.

Adaptive Semiparametric Estimators for the for the Location Shift Model

- Consequently, solutions to such estimating equations where one uses a model is a locally efficient estimator because the resulting estimator is RAL under the semiparametric model irrespective of whether the working model fpr the density of the uncentered residual is correct or not, but it only achieves the semiparametric efficiency bound if the working variance model is correctly specified.
- Note that resulting influence function has (up to a scale) an extra term $\left\{\dot{m}(Z;\beta^*) - \mathbb{E}\left\{\dot{m}(Z;\beta^*)\right\}\right\}\mathbb{E}\{\kappa(\varepsilon;\gamma^*)\}$, This term is technically the price one pays for having to estimate $\mathbb{E}\left\{\dot{m}(Z;\beta^*)\right\}$, however this term is exactly zero at the submodel where $\kappa(\varepsilon;\gamma^*)$ and therefore $\mathbb{E}\{\kappa(\varepsilon;\gamma^*)\} = \mathbb{E}\{\kappa(\varepsilon)\} = 0$ but not otherwise.

Cox Proportional Hazards Model with Censored Data

- Suppose we wish to make inferences about the log hazards ratio parameter β in a Cox PH model

$$\begin{aligned}\lambda_{T|Z}(t|z) &= \lim_{h\to 0}\left\{\frac{\Pr(t\leq T\leq t+h|T\geq t, Z=z)}{h}\right\} \\ &= \lambda_0(t)\exp\left\{\beta^T z\right\}\end{aligned}$$

where $\lambda_0(t)$ is an unrestricted function of t.

Cox Proportional Hazards Model with Censored Data

- However, we observe data (V, Δ, Z) where

$$V = \min(T, C)$$

with T the underlying failure time and C time to censoring, or time on study, and

$$\Delta = I(T\leq C)$$

is the failure indicator, and throughout we assume conditional independent censoring

$$T\perp C|Z$$

which is needed for identification.

Cox Proportional Hazards Model with Censored Data

- The goal is to find semiparametric consistent and asymptotically normal and efficient estimators for β using data (V_i, Z_i, Δ_i) without making any additional assumptios on the underlyng baseline hazard function λ_0, on the conditional density of C given Z or on the distribution of Z.
- We will use the semiparametric theory we have developed to find the space of all influence functions of RAL estimators for β in this model which will motivate a class of RAL estimators among which we will derive the semiparametric efficient estimator.
- This will be accomplished by deriving the semiparametric nuisance tangent space and its orthogonal complement when influence functions lie and the efficient score.

Cox Proportional Hazards Model with Censored Data

- The density for a single observation is given by (check this)

$$\begin{aligned}p_{V,\Delta,Z}(V,\delta,Z) &= \left\{\lambda_0(v)\exp\left(\beta^T z\right)\right\}^{\delta}\exp\left\{-\Lambda_0(v)\exp\left(\beta^T z\right)\right\} \\ &\quad\times\left\{p_{C|Z}(v|z)\right\}^{1-\delta}\left\{\int_v^{\infty} p_{C|Z}(u|z)\,du\right\}^{\delta} p_Z(z)\end{aligned}$$

where

$$\Lambda_0(v) = \int_0^v \lambda_0(u)\,du$$

is the cumulative baseline hazard function; likewise, let $\lambda_{C|Z}(v|z)$ and $\Lambda_{C|Z}(v|z)$ denote the conditional hazard and cumulative hazard functions of C given Z.

Cox Proportional Hazards Model with Censored Data

- The density may then be written

$$\begin{aligned} p_{V,\Delta,Z}(V,\delta,Z) &= \left\{\lambda_0(v)\exp\left(\beta^T z\right)\right\}^{\delta}\exp\left\{-\Lambda_0(v)\exp\left(\beta^T z\right)\right\} \\ &\quad\times\left\{\lambda_{C|Z}(v|z)\right\}^{1-\delta}\left\{\exp\left\{-\Lambda_{C|Z}(v|z)\right\}\right\}p_Z(z) \end{aligned}$$

- The semiparametric model is characterized by an $r-$dimensional parameter β of primary interest, and a infinite dimensional nuisance parameter $\eta = \left(\lambda_0(v), \lambda_{C|Z}(v|z), p_Z(z)\right)$ where $\lambda_0(v)$ is an unrestricted positive function of v, $\lambda_{C|Z}(v|z)$ is an unrestricte positive function of v and z, and $p_Z(z)$ is an unrestricted density of z such that $\int p_Z(z)\,dz = 1$.
- The goal is to characterize the orthogonal complement of the nuisance tangent space which contains all influence functions of regular and asymptotically linear estimators of β, and to characterize the efficient score for β.

Cox Proportional Hazards Model with Censored Data

- Let $N(u) = I(T \leq u, \Delta = 1)$ denote the indicator of whether the person is observed to have an event prior to time u and $Y(u) = I(V \geq u)$ denote the indicator of being at risk at time u. The corresponding martingale increment is

$$dM(u) = dN(u) - \lambda_0(u)\exp\left(\beta^{*T} z\right)Y(u)du.$$

- Likewise, let $N_C(u) = I(C \leq u, \Delta = 0)$ denote the indicator of whether the person is observed to be censored prior to time u and $Y(u) = I(V \geq u)$ denote the indicator of being at risk at time u. The corresponding martingale increment is

$$dM_C(u) = dN_C(u) - \lambda_{C|Z}(u|z)\,Y(u)du.$$

Cox Proportional Hazards Model with Censored Data

- Result 5.9.The nuisance tangent space

$$\Lambda_{\text{nuis}} = \Lambda_1 \oplus \Lambda_2 \oplus \Lambda_3$$

where

$$\begin{aligned} \Lambda_1 &= \left\{ \begin{array}{c} \int a_1(u)dM(u): \\ \text{for all } r\text{-dimensional functions } a_1(u) \end{array} \right\} \cap L_2^0 \\ \Lambda_2 &= \left\{ \begin{array}{c} \int a_2(u,Z)dM_C(u): \\ \text{for all } r\text{-dimensional functions } a_2(u,z) \end{array} \right\} \cap L_2^0 \\ \Lambda_3 &= \left\{ \begin{array}{c} a_3(Z) \\ \text{for all } r\text{-dimensional functions } a_3(z) \text{ with} \\ \mathbb{E}\left(a_3(Z)\right) = 0 \end{array} \right\} \cap L_2^0 \end{aligned}$$

