Part 5
Semiparametric Inference

Semiparametric Inference

The *semiparametric theory* of estimation that can be used to justify several of the previous causal inference methods.

We consider models that include

- parametric components (with a finite dimensional parameter of interest)
- nonparametric components (with an infinite dimensional parameter that is usually a nuisance parameter)

Example: Ordinary Least Squares

Ordinary least squares is a semiparametric approach:

$$\widehat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} (y_i - \mu(\mathbf{x}_i; \theta))^2$$

- parametric mean model $\mu(\mathbf{x}; \theta)$, $\theta \in \mathbb{R}^d$ say
- nonparametric distributional assumption for the conditional distribution of Y given $X = x$.

Example: Ordinary Least Squares

If

$$\varepsilon_i = Y_i - \mu(X_i; \theta)$$

we assume only that

$$\int f_{\varepsilon|X}(t|x)\, dt = 1 \qquad \int t\, f_{\varepsilon|X}(t|x)\, dt = 0$$

and

$$\int t^2\, f_{\varepsilon|X}(t|x)\, dt < \infty$$

but make no other assumptions.

Linear models: a recap

Suppose we have a linear model

$$Y_i = \mathbf{x}_{i1}\beta + \mathbf{x}_{i2}\psi + \varepsilon_i$$

where

- $\mathbf{x}_{i1}$ is $1 \times r$
- β is $r \times 1$
- $\mathbf{x}_{i2}$ is $1 \times q$
- ψ is $q \times 1$

Let θ be the concatenation of β and ψ, so that θ is $p \times 1$ with $p = q + r$.

Linear models: a recap

In vector form, for a random sample of size n, we have

$$\mathbf{Y} = \mathbf{X}_1\beta + \mathbf{X}_2\psi + \varepsilon$$

We require *linear independence* of the columns of the combined matrix

$$[\mathbf{X}_1\ \mathbf{X}_2]\,.$$

Linear models: a recap

Note that

- $\mathbf{X}_1\beta$ is an arbitrary point in the linear subspace of $\mathbb{R}^n$ spanned by the columns of $\mathbf{X}_1$ – denote this linear subspace Λ.
- $\mathbf{X}_2\psi$ is an arbitrary point in the linear subspace of $\mathbb{R}^n$ spanned by the columns of $\mathbf{X}_2$.

Linear models: a recap

For OLS estimation we need to solve

$$\begin{pmatrix} \mathbf{X}_1^\top \\ \mathbf{X}_2^\top \end{pmatrix} (\mathbf{y} - \mathbf{X}_1\beta - \mathbf{X}_2\psi) = \mathbf{0}_p$$

We can write the left hand side

$$\begin{pmatrix} \mathbf{X}_1^\top \\ \mathbf{X}_2^\top \end{pmatrix} \varepsilon$$

which is an element in a p-dimensional space.

Linear models: a recap

Suppose that ψ is the *parameter of interest*. For *any* ψ, we may write the model

$$(\mathbf{Y} - \mathbf{X}_2\psi) = \mathbf{X}_1\beta + \varepsilon$$

and hence for *nuisance parameter* β deduce that

$$\widetilde{\beta} = (\mathbf{X}_1^\top \mathbf{X}_1)^{-1}\mathbf{X}_1^\top(\mathbf{y} - \mathbf{X}_2\psi).$$

This follows from the usual OLS result; note that

$$\widetilde{\beta} \equiv \widetilde{\beta}(\psi)$$

is a function of ψ.

Linear models: a recap

We then have that

$$\mathbf{X}_1\widetilde{\beta} = \mathbf{X}_1(\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\mathbf{X}_1^\top(\mathbf{y} - \mathbf{X}_2\psi) = \mathbf{H}_1(\mathbf{y} - \mathbf{X}_2\psi)$$

say, with

$$\mathbf{H}_1 = \mathbf{X}_1(\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\mathbf{X}_1^\top \qquad (n \times n).$$

Note that $\mathbf{H}_1$ is *symmetric*, and that

$$\mathbf{H}_1\mathbf{H}_1 = \mathbf{H}_1^\top\mathbf{H}_1 = \mathbf{H}_1.$$

Linear models: a recap

Note that

- the $n \times 1$ vector $\mathbf{X}_1\widetilde{\beta}$ is a *specific* point in the space Λ.
- for any ψ, $\mathbf{X}_1\widetilde{\beta}$ is the *closest* point in Λ to $(\mathbf{y} - \mathbf{X}_2\psi)$;
- the matrix $\mathbf{H}_1$ *projects* $(\mathbf{y} - \mathbf{X}_2\psi)$ onto Λ to compute $\mathbf{X}_1\widetilde{\beta}$.
- the '*residual*' vector after finding $\mathbf{X}_1\widetilde{\beta}$ is

$$\begin{aligned}(\mathbf{y} - \mathbf{X}_2\psi) - \mathbf{X}_1\widetilde{\beta} &= (\mathbf{y} - \mathbf{X}_2\psi) - \mathbf{H}_1(\mathbf{y} - \mathbf{X}_2\psi) \\ &= (\mathbf{I}_n - \mathbf{H}_1)(\mathbf{y} - \mathbf{X}_2\psi)\end{aligned}$$

Linear models: a recap

- Note that for any values of β and ψ

$$\{(\mathbf{I}_n - \mathbf{H}_1)(\mathbf{y} - \mathbf{X}_2\psi)\}^\top \{\mathbf{X}_1\beta\} = (\mathbf{y} - \mathbf{X}_2\psi)^\top (\mathbf{I}_n - \mathbf{H}_1)^\top \mathbf{X}_1\beta$$
$$= 0$$

where $\mathbf{X}_1\beta$ is an arbitrary point in Λ, as

$$\begin{aligned}(\mathbf{I}_n - \mathbf{H}_1)^\top \mathbf{X}_1 &= (\mathbf{I}_n - \mathbf{H}_1^\top)\mathbf{X}_1 \\ &= \mathbf{X}_1 - \mathbf{H}_1\mathbf{X}_1 && \text{as } \mathbf{H}_1^\top = \mathbf{H}_1 \\ &= \mathbf{X}_1 - \{\mathbf{X}_1(\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\mathbf{X}_1^\top\}\mathbf{X}_1 \\ &= \mathbf{X}_1 - \mathbf{X}_1 = \mathbf{0}.\end{aligned}$$

Linear models: a recap

Therefore the residual vector

$$(\mathbf{I}_n - \mathbf{H}_1)(\mathbf{y} - \mathbf{X}_2\psi)$$

is *orthogonal* to Λ.

Linear models: a recap

Returning to the estimation of ψ, we now have the reduced form model

$$(\mathbf{Y} - \mathbf{X}_2\psi) = \mathbf{X}_1\widetilde{\beta} + \varepsilon$$

or equivalently

$$(\mathbf{Y} - \mathbf{X}_2\psi) = \mathbf{H}_1(\mathbf{Y} - \mathbf{X}_2\psi) + \varepsilon$$

that is

$$(\mathbf{I}_n - \mathbf{H}_1)\mathbf{Y} = (\mathbf{I}_n - \mathbf{H}_1)\mathbf{X}_2\psi + \varepsilon.$$

which is a linear form in ψ that can be solved using OLS as usual.

Linear models: a recap

Note that the estimating equation is

$$\mathbf{X}_2^\top(\mathbf{I}_n - \mathbf{H}_1)(\mathbf{y} - \mathbf{X}_2\psi) = \mathbf{0}$$

again implying a need for orthogonality between the columns of $\mathbf{X}_2$ and the residual quantity $(\mathbf{I}_n - \mathbf{H}_1)(\mathbf{y} - \mathbf{X}_2\psi)$.

Note also that for any β

$$(\mathbf{I}_n - \mathbf{H}_1)(\mathbf{y} - \mathbf{X}_2\psi) = (\mathbf{I}_n - \mathbf{H}_1)(\mathbf{y} - \mathbf{X}_1\beta - \mathbf{X}_2\psi).$$

as

$$(\mathbf{I}_n - \mathbf{H}_1)\mathbf{X}_1 = (\mathbf{I}_n - \mathbf{H}_1)^\top\mathbf{X}_1 = \mathbf{0}.$$

Hilbert spaces

Hilbert space: A (real) Hilbert space $\mathcal{H}$ is a vector space that

- has an associated *inner product* $\langle ., . \rangle$, such that for elements $h_1, h_2, h_3 \in \mathcal{H}$

$$\langle h_1, h_2 \rangle \in \mathbb{R}$$

can be computed and has certain properties.

- $\langle h_1, h_2 \rangle = \langle h_2, h_1 \rangle$
- $\langle a h_1 + b h_2, h_3 \rangle = a \langle h_1, h_3 \rangle + b \langle h_2, h_3 \rangle$
- $\langle h, h \rangle \geqslant 0$

Hilbert spaces

- has a *metric* defined in terms of the inner product that defines the concept of *norm*, *distance* and hence *convergence* in the space.
  - *norm*:
    $$\|h\| = \sqrt{\langle h, h \rangle}$$
  - *distance*:
    $$d(h_1, h_2) = \|h_1 - h_2\| = \sqrt{\langle h_1 - h_2, h_1 - h_2 \rangle}$$

Hilbert spaces

We may deduce that for $h_1, h_2, h_3 \in \mathcal{H}$, for example

$$d(h_1, h_3) \leqslant d(h_1, h_2) + d(h_2, h_3)$$

$$\|h_1 + h_2\| \leqslant \|h_1\| + \|h_2\|$$

$$|\langle h_1, h_2 \rangle| \leqslant \|h_1\| \|h_2\|$$

- is *complete*: essentially, Cauchy sequences of elements of $\mathcal{H}$ converge to a limit point that is also in $\mathcal{H}$.

Hilbert spaces

We are most used to dealing with Hilbert spaces when the elements h are d-dimensional real vectors, with

$$\langle h_1, h_2 \rangle = h_1^\top h_2 = \sum_{j=1}^{d} h_{1j} h_{2j}$$

for which the usual concepts of Euclidean distance etc apply.

Hilbert spaces

Consider the Hilbert space $\mathcal{H}_q$ of q-dimensional functions of random variable Z, with

$$\mathbb{E}[h(Z)] = \mathbf{0}_q \qquad \mathbb{E}[\{h(Z)\}^\top h(Z)] < \infty$$

for $h \in \mathcal{H}_q$, with the *covariance inner product*

$$\langle h_1, h_2 \rangle = \mathbb{E}[\{h_1(Z)\}^\top h_2(Z)] \qquad h_1, h_2 \in \mathcal{H}_q.$$

so that $\|h\|^2 = \langle h, h \rangle$. We say that h_1 and h_2 are *orthogonal* if

$$\langle h_1, h_2 \rangle \equiv \mathbb{E}[h_1^\top h_2] = 0.$$

In this formulation, all expectations are taken with respect to the distribution of random variable Z.

Subspaces and Projections

Consider arbitrary Hilbert space $\mathcal{H}$. For k linearly independent functions $h_1, \ldots, h_k \in \mathcal{H}$ consider the *linear subspace*, $\mathcal{U}$ of $\mathcal{H}$ defined by

$$\mathcal{U} \equiv \left\{ u : u = \sum_{j=1}^{k} a_j h_j, (a_1, \ldots, a_k) \in \mathbb{R}^k \right\}.$$

$\mathcal{U}$ is the subspace *spanned* by $h_1, \ldots, h_k$. Points in the subspace are identified uniquely by the coefficients

$$(a_1, \ldots, a_k).$$

Subspaces and Projections

For each $h \in \mathcal{H}$, there is a *unique value* $u_0 \in \mathcal{U}$ such that

$$\|h - u_0\| < \|h - u\| \qquad \forall u \in \mathcal{U}, u \neq u_0.$$

that is, u_0 is the nearest point in $\mathcal{U}$ to h.

- u_0 is the *projection* of h onto $\mathcal{U}$, sometimes denoted

$$\Pi(h|\mathcal{U});$$

- the element $h - u_0 \in \mathcal{H}$ is *orthogonal* to $\mathcal{U}$, that is for all $u \in \mathcal{U}$

$$\mathbb{E}[(h - u_0)^\top u] = 0.$$

- $\|h\|^2 = \|u_0\|^2 + \|h - u_0\|^2$.

Subspaces and Projections

Suppose that $v \equiv v(Z)$ is an r-dimensional function of Z with

$$\mathbb{E}[v(Z)] = \mathbf{0}_r \qquad \mathbb{E}[\{v(Z)\}^\top v(Z)] < \infty.$$

Let

$$\mathcal{U}_v \equiv \{\mathbf{B}v : \mathbf{B} \text{ an arbitrary } q \times r \text{ real matrix}\}$$

define a q-dimensional linear subspace of $\mathcal{H}_q$ derived from v.

That is, all elements of $\mathcal{U}_v$ can be expressed

$$\mathbf{B}v$$

for some (non-stochastic) matrix $\mathbf{B}$.

Subspaces and Projections

For $h \in \mathcal{H}_q$, the projection of h onto $\mathcal{U}_v$ must take the form $\mathbf{B}_0 v$, with $\mathbf{B}_0$ satisfying

$$\langle h - \mathbf{B}_0 v, \mathbf{B} v \rangle = 0 \qquad \forall \mathbf{B} \in \mathbb{R}^{q \times r}$$

that is,

$$\mathbb{E}[(h - \mathbf{B}_0 v)^\top \mathbf{B} v] = 0$$

Now, this equation can be re-written elementwise as

$$\sum_{i=1}^{q}\sum_{j=1}^{r} B_{ij}\mathbb{E}\left[(h-\mathbf{B}_0 v)_i v_j\right] = 0$$

where subscripting identifies elements of the vector/matrix.