Cox Proportional Hazards Model with Censored Data

- Result 5.10.The orthogonal complement to the nuisance tangent space is given by

$$\begin{aligned} \Lambda_{\text{nuis}}^{\perp} &= \Lambda_1^{\perp} \cap \Lambda_2^{\perp} \cap \Lambda_3^{\perp} \\ &= \left\{ \begin{array}{c} \int \left[a(u,Z) - a^{\dagger}(u)\right] dM(u): \\ \text{for all } r\text{-dimensional functions } a(u,Z) \end{array} \right\} \cap L_2^0 \end{aligned}$$

where $a^{\dagger}(u) = \dfrac{\mathbb{E}\left\{a(u,Z)\exp\left(\beta^{*T}Z\right)Y(u)\right\}}{\mathbb{E}\left\{\exp\left(\beta^{*T}Z\right)Y(u)\right\}}$

The efficient score of β is obtained for the choice $a(u,Z) = Z, i.e.$

$$\begin{aligned} S_{\beta}^{\text{eff}} &= \Pi\left(S_{\beta}|\Lambda_{\text{nuis}}^{\perp}\right) \\ &= \int \left[Z - \frac{\mathbb{E}\left\{Z\exp\left(\beta^{*T}Z\right)Y(u)\right\}}{\mathbb{E}\left\{\exp\left(\beta^{*T}Z\right)Y(u)\right\}}\right] dM(u) \end{aligned}$$

Cox Proportional Hazards Model with Censored Data

- Proof of Result 5.9. Recall that the likelihood for one observation is given by

$$\begin{aligned} p_{V,\Delta,Z}(V,\delta,Z) &= \left\{\lambda_0(v)\exp\left(\beta^T z\right)\right\}^{\delta}\exp\left\{-\Lambda_0(v)\exp\left(\beta^T z\right)\right\} \\ &\quad\times\left\{\lambda_{C|Z}(v|z)\right\}^{1-\delta}\left\{\exp\left\{-\Lambda_{C|Z}(v|z)\right\}\right\}p_Z(z) \end{aligned}$$

- Consider a parametric submodel in the direction of $\lambda_{C|Z}(v|z)$ assuming the rest of the likelihood is known

$$\left\{\lambda_{C|Z,\theta}(v|z)\right\}^{1-\delta}\left\{\exp\left\{-\Lambda_{C|Z,\theta}(v|z)\right\}\right\},$$

such that $\lambda_{C|Z,0}(v|z)=\lambda_{C|Z}(v|z)$. Let $\nabla_\theta()=\partial()/\partial\theta|_{\theta=0}$.

Cox Proportional Hazards Model with Censored Data

- The corresponding score equation is given by

$$\begin{aligned} &(1-\Delta)\frac{\nabla_\theta\lambda_{C|Z,\theta}(V|Z)}{\lambda_{C|Z,\theta}(V|Z)}-\int_0^V\frac{\nabla_\theta\lambda_{C|Z,\theta}(u|Z)}{\lambda_{C|Z,\theta}(u|Z)}\lambda_{C|Z,\theta}(u|z)\,du \\ =\;&(1-\Delta)\frac{\nabla_\theta\lambda_{C|Z,\theta}(V|Z)}{\lambda_{C|Z,\theta}(V|Z)} \\ &-\int\frac{\nabla_\theta\lambda_{C|Z,\theta}(u|Z)}{\lambda_{C|Z,\theta}(u|Z)}\lambda_{C|Z,\theta}(u|z)\,Y(u)\,du \\ =\;&\int\frac{\nabla_\theta\lambda_{C|Z,\theta}(u|Z)}{\lambda_{C|Z,\theta}(u|Z)}dM_C(u) \end{aligned}$$

Cox Proportional Hazards Model with Censored Data

- Since the Cox semiparametric model for $T|Z$ places no restriction on $\nabla_\theta\lambda_{C|Z,\theta}(u|z)$, the function $\nabla_\theta\lambda_{C|Z,\theta}(u|z)/\lambda_{C|Z,\theta}(u|z)$ can be any function of (v,z) and therefore the tangent space for the censoring mechanism is given by

$$\Lambda_2=\left\{\begin{array}{c}\int a_2(u,Z)dM_C(u):\\ \text{for all } r\text{-dimensional functions } a_2(u,z)\end{array}\right\}\cap L_2^0$$

- Next consider submodels $\lambda_{0,\theta}(v)$ in the direction of $\lambda_0(v)$, the corresponding likelihood contribution is given by

$$\left\{\lambda_{0,\theta}(v)\exp\left(\beta^{*T}z\right)\right\}^{\delta}\exp\left\{-\Lambda_{0,\theta}(v)\exp\left(\beta^T z\right)\right\}$$

with score equation given by

$$\begin{aligned} &\Delta\frac{\nabla_\theta\lambda_{0,\theta}(V)}{\lambda_{0,\theta}(V)}-\int_0^V\frac{\nabla_\theta\lambda_{0,\theta}(u)}{\lambda_{0,\theta}(u)}\exp\left(\beta^T Z\right)\lambda_{0,\theta}(u)\,du \\ =\;&\int\frac{\nabla_\theta\lambda_{0,\theta}(u)}{\lambda_{0,\theta}(u)}dM(u) \end{aligned}$$

Cox Proportional Hazards Model with Censored Data

- Since the Cox semiparametric model for $T|Z$ places no restriction on the baseline hazard, the function $\nabla_\theta\lambda_{0,\theta}(u)/\lambda_{0,\theta}(u)$ can be any function of v therefore the tangent space for the baseline hazard function is given by

$$\Lambda_1=\left\{\begin{array}{c}\int a_1(u)dM(u,Z):\\ \text{for all } r\text{-dimensional functions } a_1(u)\end{array}\right\}\cap L_2^0$$

- It is straightforward to show that scores for the density of Z span the space

$$\Lambda_3=\{\text{All } r\text{-dimensional functions } a_3(Z)\}\cap L_2^0$$

Cox Proportional Hazards Model with Censored Data

- Therefore the nuisance tangent space is given by

$$\Lambda_{\text{nuis}} = \Lambda_1 + \Lambda_2 + \Lambda_3$$

- Note however that because $M_C(u)$ and $M(u)$ are martingales, $0 = \mathbb{E}\left(M_C(u)|Z\right) = \mathbb{E}\left(M(u)|Z\right)$.
- Furthermore, because T and Z are conditionally independent, it is straightforward to show that

$$\mathbb{E}\left\{\int a_1(u)dM(u,Z)\int a_2(u,Z)dM_C(u)\right\} = 0$$

for all $a_1(u)$ and $a_2(u,Z)$. Together, these properties imply that

$$\Lambda_{\text{nuis}} = \Lambda_1 \oplus \Lambda_2 \oplus \Lambda_3$$

proving the result.