As this must hold for all $\mathbf{B}$, it needs to hold for such matrices that have 1 in position (i,j) and zero elsewhere, which implies

$$\mathbb{E}\left[(h-\mathbf{B}_0 v)_i v_j\right] = 0 \qquad \forall (i,j)$$

and hence that

$$\mathbb{E}\left[(h-\mathbf{B}_0 v)v^\top\right] = \mathbf{0}.$$

Subspaces and Projections

Thus we may deduce that

$$\mathbf{B}_0 = \mathbb{E}[hv^\top]\left\{\mathbb{E}[vv^\top]\right\}^{-1} \qquad (\blacksquare)$$

so that

$$\mathbf{B}_0 v = \mathbb{E}[hv^\top]\left\{\mathbb{E}[vv^\top]\right\}^{-1} v. \qquad (\blacklozenge)$$

- $\mathbb{E}[hv^\top] \equiv \mathrm{Cov}_Z[h(Z), v(Z)]$ is $q \times r$;
- $\mathbb{E}[vv^\top] \equiv \mathrm{Var}_Z[v(Z)]$ is $r \times r$;
- it is assumed that $\mathbb{E}[vv^\top]$ is non-singular.

Parametric Inference

Suppose $Z_1, \ldots, Z_n$ are i.i.d. with pdf $f_Z(z;\theta)$ with $\theta \in \mathbb{R}^p$, and

$$\theta = \begin{bmatrix} \psi \\ \\ \beta \end{bmatrix} \qquad \begin{matrix} q \times 1 \\ \\ r \times 1 \end{matrix}$$

with true value θ_0 comprised of ψ_0 and β_0.

- ψ – parameter of interest;
- β – nuisance parameter.

The data generating model is thus $f_Z(z;\theta_0)$.

Step 1: Asymptotically Linear Estimators

An estimator $\widehat{\psi}_n$ of parameter ψ_0 is *asymptotically linear* if there exists a q-dimensional function, $\varphi(Z)$, with

$$\mathbb{E}[\varphi(Z)] = \mathbf{0}_q \qquad \mathbb{E}[\varphi(Z)\{\varphi(Z)\}^\top] < \infty, \text{ nonsingular}$$

such that

$$\sqrt{n}(\widehat{\psi}_n - \psi_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi(Z_i) + \mathrm{o}_p(1).$$

Step 1: Asymptotically Linear Estimators

Then as

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi(Z_i) \xrightarrow{d} \text{Normal}\left(\mathbf{0}_q, \mathbb{E}[\varphi(Z)\{\varphi(Z)\}^\top]\right)$$

$\sqrt{n}(\widehat{\psi}_n - \psi_0)$ also has this asymptotic distribution.

If such a $\varphi(.)$ exists, it is unique; it is termed the *influence function* for the estimator.

Step 1: Asymptotically Linear Estimators

Example: Sample mean

If $Z_1, \ldots, Z_n$ are a random sample from a population with finite mean θ and variance σ^2, then the estimator

$$\widehat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} Z_i$$

yields the representation

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Z_i - \theta_0)$$

so we can deduce the influence function

$$\varphi(Z) \equiv \varphi(Z, \theta_0) = Z - \theta_0.$$

Step 1: Asymptotically Linear Estimators

Example: Sample mean

By the Central Limit Theorem, we have that as $n \longrightarrow \infty$

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} Normal(0, \sigma^2)$$

where

$$\sigma^2 \equiv \mathbb{E}[(Z - \theta_0)^2] \equiv \mathbb{E}\left[\{\varphi(Z)\}^2\right].$$

Step 1: Asymptotically Linear Estimators

Recall that in *likelihood estimation*, we have the log-likelihood

$$\ell_n(\theta) = \sum_{i=1}^{n} \log f_Z(z_i;\theta) = \sum_{i=1}^{n} \ell(z_i,\theta).$$

say. Under regularity conditions, by the *mean-value theorem*,

$$\dot{\ell}_n(\theta) = \dot{\ell}_n(\theta_0) + \ddot{\ell}_n(\theta')(\theta - \theta_0)$$

where $\|\theta' - \theta_0\| < \|\theta - \theta_0\|$, and where

$$\dot{\ell}_n(\theta) = \frac{\partial \ell_n(\theta)}{\partial \theta} \qquad \ddot{\ell}_n(\theta) = \frac{\partial^2 \ell_n(\theta)}{\partial \theta \partial \theta^\top}.$$

are $(p \times 1)$ and $(p \times p)$ respectively.

Step 1: Asymptotically Linear Estimators

Evaluating at $\theta = \widehat{\theta}_n$, and noting that $\dot{\ell}_n(\widehat{\theta}_n) = \mathbf{0}_p$, we have on rearrangement and multiplying through by $1/\sqrt{n}$ that

$$\left\{ -\frac{1}{n}\ddot{\ell}_n(\theta') \right\} \sqrt{n}(\widehat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}}\dot{\ell}_n(\theta_0)$$

where

$$\|\theta' - \theta_0\| < \|\widehat{\theta}_n - \theta_0\|$$

Step 1: Asymptotically Linear Estimators

As $n \longrightarrow \infty$, we have for the random variable version

$$\left\{ \frac{1}{n} \ddot{\ell}_n(\theta') \right\} \xrightarrow{p} \mathbb{E}\left[\left. \frac{\partial^2 \log f_Z(Z;\theta)}{\partial\theta\partial\theta^\top} \right|_{\theta=\theta_0} \right] = \mathbb{E}\left[\ddot{\ell}(Z,\theta_0) \right]$$

say, as

$$\widehat{\theta}_n \xrightarrow{p} \theta_0.$$

Step 1: Asymptotically Linear Estimators

Therefore

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = \left\{-\mathbb{E}\left[\ddot{\ell}(Z, \theta_0)\right]\right\}^{-1} \frac{1}{\sqrt{n}} \dot{\ell}_n(\theta_0) + \mathrm{o}_p(1)$$

which yields the influence function

$$\begin{aligned}\varphi(Z) \equiv \varphi(Z, \theta_0) &= \left\{-\mathbb{E}\left[\ddot{\ell}(Z, \theta_0)\right]\right\}^{-1} \dot{\ell}(Z, \theta_0) \\ &= \mathcal{J}^{-1}(\theta_0)\dot{\ell}(Z, \theta_0)\end{aligned}$$

say.

Step 1: Asymptotically Linear Estimators

To obtain the asymptotic variance, we compute

$$\text{Var}[\varphi(Z)] = \mathcal{J}^{-1}(\theta_0)\text{Var}[\dot{\ell}(Z,\theta_0)]\mathcal{J}^{-1}(\theta_0)$$

where, under standard likelihood theory,

$$\mathcal{I}(\theta_0) = \text{Var}[\dot{\ell}(Z,\theta_0)] = \mathbb{E}\left[\dot{\ell}(Z,\theta_0)\{\dot{\ell}(Z,\theta_0)\}^\top\right] \equiv -\mathbb{E}\left[\ddot{\ell}(Z,\theta_0)\right]$$

that is $\mathcal{I}(\theta_0) = \mathcal{J}(\theta_0)$ so that

$$\text{Var}[\varphi(Z)] = \left\{\mathbb{E}\left[\dot{\ell}(Z,\theta_0)\{\dot{\ell}(Z,\theta_0)\}^\top\right]\right\}^{-1} = \mathcal{I}^{-1}(\theta_0)$$

which is the asymptotic variance of $\widehat{\theta}_n$.