Cox Proportional Hazards Model with Censored Data

- Proof of Result 5.10. The orthogonal complement to the nuisance tangent space is given by

$$\Lambda_{\text{nuis}}^{\perp} = \Lambda_1^{\perp} \cap \Lambda_2^{\perp} \cap \Lambda_3^{\perp}$$

- We first show that

$$\Lambda_2^{\perp} \cap \Lambda_3^{\perp} = \left\{ \begin{array}{c} \int \left[a\left(u,Z\right)\right] dM(u) : \\ \text{for all } r\text{-dimensional functions } a\left(u,Z\right) \end{array} \right\}$$

- To see this, consider the nonparametric model where $\lambda_{T|Z}\left(t|z\right)$ is allowed to remain unrestricted. Then it is possible to show that the model places no restriction on the joint density of the observed data $p(v,\delta,z)$.

Cox Proportional Hazards Model with Censored Data

- One can use the previous proof to show that in this model, the *maximal* tangent space is given by

$$\Lambda^* = \Lambda_1^* \oplus \Lambda_2 \oplus \Lambda_3 = L_2^0$$

where

$$\Lambda_1^* = \left\{ \begin{array}{c} \int a\left(u,Z\right) dM(u) : \\ \text{for all } r\text{-dimensional functions } a\left(u,Z\right) \end{array} \right\} \cap L_2^0$$

Therefore $\Lambda_1^* = \left(\Lambda_2 \oplus \Lambda_3\right)^{\perp} = \Lambda_2^{\perp} \cap \Lambda_3^{\perp}$.

- Finally, we have that $\Lambda_{\text{nuis}}^{\perp}$ is the subspace of Λ_1^* also orthogonal to Λ_1,

$$\left\{ \begin{array}{c} \int a^*\left(u,Z\right) dM(u) : \text{for all } a^*\left(u,Z\right) \text{ st} \\ \mathbb{E}\left\{\int a^*\left(u,Z\right) dM(u) \int a^*\left(u\right) dM(u)\right\} = 0 \text{ for all } a^*\left(u\right) \end{array} \right\} \cap L_2^0$$

Cox Proportional Hazards Model with Censored Data

- Note that all such functions can be written $\int \left\{a\left(u,Z\right) - a^{\dagger}(u)\right\} dM(u)$ for arbitrary $a(u,Z)$ where $\int \left\{a^{\dagger}(u)\right\} dM(u)$ is the projection of $\int a\left(u,Z\right) dM(u)$ onto the Hilbert subspace $\left\{\int a(u)dM(u) : a\left(u\right)\right\} \cap L_2^0$.
- That is for any choice of $a(u,Z)$,

$$\mathbb{E}\left\{\int \left\{a\left(u,Z\right) - a^{\dagger}(u)\right\} dM(u) \int a(u)dM(u)\right\} = 0, \text{ for all } a$$

Cox Proportional Hazards Model with Censored Data

- Using the fact that the covariance of martingale stochastic integrals is given by the expectaction of the predictable covariance process, one can show that this implies that for all a

$$\begin{aligned}\mathbb{E}\left\{\int\left\{a\left(u,Z\right)-a^{\dagger}(u)\right\}a(u)Y(u)\lambda_0\left(u\right)\exp\left(\beta^{*T}Z\right)du\right\} &= 0, \text{ that is}\\ \mathbb{E}\left\{\left\{a\left(u,Z\right)-a^{\dagger}(u)\right\}Y(u)\exp\left(\beta^{*T}Z\right)\right\} &= 0\end{aligned}$$

proving the result

$$a^{\dagger}(u)=\frac{\mathbb{E}\left\{a\left(u,Z\right)Y(u)\exp\left(\beta^{*T}Z\right)\right\}}{\mathbb{E}\left\{Y(u)\exp\left(\beta^{*T}Z\right)\right\}}$$

and therefore

$$\begin{aligned}\Lambda_{\text{nuis}}^{\perp} &= \Lambda_1^{\perp}\cap\Lambda_2^{\perp}\cap\Lambda_3^{\perp}\\ &= \left\{\begin{array}{c}\int\left[a\left(u,Z\right)-a^{\dagger}\left(u\right)\right]dM(u):\\ \text{for all } r\text{-dimensional functions } a\left(u,Z\right)\end{array}\right\}\cap L_2^0\end{aligned}$$

Cox Proportional Hazards Model with Censored Data

- A globally semiparametric efficient estimator is obtained by deriving an empirical efficient estimating function

$$\int\left[Z-\frac{\mathbb{P}_n\left\{Z\exp\left(\beta^TZ\right)Y\left(u\right)\right\}}{\mathbb{P}_n\left\{\exp\left(\beta^TZ\right)Y\left(u\right)\right\}}\right]dM(u)$$

where recall $dM(u)=\left\{dN(u)-\lambda_0\left(u\right)\exp\left(\beta^Tz\right)Y(u)du\right\}$. The empirical estimating equation simplifies

$$0=\mathbb{P}_n\int\left[Z-\frac{\mathbb{P}_n\left\{Z\exp\left(\beta^TZ\right)Y\left(u\right)\right\}}{\mathbb{P}_n\left\{\exp\left(\beta^TZ\right)Y\left(u\right)\right\}}\right]dN(u)$$

so that it does depend on the baseline hazard function, which is remarkable. why?

- The solution to this equation is therefore globally semiparametric efficient. Why?

Efficiency in randomized trials w baseline covariates

- Consider a double-blind placebo controlled randomized trial to evaluate the average causal effect of an intervention A which is binary on an outcome Y.
- By randomization the average causal effect on the additive scale is given by

$$\beta=\mathbb{E}(Y|A=1)-\mathbb{E}\left(Y|A=0\right)$$

the difference in average outcome between the two treatment arms.

- Throughout we will assume that $\Pr(A=1)=1/2$, although other randomization schemes are easily handled.
- Suppose that in addition to A and Y, as routinely done in randomized trials, we also observe a large collection of baseline characteristics L.

Efficiency in randomized trials w baseline covariates

- By randomization, we have that treatment assignment is independent of all pretreatment characteristics, i.e. $A\perp L$.
- Our semiparametric model $\mathcal{M}$ is therefore the set of all regular law $f_{Y,A,L}(y,a,l)$, such that

$$f_{Y,A,L}(y,a,l)=f_{Y|A,L}(y|a,l)f_{A|L}(a)f_L(l)$$

- Our goal is to characterize all influence functions and therefore all RAL estimators of $\beta=\mathbb{E}(Y|A=1)-\mathbb{E}\left(Y|A=0\right)$ in this model.
- Additionally, we wish to characterize the efficient influence function in this model

Efficiency in randomized trials w baseline covariates

- Tangent space of the model:
- Result 5.11. The tangent space of this semiparametric model is given by

$$\Lambda = \Lambda_1 \oplus \Lambda_2 \oplus \Lambda_3$$

where

$$\begin{aligned}
\Lambda_1 &= \{b_1(Y,A,L) : \mathbb{E}(B_1|A,L) = 0\} \cap L_2^0 \\
\Lambda_2 &= \{b_2(A) : \mathbb{E}(B_2|L) = \mathbb{E}(B_2) = 0\} \cap L_2^0 \\
\Lambda_3 &= \{b_3(L) : \mathbb{E}(B_3) = 0\} \cap L_2^0
\end{aligned}$$

- Result 5.12. The orthocomplement to the tangent space is given by

$$\begin{aligned}
\Lambda^{\perp} &= \Lambda_1^{\perp} \cap \Lambda_2^{\perp} \cap \Lambda_3^{\perp} \\
&= \left\{ \left(A - \frac{1}{2} \right) (b(L) - \mathbb{E}(b(L))) : b(L) \right\} \cap L_2^0
\end{aligned}$$

Efficiency in randomized trials w baseline covariates

- Influence functions:
- Consider the nonparametric model $\mathcal{M}_{np}$ given by the set of all regular law $f_{Y,A,L}(y,a,l)$ where A and L are no longer assumed to be independent. As this is a nonparametric model the following result provides the unique influence function of $\beta = \mathbb{E}(Y|A=1) - \mathbb{E}(Y|A=0)$
- Result 5.13. The efficient IF of β in the nonparametric model $\mathcal{M}_{np}$ is given by

$$IF_{\beta,np} = 2I(A=1)(Y-\mu_1) - 2I(A=0)(Y-\mu_0)$$

- Result 5.14. The set of IFs of β in the semiparametric model $\mathcal{M}$ is given by

$$IF_{\beta,np} \oplus \Lambda^{\perp} = \left\{ \begin{array}{c} 2I(A=1)(Y-\mu_1) - 2I(A=0)(Y-\mu_0) \\ + \left(A - \frac{1}{2}\right)(b(L) - \mathbb{E}(b(L))) \end{array} : b(L) \right\} \cap L_2^0$$

where $\mu_0 = \mathbb{E}(Y|A=0), \mu_1 = \mathbb{E}(Y|A=1)$.

Efficiency in randomized trials w baseline covariates

- Efficient IF:
- Result 5.15. The efficient IF of β in the semiparametric model $\mathcal{M}$ is given by

$$\begin{aligned}
IF_{\beta}^{\text{eff}} &= \Pi(IF_{\beta,np}|\Lambda) \\
&= \Pi(IF_{\beta,np}|\Lambda_1 \oplus \Lambda_2 \oplus \Lambda_3) \\
&= \Pi(IF_{\beta,np}|\Lambda_1) \oplus \Pi(IF_{\beta,np}|\Lambda_2) \oplus \Pi(IF_{\beta,np}|\Lambda_3) \\
&= 2I(A=1)(Y - E(Y|A=1,L)) \\
&\quad + E(Y|A=1,L) - \mu_1 \\
&\quad - 2I(A=0)(Y - E(Y|A=0,L)) \\
&\quad - E(Y|A=0,L) + \mu_0
\end{aligned}$$

Efficiency in randomized trials w baseline covariates

- Efficient IF:
- The efficient IF of β in the semiparametric model $\mathcal{M}$ may also be written

$$\begin{aligned}
& IF_{\beta}^{\text{eff}} \\
&= 2I(A=1)(Y-\mu_1) - (2I(A=1)-1)(E(Y|A=1,L)-\mu_1) \\
&\quad -2I(A=0)(Y-\mu_0) + (2I(A=0)-1)(E(Y|A=0,L)-\mu_0) \\
&= 2I(A=1)(Y-\mu_1) - 2I(A=0)(Y-\mu_0) \\
&\quad -(2I(A=1)-1)(E(Y|A=1,L) - E(Y|A=0,L) - \mu_0 + \mu_1)
\end{aligned}$$

which is in the set of influence functions

$$IF_{\beta,np} \oplus \Lambda^{\perp} = \left\{ \begin{array}{c} 2I(A=1)(Y-\mu_1) - 2I(A=0)(Y-\mu_0) \\ + \left(A - \frac{1}{2}\right)(b(L) - \mathbb{E}(b(L))) \end{array} : b(L) \right\} \cap L_2^0$$

with $b(L) = b^{opt}(L) = -2(E(Y|A=1,L) - E(Y|A=0,L))$.

Efficiency in randomized trials w baseline covariates

- Efficient IF:
- Now we note that $IF_{\beta,np} = IF_{\beta,np} - \Pi\left(IF_{\beta,np}|\Lambda^{\perp}\right) + \Pi\left(IF_{\beta,np}|\Lambda^{\perp}\right)$
- Because $IF_{\beta,np} - \Pi\left(IF_{\beta,np}|\Lambda^{\perp}\right) \in \Lambda$ and $\Pi\left(IF_{\beta,np}|\Lambda^{\perp}\right) \in \Pi\left(IF_{\beta,np}|\Lambda^{\perp}\right)$, we conclude that

$$IF_{\beta}^{\text{eff}} = IF_{\beta,np} - \Pi\left(IF_{\beta,np}|\Lambda^{\perp}\right)$$

which implies that

$$\begin{aligned} & \Pi\left(IF_{\beta,np}|\Lambda^{\perp}\right) \\ = & \left(I\left(A=1\right) - \frac{1}{2}\right) \\ & \times 2\left(E\left(Y|A=1,L\right) - E\left(Y|A=0,L\right) - \beta\right) \end{aligned}$$

- This further implies that

$$\begin{aligned} var\left(IF_{\beta}^{\text{eff}}\right) &= var\left(IF_{\beta,np} - \Pi\left(IF_{\beta,np}|\Lambda^{\perp}\right)\right) \\ &\leq var\left(IF_{\beta,np}\right) \end{aligned}$$

Efficiency in randomized trials w baseline covariates

- Efficient IF:
- It is straightforward to establish that $IF_{\beta,np}$ is the influence function of the nonparametric estimator of β given by

$$\widehat{\beta} = \frac{\mathbb{P}_n(YA)}{\mathbb{P}_n(A)} - \frac{\mathbb{P}_n(Y\left(1-A\right))}{\mathbb{P}_n(\left(1-A\right))}$$

Check this.

- As a consequence, we have established that the difference in sample means which is the standard to analyze randomized trial is inefficient when as ususally the case baseline characteristics are available in a randomized trial.

Efficiency in randomized trials w baseline covariates

- Efficient IF:
- A semiparametric locally efficient estimator can be obtained by using the EIF as an estimating equation,
- First note that

$$\begin{aligned} IF_{\beta}^{\text{eff}} &= 2I\left(A=1\right)\left(Y - E(Y|A=1,L)\right) \\ & -2I\left(A=0\right)\left(Y - E(Y|A=0,L)\right) \\ & +E\left(Y|A=1,L\right) - E\left(Y|A=0,L\right) - \beta \end{aligned}$$

depends on the unknown conditional mean function $E\left(Y|A,L\right)$ which we must estimate in order to obtain a feasible estimating function. Suppose we posit a simple parametric model $b(A,L;\eta)$, say

$$b(A,L;\eta) = \left(1, A, AL', L'\right)\eta$$

and suppose that η is estimated by OLS.