Step 2: Regular Estimators

Consider the *local data generating process* (LDGP)

$$Z_{1n}, \ldots, Z_{nn} \sim f_Z(z; \theta_n) \qquad \text{i.i.d.}$$

with

$$\sqrt{n}(\theta_n - \theta^*) \longrightarrow \text{constant}$$

as $n \longrightarrow \infty$.

Estimator $\widehat{\psi}_n$ of ψ_n is *regular* if its limiting distribution does not depend on θ_n.

Step 2: Regular Estimators

That is, in the Normal case, $\widehat{\psi}_n$ is *regular* if

$$\sqrt{n}(\widehat{\psi}_n - \psi^*) \xrightarrow{d} Normal(0, \Sigma^*)$$

under $f_Z(z; \theta^*)$ implies that

$$\sqrt{n}(\widehat{\psi}_n - \psi_n) \xrightarrow{d} Normal(0, \Sigma^*)$$

under $f_Z(z; \theta_n)$.

Step 3: Score Functions

The *score function* $S_\theta(z, \theta_0)$ is defined by

$$S_\theta(z, \theta_0) = \frac{\partial}{\partial \theta} \{\log f_Z(z; \theta)\}_{\theta=\theta_0} = \begin{bmatrix} S_\psi(z, \theta_0) \\ S_\beta(z, \theta_0) \end{bmatrix} \begin{matrix} q \times 1 \\ r \times 1 \end{matrix}$$

with the subscript denoting the variable with respect to which the derivative is being taken.

Step 4: Key Theorem for RAL Estimators

Consider a q-dimensional parameter of interest $\Psi(\theta)$, and let

$$\Gamma(\theta) = \frac{\partial \Psi(\theta)}{\partial \theta^\top} \qquad q \times p.$$

Suppose that $\widehat{\Psi}_n$ is *regular and asymptotically linear* (RAL) with influence function $\varphi(.)$ such that $\mathbb{E}[\{\varphi(Z)\}^\top \varphi(Z)] < \infty$.

Then

$$\mathbb{E}[\varphi(Z)\{S_\theta(Z, \theta_0)\}^\top] = \Gamma(\theta_0).$$

Step 4: Key Theorem for RAL Estimators

Special case: If

$$\Psi(\theta) \equiv \psi$$

then

$$\Gamma(\theta_0) = \begin{bmatrix} \mathbf{I}_{q\times q} & \mathbf{0}_{q\times r} \\ \mathbf{0}_{r\times q} & \mathbf{0}_{r\times r} \end{bmatrix}$$

and so

(i) $\mathbb{E}[\varphi(Z)\{S_\psi(Z,\theta_0)\}^\top] = \mathbf{I}_{q\times q}$;

(ii) $\mathbb{E}[\varphi(Z)\{S_\beta(Z,\theta_0)\}^\top] = \mathbf{0}_{q\times r}$.

That is, $\varphi(Z)$ is *orthogonal* to $S_\beta(Z,\theta_0)$.

Step 5: m-estimation

In m-estimation, we replace the *score equation* for θ

$$\sum_{i=1}^{n} S_\theta(Z_i, \theta) \equiv \sum_{i=1}^{n} \dot{\ell}(Z_i, \theta) = \mathbf{0}_p$$

by the more general form

$$\sum_{i=1}^{n} m(Z_i, \theta) = \mathbf{0}_p$$

for function $m(.,.)$

Step 5: m-estimation

We must have

$$\text{(i)} \quad \mathbb{E}[m(Z,\theta)] = \mathbf{0}_p$$

$$\text{(ii)} \quad \mathbb{E}[\{m(Z,\theta)\}^\top m(Z,\theta)] < \infty$$

$$\text{(iii)} \quad \mathbb{E}[m(Z,\theta)\{m(Z,\theta)\}^\top] \text{ nonsingular}$$

for all possible data generating θ.

Step 5: m-estimation

Using the same logic as in the likelihood case, we have

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi_m(Z_i) + \mathrm{o}_p(1)$$

for the influence function associated with m

$$\varphi_m(Z) = \{-\mathbb{E}\left[\dot{m}(Z, \theta_0)\right]\}^{-1} m(Z, \theta_0).$$

Step 5: m-estimation

We have by the CLT and elementary results for the Normal distribution that

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \longrightarrow Normal(\mathbf{0}_p, \mathcal{J}^{-1}\mathcal{I}\mathcal{J}^{-\top})$$

where

$$\mathcal{I} \equiv \mathcal{I}(\theta_0) = \mathbb{E}[m(Z, \theta_0)\{m(Z, \theta_0)\}^\top]$$

and

$$\mathcal{J} \equiv \mathcal{J}(\theta_0) = -\mathbb{E}[\dot{m}(Z, \theta_0)].$$

Step 5: m-estimation

These $p \times p$ matrices are typically estimated by

$$\widehat{I}_n = \frac{1}{n} \sum_{i=1}^{n} m(Z_i, \theta_0)\{m(Z_i, \theta_0)\}^\top \qquad \widehat{J}_n = -\frac{1}{n} \sum_{i=1}^{n} \dot{m}(Z_i, \theta_0)$$

with θ_0 replaced by estimator $\widehat{\theta}_n$.

Step 6: The Geometry of Influence Functions

Consider probability model $f_Z(z;\theta)$, $\theta^\top = (\psi^\top, \beta^\top)$.

- $\mathcal{H}_q$: Hilbert space of q-dimensional zero-mean functions with finite variance, with the covariance inner product.
- $\mathcal{T}$: linear subspace of $\mathcal{H}_q$ – the *tangent space*

$$\mathcal{T} \equiv \{\mathbf{B}S_\theta(z,\theta) : \mathbf{B} \text{ a } q \times p \text{ matrix}\}$$

  that is, the space spanned by the score S_θ.
- Λ: linear subspace of $\mathcal{H}_q$ – the *nuisance tangent space*

$$\Lambda \equiv \{\mathbf{B}S_\beta(z,\theta) : \mathbf{B} \text{ a } q \times r \text{ matrix}\}$$

  that is, the space spanned by the nuisance score S_β.

Step 6: The Geometry of Influence Functions

By the previous theorem, if $\widehat{\psi}_n$ is an RAL estimator of ψ with influence function $\varphi(Z) \in \mathcal{H}_q$, then we must have

$$\mathbb{E}\left[\varphi(Z)\left\{S_\beta(Z, \theta_0)\right\}^\top\right] = \mathbf{0}_{q \times r}$$

that is, $\varphi(Z)$ is *orthogonal to the nuisance tangent space*.