Efficiency in randomized trials w baseline covariates

- Efficient IF:
- $\widetilde{\beta}$ can be shown to be consistent and asymptotically normal in model $\mathcal{M}$ and to achieve the semiparametric efficiency bound if $b(1,L;\eta)$ is correctly specified, so that $b(a,l;\widehat{\eta})$ converges to $\mathbb{E}\left(Y|a,l\right)$ for all (a,l) in probability.
- That is $\widetilde{\beta}$ is completely robust to model misspecification of $b(1,L;\eta)$ and achieve the efficiency bound if $b(1,L;\eta)$ is correct model. In other words $\widetilde{\beta}$ is a semiparametric locally efficient estimator.
- One may be tempted to use a nonparametric model for $b(1,L;\eta)$, this will typically only perform well if L is very low dimensional, e.g.continuous w one or at most 2 components, or discrete w few levels.

Outline

Observational study: The case of point exposure

- Suppose that randomization no longer holds, because the observed data comes from a point exposure/cross-sectional observational study, with observed data $\{L, A, Y\}$.
- L is a rich vector of covariates that satisfies :
  (NUCA) No unmeasured confounding assumption holds: $\{Y_0, Y_1\} \perp A|L$ where Y_a is the potential outcome had exposure been a
- Then we say that there are no unmeasured counfounders for the effect of A on Y.

Observational study: The case of point exposure

- The intuition behind (NUCA) is similar to that of RA. Mainly, that we have measured enough covariates L, so that within levels of L, the data mimicks a randomized trial with the randomization probabilities now allowed to depend on L.
- Conceptually, this can be achieved only if we are able to measure all common causes of A and Y (that is all risk factors for Y that also determine A).

Observational study: The case of point exposure

- Next we show that the no unmeasured confounding assumption is sufficient to again identify $\{E(Y_a) : a\}$ and thus $\psi = E(Y_1) - E(Y_0)$
  Without loss of generality, suppose L is categorical; then
  $E(Y_a) = E(E(Y_a|L)) = \sum_l E(Y_a|L = l) f_L(l)$
  $\overset{NUCA}{=} \sum_l E(Y_a|A = a, L = l) f_L(l)$
  $\overset{CA}{=} \sum_l E(Y|A = a, L = l) f_L(l)$
  $\equiv g(a)$

Observational study: The case of point exposure

- $g(a)$ is known as the *direct standardization* of $E(Y|A=a,L)$. It is a special case of Robins' *G-formula* (which we will discuss in the longitudinal case).
- Thus

$$\psi = g(1) - g(0) = \sum_l \{E(Y|A=1,L=l) - E(Y|A=0,L=l)\} f_L(l)$$

is the *standardized risk difference*.

- Under NUCA, we see that crude association$\neq$causation, as
$\sum_l E(Y_a|A=a,L=l) f_L(l) = E(Y_a)$
$\neq E(Y|A=a) = \sum_l E(Y|A=a,L=l) f_L(l|A=a)$.

Observational study: The case of point exposure

- So that the crude risk difference does not have a causal interpretation. However, if NUCA holds, and either of the following conditions holds:

$$Y \amalg L|A \text{ or } A \amalg L \tag{2}$$

then $E(Y_a) = E(Y|A=a)$ and L is a non-confounder, so that this implies that RA actually holds.

Observational study: The case of point exposure

Proof:

- In the first case,
$E(Y_a) = \sum_l E(Y|A=a,L=l) f_L(l)$
$= \sum_l E(Y|A=a) f_L(l) = E(Y|A=a)$;
- In the second case,
$E(Y_a) = \sum_l E(Y|A=a,L=l) f_L(l)$
$= \sum_l E(Y|A=a,L=l) f_L(l|A=a) = E(Y|A=a)$

Observational study: The case of point exposure

- In general,the point exposure G-formula is written

$$E(Y_a) = \int E(Y|A=a,L=l)\, dF(l)$$

- The left-hand side is the mean of a counterfactual (latent variable), the right–hand side is a functional of the observed data, which is always well defined but only has a causal interpretation under the no unmeasured assumption.

Observational study: The case of point exposure

- This functional is not a conditional expectation of the observed data, therefore, cannot be estimated directly such as the crude. However, the pluggin principle may be used as discussed below.
- The G-formula is not restricted to the mean,other versions:
  - $E\left(Y_a|V=v\right)=\int E\left(Y|A=a,L=l\right)dF(l|V=v)$, where V is contained in L
  - $f\left(Y_a|V=v\right)=\int f\left(Y|A=a,L=l\right)dF(l|V=v)$.

G-computation

- Given the observed data $O_i=(Y_i,A_i,L_i)$, G-computation generally refers to nonparametric inference on the G-formula

$$g(a)=\sum_l E\left(Y|A=a,L=l\right)f_L\left(l\right).$$

- A natural nonparametric estimator of $g(a)$ is given by the nonparametric pluggin estimator, which requires nonparametric estimates of $E\left(Y|A=a,L=l\right)=b(a,l)$ written $\widehat{b}(a,l)$ and of $F_L\left(l\right)$ written $\widehat{F}_L\left(l\right)$.

G-computation

- Until otherwise stated, assume both A and L are categorical variables with low to moderate number of levels, so that $\widehat{b}(a,l)$ is given by the stratified sample average:

$\widehat{b}(a,l)=\sum_{i=1}^n I\left(A_i=a,L_i=l\right)Y_i/\sum_{i=1}^n I\left(A_i=a,L_i=l\right)$

and $\widehat{f}_L\left(l\right)=n^{-1}\sum_{i=1}^n I\left(L_i=l\right)$

- The nonparametric estimator of the G-formula is given by:

$$\begin{aligned}\widehat{g}(a) &= \sum_l \widehat{b}(a,l)\widehat{f}_L\left(l\right)=\sum_l \widehat{b}(a,l)n^{-1}\sum_{i=1}^n I\left(L_i=l\right)\\ &= n^{-1}\sum_{i=1}^n\sum_l \widehat{b}(a,l)I\left(L_i=l\right)=n^{-1}\sum_{i=1}^n \widehat{b}(a,L_i)\end{aligned}$$

G-computation: Mathematical Interlude

- How to use $\widehat{g}(a)$ to make inference on $\psi=g\left(1\right)-g\left(0\right)$?
- e.g. Want 95%CI of the average causal effect of A on Y.
- Wald type CI:
  - $\widehat{\psi}\pm 1.96\sqrt{\widehat{var}\left(\widehat{\psi}\right)}$, where $\widehat{\psi}=\widehat{g}(1)-\widehat{g}(0)$ and $\widehat{var}\left(\widehat{\psi}\right)$ is a consistent estimator of the variance of $\widehat{\psi}$.
  - Coverage of this CI depends on the asymptotic distribution of $\widehat{\psi}$ and on identifying a reasonable estimator of $\widehat{var}\left(\widehat{\psi}\right)$.
  - Difficulty, $\widehat{g}(a)=n^{-1}\sum_{i=1}^n\widehat{b}(a,L_i)$ is a sample average, however of terms that are estimated using $\{Y_i,A_i,L_i:i=1,...,n\}$ so that though $b(a,L_i)$ and $b(a,L_j),i\neq j$ are independent, $\widehat{b}(a,L_i)$ and $\widehat{b}(a,L_j),i\neq j$ are highly dependent and standard $i.i.d$ Central Limit Theorem doesn't apply.