We write

$$\varphi(Z) \in \Lambda^\perp.$$

where

$$\Lambda^\perp \equiv \quad \text{“space orthogonal to } \Lambda\text{.”}$$

Step 6: The Geometry of Influence Functions

We have that

$$\mathcal{H}_q = \Lambda \oplus \Lambda^{\perp}$$

that is, any $h \in \mathcal{H}_q$ can be written

$$h = s_1 + s_2$$

for $s_1 \in \Lambda, s_2 \in \Lambda^{\perp}$, with s_1 orthogonal to s_2.

Step 6: The Geometry of Influence Functions

For any $h \in \mathcal{H}$, write s_0 for the projection of h onto Λ

$$\Pi(h|\Lambda) = s_0$$

that is, s_0 is the *unique* point in Λ that is at the "foot" of the perpendicular dropped from h onto Λ.

Then as $h = s_0 + (h - s_0)$, we can deduce that

$$\Pi(h|\Lambda^{\perp}) = h - s_0$$

by the projection theorem.

Step 6: The Geometry of Influence Functions

By the results on page 389, we know how to compute s_0 explicitly; we have that for any $h \in \mathcal{H}_q$

$$\Pi(h|\Lambda) = \mathbb{E}[hS_\beta^\top]\left\{\mathbb{E}[S_\beta S_\beta^\top]\right\}^{-1} S_\beta$$

so therefore all elements in the space $\Lambda^\perp$ can be written

$$h - \Pi(h|\Lambda) = h - \mathbb{E}[hS_\beta^\top]\left\{\mathbb{E}[S_\beta S_\beta^\top]\right\}^{-1} S_\beta.$$

Step 6: The Geometry of Influence Functions

Finally, as

$$\mathbf{B} S_\theta = \mathbf{B} \begin{bmatrix} S_\psi \\ S_\beta \end{bmatrix}$$

we can further decompose

$$\mathcal{T} = \mathcal{T}_\psi \oplus \Lambda$$

where

$$\mathcal{T}_\psi \equiv \{\mathbf{B}_1 S_\psi(Z, \theta_0) : \mathbf{B}_1 \text{ a } q \times q \text{ matrix}\}.$$

Step 7: Constructing an Estimating Function

Let influence function $\varphi(Z)$ satisfy the Theorem on page 403. Set

$$m(Z, \psi, \beta) = \varphi(Z) - \mathbb{E}[\varphi(Z)]$$

where the expectation is taken with respect to $f_Z(.; \psi, \beta)$. The function $m(Z, \psi, \beta)$ has mean zero and finite variance

Then $\widehat{\psi}_n$ satisfying

$$\sum_{i=1}^{n} m(Z_i, \widehat{\psi}_n, \widehat{\beta}_n(\widehat{\psi}_n)) = \mathbf{0}_p$$

is RAL with influence function $\varphi(Z)$.

Step 8: Efficient Influence Function

The *efficient influence function*, $\varphi^{\text{eff}}(Z)$, is the influence function with the smallest variance.

For an arbitrary influence function $\varphi(Z)$, using the projection theorem, it is evident that

$$\varphi^{\text{eff}}(Z) = \varphi(Z) - \Pi(\varphi(Z)|\mathcal{T}^{\perp}) \equiv \Pi(\varphi(Z)|\mathcal{T}).$$

Step 8: Efficient Influence Function

If we denote

$$\mathcal{I}(\theta_0) = \mathbb{E}\left[S_\theta(Z, \theta_0)\left\{S_\theta(Z, \theta_0)\right\}^\top\right]$$

we have explicitly that to estimate $\Psi(\theta_0)$

$$\varphi^{\text{eff}}(Z) = \Gamma(\theta_0)\left\{\mathcal{I}(\theta_0)\right\}^{-1} S_\theta(Z, \theta_0)$$

where, recall,

$$\Gamma(\theta) = \frac{\partial \Psi(\theta)}{\partial \theta^\top} \qquad q \times p.$$

Step 8: Efficient Influence Function

The *efficient score function* for ψ is obtained by projecting the score $S_\psi(Z, \theta_0)$ onto the nuisance tangent space Λ, and taking the residual.

By the result on page 389

$$\Pi(S_\psi|\Lambda) = \mathbb{E}[S_\psi S_\beta^\top] \left\{\mathbb{E}[S_\beta S_\beta^\top]\right\}^{-1} S_\beta(Z, \theta_0)$$

so therefore the efficient score function for ψ is

$$S_\psi^{\text{eff}}(Z, \theta_0) = S_\psi(Z, \theta_0) - \mathbb{E}[S_\psi S_\beta^\top] \left\{\mathbb{E}[S_\beta S_\beta^\top]\right\}^{-1} S_\beta(Z, \theta_0)$$

Step 8: Efficient Influence Function

The *efficient influence function* for ψ is

$$\varphi_{\psi}^{\text{eff}}(Z) = \left\{ \mathbb{E}\left[S_{\psi}^{\text{eff}} \left\{ S_{\psi}^{\text{eff}} \right\}^{\top} \right] \right\}^{-1} S_{\psi}^{\text{eff}}(Z, \theta_0)$$

which has variance

$$\left\{ \mathbb{E}\left[S_{\psi}^{\text{eff}} \left\{ S_{\psi}^{\text{eff}} \right\}^{\top} \right] \right\}^{-1}$$

This result is the generalization of the earlier results.

Step 9: Semiparametric modelling

Consider the class of *semiparametric* models

$$\mathcal{P} \equiv \{f_Z(z;\psi,\beta(.)) : \psi \text{ is } q \times 1, \beta(.) \text{ is } \textit{infinite dimensional}\}$$

with true model $f_0(z) = f_Z(z;\psi_0,\beta_0(.))$.

$\beta(.)$ represents an unknown density function, say

Step 9: Semiparametric modelling

Consider the *parametric submodel*

$$\mathcal{P}_{\psi,\gamma} \equiv \{f_Z(z;\psi,\gamma) : \psi \text{ is } q \times 1, \gamma \text{ is } r \times 1\}$$

where

(i) $\mathcal{P}_{\psi,\gamma} \subset \mathcal{P}$;

(ii) $f_0(z) \in \mathcal{P}_{\psi,\gamma}$; that is

$$f_0(z) \equiv f_Z(z;\psi_0,\gamma_0).$$

The parametric submodel is *identical* to the *true model* for one setting of the parameters (ψ,γ).

Step 9: Semiparametric modelling

Note

The parametric submodel is simply a way to allow us to compute score functions explicitly.

Note that the parametric submodel is in general specified in terms of the true model.

Step 9: Semiparametric modelling

Example: Restricted Moment Model

The *restricted moment model*

$$Y_i = \mu(X_i; \psi) + \varepsilon_i$$

where $\beta(.)$ specifies the density of ε_i admits the parametric submodel where

$$\varepsilon_i \sim \text{Normal}(0, \sigma^2)$$

that is, $\gamma \equiv \sigma^2$.

Step 9: Semiparametric modelling

Example: Proportional Hazards Model

In the *proportional hazards model* with hazard function

$$\lambda(t|X;\psi,\beta) = \beta(t)\exp\{\mathbf{x}\psi\}$$

with $\beta(.)$ nonparametrically specified, with true values of the parameters ψ_0 and $\beta_0(.)$.

Step 9: Semiparametric modelling

Example: Proportional Hazards Model

A parametric submodel takes the form

$$\lambda(t|X;\beta,\gamma) = \beta_0(t)\exp\{\gamma_1 g_1(t) + \cdots + \gamma_r g_r(t)\}\exp\{\mathbf{x}\psi\}$$

for specified functions $g_1(t),\ldots,g_r(t)$.

This parametric model is specified in terms of the *true model* $\beta_0(t)$ which is not known. However, if we specify $\psi = \psi_0$, and

$$\gamma_1 = \cdots = \gamma_r = 0$$

we recover the true model.

Step 10: Semiparametric inference

Consider the Hilbert space $\mathcal{H}_q$, and the parametric submodel.

(i) *nuisance tangent space*

$$\Lambda_\gamma \equiv \{\mathbf{B}S_\gamma(Z, \psi_0, \gamma_0) : \mathbf{B} \text{ a } q \times r \text{ matrix}\}$$

where

$$S_\gamma(Z, \psi_0, \gamma_0)$$

is the score function component corresponding to γ from the parametric submodel.

Step 10: Semiparametric inference

(ii) *efficient influence function*

$$\varphi^{\text{eff}}_{\psi,\gamma}(Z) = \{\mathcal{I}(\psi_0, \gamma_0)\}^{-1} \, S^{\text{eff}}_{\psi,\gamma}(Z, \psi_0, \gamma_0)$$

where

$$\mathcal{I}(\psi_0, \gamma_0) = \mathbb{E}\left[S^{\text{eff}}_{\psi,\gamma}(Z, \psi_0, \gamma_0) \left\{ S^{\text{eff}}_{\psi,\gamma}(Z, \psi_0, \gamma_0) \right\}^{\top} \right]$$

Step 10: Semiparametric inference

(iii) *efficient score function* for ψ is

$$S_\psi^{\text{eff}}(Z, \psi_0, \gamma_0) = S_\psi(Z, \psi_0, \gamma_0) - \Pi(S_\psi(Z, \psi_0, \gamma_0)|\Lambda_\gamma).$$

As ever, the RHS is defined in terms of (ψ_0, γ_0), which is identical to $(\psi_0, \beta_0(.))$ by assumption.

Step 10: Semiparametric inference

(iv) *smallest asymptotic variance* amongst RAL estimators for ψ is

$$\left\{ \mathbb{E}\left[S_{\psi}^{\text{eff}}(Z, \psi_0, \gamma_0) \left\{ S_{\psi}^{\text{eff}}(Z, \psi_0, \gamma_0) \right\}^{\top} \right] \right\}^{-1}$$

All expectations are taken with respect to the *true* model

$$f_Z(z; \psi_0, \gamma_0) \equiv f_Z(z; \psi_0, \beta_0).$$

Step 10: Semiparametric inference

- An estimator for ψ is RAL for the semiparametric model if it is RAL for *every* parametric submodel;
- Any influence function of an RAL estimator in the semiparametric model *must* be an influence function of an RAL estimator within a parametric submodel;
- Any influence function of an RAL estimator in the semiparametric submodel must be *orthogonal* to *all* parametric submodel nuisance tangent spaces;
- The variance of any RAL semiparametric influence function must be no smaller than the variance on page 430.

Step 11: The Restricted Mean Model

Suppose scalar response Y follows the model

$$Y = \mu(X; \psi) + \varepsilon$$

with ψ a $q \times 1$ vector. Suppose $\mathcal{P} \equiv \{f_Z(z; \psi, \beta()), z = (y, x)\}$ with

$$f_{Y,X}(y, x) \equiv f_{\varepsilon,X}(y - \mu(x; \psi), x)$$

with the requirement $\mathbb{E}[\varepsilon \mid X] = 0$.

Step 11: The Restricted Mean Model

Write

$$f_{\varepsilon,X}(\varepsilon, \mathbf{x}) = \beta_1(\varepsilon, \mathbf{x})\beta_2(\mathbf{x})$$

where

- Conditional model:

$$\beta_1(\varepsilon, \mathbf{x}) \equiv f_{\varepsilon|X}(\varepsilon|\mathbf{x})$$

- Marginal model;

$$\beta_2(\mathbf{x}) \equiv f_X(\mathbf{x}).$$

Suppose the true (data generating) functions are

$$\beta_{10}(\varepsilon, \mathbf{x}) \qquad \beta_{20}(\mathbf{x}).$$

Step 11: The Restricted Mean Model

We require for all x

$$\int \beta_1(\varepsilon, \mathbf{x})\ d\varepsilon = 1 \qquad \int \varepsilon\, \beta_1(\varepsilon, \mathbf{x})\ d\varepsilon = 0$$

and that $\beta_2(x)$ is non-negative and satisfies

$$\int \beta_2(x)\ dx = 1.$$

With no further restrictions, we have a semiparametric specification with these infinite dimensional nuisance parameters.

Step 11: The Restricted Mean Model

A parametric submodel is

$$\mathcal{P}_{\psi,\gamma} \equiv \{f_Z(z;\psi,\gamma) = f_{\varepsilon|X}(y-\mu(x;\psi)|x;\gamma_1)f_X(x;\gamma_2)\}$$

where γ_1 is $r_1 \times 1$ and γ_2 is $r_2 \times 1$, with $r = r_1 + r_2$.

Denote the true model

$$f_0(z) = f_{\varepsilon|X}(y-\mu(x;\psi)|x;\gamma_{10})f_X(x;\gamma_{20}).$$

Step 12: Efficient Inference

We have in the parametric submodel

$$\log f_Z(z;\psi,\gamma) \equiv \log f_{\varepsilon|X}(\varepsilon|x;\gamma_1) + \log f_X(x;\gamma_2).$$

Therefore

$$S_{\gamma_1}(\varepsilon,\mathbf{x};\psi_0,\gamma_0) = \frac{\partial}{\partial\gamma_1}\left\{\log f_{\varepsilon|X}(\varepsilon|\mathbf{x};\gamma_1)\right\}_{\gamma_1=\gamma_{10}}$$

$$S_{\gamma_2}(\varepsilon,\mathbf{x};\psi_0,\gamma_0) = \frac{\partial}{\partial\gamma_2}\left\{\log f_X(\mathbf{x};\gamma_2)\right\}_{\gamma_2=\gamma_{20}}$$

We will suppress the dependence on (ψ_0,γ_0).

Step 12: Efficient Inference

A typical element in the parametric submodel nuisance tangent space is given by

$$\mathbf{B}S_\gamma(\varepsilon, X) = \mathbf{B}_1 S_{\gamma_1}(\varepsilon, X) + \mathbf{B}_2 S_{\gamma_2}(X)$$

where

- $\mathbf{B}_1$ is $q \times r_1$,
- $\mathbf{B}_2$ is $q \times r_2$.