- However it is still possible to show asymptotic normality of $\widehat{\psi}$;in fact, this is an instructive exercise, write:

$$n^{1/2}\widehat{g}(a) = n^{-1/2}\sum_l \sum_{i=1}^n \widehat{b}(a,l) I(L_i = l)$$
$$= n^{-1/2}\sum_l \sum_{i=1}^n I(L_i = l)\left\{\frac{\sum_{s=1}^n I(A_s=a, L_s=l) Y_s}{\sum_{j=1}^n I(A_j=a, L_j=l)}\right\}$$
$$= n^{-1/2}\sum_l \sum_{s=1}^n I(A_s = a, L_s = l) Y_s \left\{\frac{\sum_{i=1}^n I(L_i=l)}{\sum_{j=1}^n I(A_j=a, L_j=l)}\right\}$$

$$= n^{-1/2}\sum_l \sum_{s=1}^n \left\{ \begin{array}{c} I(A_s = a, L_s = l)(Y_s - b(a,l)) \\ \times \left\{\frac{n^{-1}\sum_{i=1}^n I(L_i=l)}{n^{-1}\sum_{j=1}^n I(A_j=a, L_j=l)}\right\} \end{array} \right\}$$

$$+n^{-1/2}\sum_l b(a,l)\left\{\sum_{i=1}^n I(L_i = l)\right\}$$

$$= n^{-1/2}\sum_l \sum_{s=1}^n I(A_s = a, L_s = l)(Y_s - b(a,l)) f_{A|L}^{-1}(a|l)$$

$$+n^{-1/2}\sum_l b(a,l)\left\{\sum_{i=1}^n I(L_i = l)\right\} + o_p(1)$$

$$= n^{-1/2}\sum_{i=1}^n \frac{I(A_i=a)}{f_{A|L}(A_i|L_i)}(Y_i - b(A_i, L_i)) + b(a, L_i) + o_p(1)$$

so that we finally obtain

$$\begin{aligned} & n^{1/2}\left(\widehat{g}(a) - g(a)\right) \\ = \;& n^{-1/2}\sum_{i=1}^n \left\{\frac{I(A_i = a)}{f_{A|L}(A_i|L_i)}(Y_i - b(A_i, L_i)) + b(a, L_i) - g(a)\right\} + o_p(1) \\ = \;& n^{-1/2}\sum_{i=1}^n IF_i(a) + o_p(1) \\ \sim \;& N\left(0, E\left(IF_i(a)^2\right)\right) \end{aligned}$$

A wald type 95%CI for $\psi = g(1) - g(0)$ is given by:

$$\widehat{g}(1) - \widehat{g}(0) \pm 1.96\sqrt{n^{-1}\sum_i \left(\widehat{IF}_i(0) - \widehat{IF}_i(1)\right)^2}$$

where $\widehat{IF}_i(a) = \left\{\frac{I(A_i=a)}{\widehat{f}_{A|L}(a|L_i)}\left(Y_i - \widehat{b}(A_i, L_i)\right) + \widehat{b}(a, L_i) - \widehat{g}(a)\right\}$,

and $\widehat{f}_{A|L}(a|L_i = l) = \frac{\sum_i I(A_i=a, L_i=l)}{\sum_i I(L_i=l)}$ is the nonparametric estimator of $f_{A|L}(a|L_i = l)$,the probability of receiving treatment $A = a$ given $L = l$.

- The term $b(a, L_i) - g(a)$ reflects the variability due to the estimation of $F_L(l)$, whereas $\frac{I(A_i=a)}{f_{A|L}(a|L_i)}(Y_i - b(a, L_i))$ captures the variability due to the estimation of $b(a, L_i)$.
- It is interesting to note that the influence function and thus the variance of $\widehat{\psi} = \widehat{g}(1) - \widehat{g}(0)$ depends on the treatment process $\{f_{A|L}(a|L_i) : a\}$,even though the pluggin estimator described above appears not to.
- In fact, we give a different representation of the nonparametric estimator of the G-formula which makes explicit its dependence on the estimated treatment process.

Summary

- Goal: Make inference on the causal mean difference: $\psi = E(Y_1 - Y_0) = E(Y_1) - E(Y_0)$
  - where Y_a, $a = 0, 1$ is the counterfactual/potential outcome that a patient would experience if possibly countering to fact he/she receives treatment $a = 0, 1$.
  - $E(Y_a)$, $a = 0, 1$ has the interpretation of the population outcome mean that we would observe if everyone in the population was forced to take treatment $a = 0, 1$.
  - This is to be contrasted with $E(Y|A = a)$ which is the outcome mean among those who did take treatment $A = a$, which may or may not equal $E(Y_a)$.

- Identifying assumptions:
  - Randomization:$\{Y_0, Y_1\} \amalg A \Rightarrow E(Y_a) = E(Y|A = a) \Leftrightarrow$**crude association is causal**.
  - No unmeasured confounding: $\{Y_0, Y_1\} \amalg A|L \Rightarrow E(Y_a) = g(a) = \sum_l E(Y|A = a, L = l) f_L(l) \Rightarrow$**crude association is not causal**. To get causal effect ψ, we must compute the g-formula $g(a)$, $a = 0, 1$; This naturally adjust for the confounders L
  - No unmeasured confounding essentially means that we have collected all common causes of A and Y in L. This is a strong assumption which requires subject matter expertise.

- Inference on the causal effect ψ:
  - If randomization holds standard crude analysis still applies, such as two sample t-tests ...etc .
  - under no unmeasured confounding, we must obtain point and interval estimates of ψ.
  - To get point estimates nonparametrically, we use the pluggin principle: $\widehat{g}(a) = \sum_l \widehat{b}(a, l)\widehat{f}_L(l)$, where $\widehat{b}(a, l)$ and $\widehat{f}_L(l)$ are nonparametric estimates of **nuisance parameters** not of primary interest but required to define our causal contrast.
  - Nonparametric method here is appealing because we do not want to specify models for $b(a, l)$ and $f_L(l)$ if not necessary as neither is of primary interest.