We define the spaces Λ_γ, Λ_{γ_1} and Λ_{γ_2} by

$$\mathbf{B}S_\gamma(\varepsilon, X) \in \Lambda_\gamma \qquad \mathbf{B}_1 S_{\gamma_1}(\varepsilon, X) \in \Lambda_{\gamma_1} \qquad \mathbf{B}_2 S_{\gamma_2}(X) \in \Lambda_{\gamma_2}.$$

That is, in terms of the corresponding spaces

$$\Lambda_\gamma = \Lambda_{\gamma_1} \oplus \Lambda_{\gamma_2}$$

with Λ_{γ_1} and Λ_{γ_2} orthogonal as

$$\mathbb{E}\big[S_{\gamma_1}(\varepsilon, X)\{S_{\gamma_2}(X)\}^\top\big] = \mathbf{0}_{r_1 \times r_2}$$

by iterated expectation.

Step 12: Efficient Inference

Let

- Λ be the *mean-square closure* of *all* parametric submodel nuisance tangent spaces

$$\Lambda \equiv \{\text{mean-square closure of all } \Lambda_{\gamma_1} \oplus \Lambda_{\gamma_2}\};$$

- $\Lambda_{1s} = \{\text{mean-square closure of all } \Lambda_{\gamma_1}\};$
- $\Lambda_{2s} = \{\text{mean-square closure of all } \Lambda_{\gamma_2}\}$

For Λ, the mean-square closure is the set of functions that can be represented as the limit of sequences of score functions arising from the parametric submodels.

Step 12: Efficient Inference

- Λ_{2s} is a subspace of $\mathcal{H}_q$ arising from the unknown $f_X(x)$ (which is *essentially unrestricted*)
  - comprising elements that are eligible *score functions*;
  - where every bounded element $\alpha(x) \in \Lambda_{2s}$ is the score for some parametric submodel, for example

$$f_X(\mathbf{x}; \gamma_2) = f_0(\mathbf{x})(1 + \gamma_2^\top \alpha(\mathbf{x}))$$

    for γ_2 small enough.

See Tsiatis, pp 78–79.

Step 12: Efficient Inference

- Λ_{1s} is a subspace of $\mathcal{H}_q$ comprising functions $a(\varepsilon, X)$
  - which satisfy
    (i) $\mathbb{E}[a(\varepsilon, X)|X] = \mathbf{0}_q$;
    This says that $a(\varepsilon, x)$ must be a *score function*.
    (ii) $\mathbb{E}[a(\varepsilon, X)\varepsilon|X] = \mathbf{0}_q$
    This says that $a(\varepsilon, x)$ must be a *uncorrelated* with ε for all x, and enforces the requirement

$$\mathbb{E}[\epsilon|X] = 0 \qquad \text{w.p. } 1.$$

  - which arise from some parametric submodel, for example

$$f_{\varepsilon|X}(\varepsilon|\mathbf{x}; \gamma_1) = f_0(\varepsilon|\mathbf{x})(1 + \gamma_1^\top a(\varepsilon, \mathbf{x}))$$

See Tsiatis, pp 80–81.

Step 12: Efficient Inference

Let

$$\alpha(X) \in \Lambda_{2s} \qquad \text{and} \qquad a(\varepsilon, X) \in \Lambda_{1s}.$$

Then

$$\mathbb{E}_{X,\varepsilon}[\alpha(X)^\top a(\varepsilon, X)] = \mathbb{E}_X\left[\alpha(X)^\top \mathbb{E}_{\varepsilon|X}[a(\varepsilon, X) \mid X]\right] = 0$$

by iterated expectation, as

$$\mathbb{E}_{\varepsilon|X}[a(\varepsilon, X) \mid X] = \mathbf{0}_q.$$

Therefore Λ_{1s} and Λ_{2s} are *orthogonal*, and we have precisely characterized the nuisance tangent space as

$$\Lambda = \Lambda_{1s} \oplus \Lambda_{2s}.$$

Step 12: Efficient Inference

The space Λ_{1s} is a space of q-dimensional random functions, $a(\varepsilon, X)$, say, defined by two conditions:

$$\mathbb{E}[a(\varepsilon, X)|X] = \mathbf{0}_q \tag{C1}$$

$$\mathbb{E}[a(\varepsilon, X)\varepsilon|X] = \mathbf{0}_q \tag{C2}$$

Step 12: Efficient Inference

Suppose $\alpha(X) \in \Lambda_{2s}$. Then, for any $a(\varepsilon, X)$ satisfying (C1), we have that

$$\mathbb{E}_{X,\varepsilon}[\alpha(X)^\top a(\varepsilon, X)] = 0$$

by iterated expectation. That is, $\alpha(X)$ is *orthogonal* to $a(\varepsilon, X)$.

Similarly, as

$$\mathbb{E}_{\varepsilon|X}[\alpha(X)\varepsilon|X] = \mathbf{0}_q$$

$\alpha(X)$ *also* satisfies (C2).

Step 12: Efficient Inference

It follows that the *nuisance tangent space* Λ comprises *precisely* those functions that satisfy (C2),

That is, Λ comprises functions $h \equiv h(\varepsilon, X) \in \mathcal{H}_q$ such that

$$\mathbb{E}[h(\varepsilon, X)\varepsilon|X] = \mathbf{0}_q.$$

See Tsiatis, pp 82–83.

Step 12: Efficient Inference

For the efficient score for ψ we need to *project* the ordinary score onto Λ, and take the '*residual*'

$$S_\psi^{\text{eff}}(\varepsilon, X) = S_\psi(\varepsilon, X) - \Pi(S_\psi(\varepsilon, X)|\Lambda).$$

By this construction, we see that

$$S_\psi^{\text{eff}}(\varepsilon, X) \in \Lambda^\perp$$

as

$$\begin{aligned} S_\psi(\varepsilon, X) &= \Pi(S_\psi(\varepsilon, X)|\Lambda) + \{S_\psi(\varepsilon, X) - \Pi(S_\psi(\varepsilon, X)|\Lambda)\} \\ &= \Pi(S_\psi(\varepsilon, X)|\Lambda) + S_\psi^{\text{eff}}(\varepsilon, X). \end{aligned}$$

Step 12: Efficient Inference

In the restricted mean model, the space *orthogonal to the nuisance tangent space* is seen to be

$$\Lambda^{\perp} = \{A(X)\varepsilon : \text{where } A \text{ is } q \times 1\}$$

To see this, note that for all $a(\varepsilon, X)$ satisfying condition (C2)

$$\mathbb{E}[\{a(\varepsilon, X)\}^{\top} A(X)\varepsilon] = 0$$

using iterated expectation.