- We also proved that:

$$\sqrt{n}\left(\widehat{\psi}-\psi\right)=\sqrt{n}\left(\widehat{g}\left(1\right)-\widehat{g}\left(0\right)-\psi\right)\approx N\left(0,E\left[\left(IF\left(0\right)-IF\left(1\right)\right)\right]^{2}\right)$$

where $IF_{i}\left(a\right)=\left\{\frac{I(A_{i}=a)}{\widehat{f}_{A|L}(a|L_{i})}\left(Y_{i}-b(A_{i},L_{i})\right)+b(a,L_{i})-g\left(a\right)\right\}$

- which we used to construct a Wald-type 95%CI

$$\widehat{\psi}\pm n^{-1/2}1.96\sqrt{n^{-1}\sum_{i}\left(\widehat{IF}_{i}\left(0\right)-\widehat{IF}_{i}\left(1\right)\right)^{2}}$$

where $\widehat{IF}_{i}\left(a\right)=\left\{\frac{I(A_{i}=a)}{\widehat{f}_{A|L}(a|L_{i})}\left(Y_{i}-\widehat{b}(A_{i},L_{i})\right)+\widehat{b}(a,L_{i})-\widehat{g}\left(a\right)\right\}$.

- Finally we found that :$\widehat{g}(a)=\sum_{l}\widehat{b}(a,l)\widehat{f}_{L}\left(l\right)=\frac{\sum_{s=1}^{n}I(A_{s}=a)Y_{s}\widehat{f}_{A|L}^{-1}(A_{s}|L_{s})}{\sum_{s=1}^{n}I(A_{s}=a)\widehat{f}_{A|L}^{-1}(A_{s}|L_{s})}$ has an exact dual representation as an inverse-probability of treatment weighted (iptw) estimator.

G-computation: Efficiency considerations

The observed data likelihood is given by

$$\begin{aligned}
&\prod_{i}f_{Y,L,A}\left(Y_{i},L_{i},A_{i}\right)\\
=&\prod_{i}\int f_{Y_{0},Y_{1},L,A}\left(Y_{0i},Y_{1i},L_{i},A_{i}\right)dY_{1-A_{i}}\\
=&\prod_{i}\int f_{Y_{0},Y_{1},L}\left(Y_{0i},Y_{1i},L_{i}\right)f_{A|Y_{0},Y_{1},L}\left(A_{i}|Y_{0i},Y_{1i},L_{i}\right)dY_{1-A_{i}}\\
&\overset{NUCA}{=}\prod_{i}f_{A|Y_{0},Y_{1},L}\left(A_{i}|L_{i}\right)\int f_{Y_{0},Y_{1},L}\left(Y_{0i},Y_{1i},L_{i}\right)dY_{1-A_{i}}\\
=&\left\{\prod_{i}f_{A|Y_{0},Y_{1},L}\left(A_{i}|L_{i}\right)\right\}\times\prod_{i}f_{Y_{0}|L}\left(Y_{0i}|L_{i}\right)^{(1-A_{i})}f_{Y_{1}|L}\left(Y_{1i}|L_{i}\right)^{A_{i}}f_{L}\left(L_{i}\right)
\end{aligned}$$

Thus the observed data likelihood factorizes into a part that depends on the treatment process (first factor) and a part that depends on the full data process (second factor). For simplicity, assume that Y is binary, so that the likelihood becomes

$$\begin{aligned}
&\prod_{i}f_{A|Y_{0},Y_{1},L}\left(A_{i}|L_{i}\right)\\
&\times\prod_{i}\left\{\begin{array}{c}\Pr\left(Y_{0i}=1|L_{i}\right)^{Y_{0i}(1-A_{i})}\Pr\left(Y_{0i}=0|L_{i}\right)^{(1-Y_{0i})(1-A_{i})}\\ \times\Pr\left(Y_{1i}=1|L_{i}\right)^{Y_{1i}A_{i}}\Pr\left(Y_{1i}=0|L_{i}\right)^{(1-Y_{1i})A_{i}}f_{L}\left(L_{i}\right)\end{array}\right\}\\
=&\prod_{i}f_{A|Y_{0},Y_{1},L}\left(A_{i}|L_{i}\right)\times\\
&f_{L}\left(L_{i}\right)\times\prod_{i}\left\{\begin{array}{c}b\left(0,L_{i}\right)^{Y_{i}(1-A_{i})}\left(1-b\left(0,L_{i}\right)\right)^{(1-Y_{i})(1-A_{i})}\\ \times b\left(1,L_{i}\right)^{Y_{i}A_{i}}\left(1-b\left(1,L_{i}\right)\right)^{(1-Y_{i})A_{i}}\end{array}\right\}
\end{aligned}$$

- The factorization is similar to that encountered in missing data problems, under 'missingness at random'.
- The treatment process is said to be ancillary to inferences on $\{g(0),g(1)\}$ and thus on $\psi=g(1)-g(0)$
- It is straightforward to show that the MLE of $g(a)$ is $\widehat{g}(a)=\sum_{l}\widehat{b}(a,l)\widehat{f}_{L}\left(l\right),a$ $0,1$.Thus the pluggin estimator attains the Cramer-Rao efficiency bound.
- Moreover, because the likelihood factorizes, the Cramer-Rao efficiency bound is the same whether or not the treatment process is known. Though the bound itself (the asymptotic variance of $\widehat{g}(1)-\widehat{g}(0)$) depends on the treatment process!!!

- We now show that the nonparametric G-computation estimator $\widehat{g}(1) - \widehat{g}(0)$ has a dual representation as an inverse-probability of treatment weighted (iptw) estimator:
- $\widehat{g}(a) = \sum_l \widehat{b}(a,l)\widehat{f}_L(l) = n^{-1}\sum_l \sum_{i=1}^n \widehat{b}(a,l) I(L_i = l)$
  $= n^{-1}\sum_l \sum_{i=1}^n I(L_i = l)\left\{\frac{\sum_{s=1}^n I(A_s=a, L_s=l) Y_s}{\sum_{j=1}^n I(A_j=a, L_j=l)}\right\}$
  $= n^{-1}\sum_l \sum_{s=1}^n I(A_s = a, L_s = l) Y_s \left\{\frac{\sum_{i=1}^n I(L_i=l)}{\sum_{j=1}^n I(A_j=a, L_j=l)}\right\}$
  $= n^{-1}\sum_{s=1}^n I(A_s = a) Y_s \widehat{f}^{-1}_{A|L}(A_s|L_s) = \frac{\sum_{s=1}^n I(A_s=a) Y_s \widehat{f}^{-1}_{A|L}(A_s|L_s)}{\sum_{s=1}^n I(A_s=a)\widehat{f}^{-1}_{A|L}(A_s|L_s)}$

- $\widehat{f}_{A|L}(a|L=l) = \frac{\sum_i I(A_i=a, L_i=l)}{\sum_i I(L_i=l)}$ is the nonparametric estimator of the treatment conditional probability mass function $f_{A|L}(a|L=l)$.
- So that the nonparametric G-formula estimator has a (exact) dual representation as an inverse-probability-of-treatment weighted (iptw) estimator, and this also shows that the latter is asymptotically efficient.

- One can easily show (try it) that the iptw estimator $\widehat{g}(a)$ is the solution to the estimating equation: $\sum_{i=1}^n \frac{I(A_i=a)}{\widehat{f}_{A|L}(a|L_i=l)}(Y_i - \widehat{g}(a)) = 0$.
- In contrast to the estimating equation for the conditional mean $\mu(a) = E(Y|A=a)$ given by: $\sum_{i=1}^n I(A_i = a)(Y_i - \widehat{\mu}(a)) = 0$.
- We again see that if $f_{A|L}(a|L_i = l) = f_A(a)$, then $E(Y|A=a) = E(Y_a)$ and association is causation.
- If $f_{A|L}(a|L_i = l) \neq f_A(a)$ and L is a risk factor of Y, then to obtain a counterfactual mean $E(Y_a)$, the recipe is to weight the estimating function for the conditional mean by $\widehat{f}^{-1}_{A|L}(a|L_i = l)$ to adjust for confounding by L.

- Why does weigthing adjust for confounding?
- Essentially, creates a 'pseudo-population' in which A is no longer associated with Y; so that the crude mean in this population has a counterfactual interpretation.
- For example, suppose L is binary and we observe :

N	A	L	$E(Y\|A=a, L=l)$
4000	1	0	24
3000	1	1	36
8000	0	0	10
9000	0	1	22

- So that $f_{A|L}(A=1|L=1) = 1/4, f_{A|L}(A=1|L=0) = 1/3$ and thus L predicts A.
- Moreover $E(Y|A=1, L=l) = 24 + 12l$ so that L predict Y given A.

- The crude mean $E(Y|A=1) = \sum_l E(Y|A=1, L=l) f(L=l|A=1) = 24 \times 4/7 + 36 \times 3/7 = 204/7$
- Whereas $E(Y_1) = \sum_l E(Y|A=1, L=l) f(L=l) = 24 \times 1/2 + 36 \times 1/2 = 210/7$
- Create a pseudo population by reweigthing the numbers in each row by $f^{-1}_{A|L}$

N	$f(A\|L)$	$Pseudo-N$	A	L	$E(Y\|A=a, L=l)$
4000	1/3	$4000 \times 3 = 12,000$	1	0	24
3000	1/4	$3000 \times 4 = 12,000$	1	1	36
8000	2/3	$8000 \times 3/2 = 12,000$	0	0	10
9000	3/4	$9000 \times 4/3 = 12,000$	0	1	22

- So that in the Pseudo population, the crude analysis gives:$E^*(Y|A=1) = 24 \times 1/2 + 36 \times 1/2 = 210/7 = E(Y_1)$
- Do a similar calculation for $E(Y_0)$

IPTW estimator in a randomized trial: an efficiency paradox

- Lets revisit the familiar setting where A is randomized. Denote by L a pretreatment categorical covariate that is a strong predictor of the outcome Y.
- By randomization, we have :$\{Y_0, Y_1, L\} \amalg A$
- And as before, the crude estimator

$$\widetilde{\psi} = \frac{\sum_{s=1}^n I(A_s=1) Y_s}{\sum_{s=1}^n I(A_s=1)} - \frac{\sum_{s=1}^n I(A_s=0) Y_s}{\sum_{s=1}^n I(A_s=0)}$$

is a valid estimator of the average causal effect ψ_0.

- Note that this estimator may also be written as an iptw estimator with known weights $f^{-1}_{A|L}(A_s|L_s) = (1/2)^{-1}$

$$\widetilde{\psi} = \frac{\sum_{s=1}^n I(A_s=1) Y_s (1/2)^{-1}}{\sum_{s=1}^n I(A_s=1)(1/2)^{-1}} - \frac{\sum_{s=1}^n I(A_s=0) Y_s (1/2)^{-1}}{\sum_{s=1}^n I(A_s=0)(1/2)^{-1}}$$

- It is easy to show that the asymptotic variance of this estimator is given by

$$var\left(\sqrt{n}\widetilde{\psi}\right) = E\left((Y-g(1))^2 | A=1\right) + E\left((Y-g(0))^2 | A=0\right)$$

- Compare this estimator to the one used by a statistician who decides to ignore randomization but rather assumes that $\{Y_0, Y_1\} \perp A|L$, so that she may use G-computation

$$\begin{aligned}\widehat{\psi} &= \hat{g}(1) - \hat{g}(0) = \sum_l \left(\hat{b}(1,l) - \hat{b}(0,l)\right) \hat{f}_L(l) \\ &= \frac{\sum_{s=1}^n I(A_s=1) Y_s \hat{f}^{-1}_{A|L}(A_s|L_s)}{\sum_{s=1}^n I(A_s=1) \hat{f}^{-1}_{A|L}(A_s|L_s)} - \frac{\sum_{s=1}^n I(A_s=0) Y_s \hat{f}^{-1}_{A|L}(A_s|L_s)}{\sum_{s=1}^n I(A_s=0) \hat{f}^{-1}_{A|L}(A_s|L_s)}\end{aligned}$$

with asymptotic variance $var\left(\widehat{\psi}\right) \neq var\left(\widetilde{\psi}\right)$

- In fact, it can be shown (check this) that $\widehat{\psi}$ is still asymptotically efficient and therefore $var\left(\widehat{\psi}\right) \leq var\left(\widetilde{\psi}\right)$

- But this is paradoxal as we find that the estimator $\widetilde{\psi}$ that uses available information on the treatment mechanism, is less efficient that $\widehat{\psi}$ which ignores this information
- However, as argued with the likelihood on slide (97), the treatment process is ancillary, that is the efficiency bound for estimating ψ is the same whether or not one knows the treatment mechanism.
- The paradox is resolved by realizing that $\widetilde{\psi}$ uses the available information, but in a very inefficient way as it ignores the fact that L is a strong correlate of Y.

- In contrast, the G-formula and the iptw with nonparametrically estimated weights correctly incorporates this information.
- Take home messages:
  - More information is generally a good thing, but not helpful if used inefficiently.
  - Standard methods for analyzing clinical trials such as simple two sample z-test, may be highly inefficient as they ignore potentially highly informative baseline covariates.