Step 12: Efficient Inference

The projection of an arbitrary h onto $\Lambda^{\perp}$ is

$$h - \mathbb{E}[h\varepsilon|X]\left\{\mathbb{E}[\varepsilon^2|X]\right\}^{-1}\varepsilon.$$

We can then characterize elements of $\Lambda^{\perp}$ as taking the form

$$h(\varepsilon, X) - \Pi(h(\varepsilon, X)|\Lambda) = g(X)\varepsilon$$

for arbitrary $h \in \mathcal{H}$ and for $g(X)$ the q-dimensional vector

$$g(X) = \mathbb{E}[h\varepsilon|X]\left\{\mathbb{E}[\varepsilon^2|X]\right\}^{-1}$$

Taking $h(\varepsilon, X) \equiv S_{\psi}(\varepsilon, X)$, we obtain the efficient score for ψ.

Step 12: Efficient Inference

The efficient score is then

$$\begin{aligned} S_\psi^{\text{eff}}(\varepsilon, X) &= S_\psi(\varepsilon, X) - \Pi(S_\psi(\varepsilon, X)|\Lambda) \\ &\equiv \mathbb{E}[S_\psi(\varepsilon, X)\varepsilon|X] \left\{\mathbb{E}[\varepsilon^2|X]\right\}^{-1} \varepsilon \\ &= D(X)^\top \left\{V(X)\right\}^{-1} \varepsilon \end{aligned}$$

say, where

$$D(X) \equiv D(X; \psi_0) = \left.\frac{\partial \mu(X; \psi)}{\partial \psi^\top}\right|_{\psi=\psi_0}$$

is a $1 \times q$ vector.

Step 12: Efficient Inference

Hence we must (in principle) solve the estimating equation

$$\sum_{i=1}^{n} D(X_i;\psi)^\top \{V(X_i)\}^{-1} (Y_i - \mu(X_i;\psi)) = \mathbf{0}_q.$$

This requires knowledge of the true model $f_0(.)$, as the formula depends on

$$V(X) \equiv V(X;\psi_0,\beta_0) = \mathbb{E}[\varepsilon^2|X]$$

In practice we cannot implement the estimation procedure without further modelling.

Step 12: Efficient Inference

However, we can implement estimation based on

$$\sum_{i=1}^{n} A(X_i)(Y_i - \mu(X_i;\psi)) = \mathbf{0}_q.$$

where $A(.)$ is some pre-specified $q \times 1$ function of X.

If

$$A(X) = D(X;\psi_0)^\top \{V(X)\}^{-1}$$

we obtain optimal inference.

G-estimation

Consider the simple structural (causal) specification given by

$$\mathbb{E}[Y(z) - Y(0)|X] \equiv \mathbb{E}[Y(z) - Y(0)] = z\psi_0.$$

This states that

- compared to the 'baseline' case of $z = 0$, exposure at level $z = z$ leads to a change $z\psi_0$ in the expected (potential) outcome, and
- that this quantity does not depend on confounders X.

We proceed assuming Z is binary.

G-estimation

The model can be considered as one that specifies that

$$Y(z) = \mu_0(X) + z\psi_0 + \varepsilon$$

where $\mu_0(X) = \mathbb{E}[Y(0)|X]$ is a nuisance component.

Suppose a proposed parametric submodel is

$$\mu_0(X) = \mu_0(X; \beta_0).$$

G-estimation

We consider the probability model based on the semiparametric regression formulation

$$f_{X,Y,Z}(x,y,z) = f_{\varepsilon|X,Z}(y - \mu_0(x;\beta) - z\psi|x,z) f_{X,Z}(x,z)$$

and for simplicity focus on the parametric submodel for the first component

$$f_{\varepsilon|X,Z}(\varepsilon|x,x) \equiv f_{\varepsilon}(\varepsilon;\sigma_0) \equiv \text{Normal}(0,\sigma_0^2).$$

G-estimation

Then the score function for this submodel is

$$\begin{pmatrix} S_\beta \\ S_\psi \end{pmatrix} = \frac{1}{\sigma_0^2} \begin{pmatrix} \mu_{0\beta}(X;\beta_0) \\ Z \end{pmatrix} (Y - \mu_0(X;\beta_0) - Z\psi_0)$$

$$= \frac{1}{\sigma_0^2} \begin{pmatrix} \mu_{0\beta}(X;\beta_0) \\ Z \end{pmatrix} \varepsilon$$

say.

G-estimation

The nuisance tangent space corresponds to the S_β component, and can be seen to take the form

$$\{A(X)\varepsilon : A(X) \text{ arbitrary}\}$$

where $A(X)$ has the same dimension as β.

We then need to project S_ψ onto this tangent space: the projection is the quantity $A_0(X)\varepsilon$, where we must have

$$\mathbb{E}[(Z\varepsilon - A_0(X)\varepsilon)^\top A(X)\varepsilon] = 0 \qquad \forall A(X)$$

G-estimation

By iterated expectation, if

$$e(X) \equiv \mathbb{E}[Z|X]$$

we therefore must have

$$\mathbb{E}[(e(X)\varepsilon - A_0(X)\varepsilon)^\top A(X)\varepsilon] = 0 \qquad \forall A(X)$$

implying that

$$A_0(X) \equiv e(X).$$

G-estimation

Therefore the projection of the score S_ψ onto the nuisance tangent space yields the efficient score function

$$S_\psi^{\text{eff}} = S_\psi - \Pi(S_\psi|\Lambda) = (Z - e(X))\varepsilon.$$

G-estimation

Therefore efficient estimation is achieved by solving

$$\sum_{i=1}^{n} \begin{pmatrix} \dot{\mu}_{0\beta}(\mathbf{x}_i;\beta) \\ z_i - e(\mathbf{x}_i) \end{pmatrix} (y_i - \mu_0(\mathbf{x}_i;\beta) - z_i\psi) = \mathbf{0}$$

which corresponds to G-estimation.

Inverse Probability Weighting

Weighting estimators can be justified using semiparametric theory by adopting a *missing* (or *coarsened*) data strategy.

Recall that for the ATE in the binary case, using the potential outcome notation

$$\delta = \mu(1) - \mu(0) = \mathbb{E}[Y(1) - Y(0)]$$

Inverse Probability Weighting

If such data were available, we would use the estimator

$$\mu(z) = \frac{1}{n} \sum_{i=1}^{n} Y_i(z).$$

based on the *complete data* estimating equation

$$\sum_{i=1}^{n} (Y_i(z) - \mu(z)) = 0.$$

However, we do not observe the counterfactual outcomes.

Inverse Probability Weighting

We need to utilize the *observed data* estimating equation:

$$\sum_{i=1}^{n} \left(\frac{\mathbb{1}_{\{z\}}(Z_i) Y_i}{f_{Z|X}(Z_i|X_i)} - \mu(z) \right) = 0.$$

With $z = 1$, this becomes

$$\sum_{i=1}^{n} \left(\frac{Z_i Y_i}{e(X_i)} - \mu(1) \right) = 0.$$

Inverse Probability Weighting

To construct the efficient (observed data) score and influence function, it can be shown first that all observed data influence functions can be written

$$h(X, Y, Z) + L(X, Y, Z)$$

where $h(X, Y, Z)$ satisfies

$$\mathbb{E}[h(X, Y, Z)|Y(0), Y(1), X] = Y(1) - \mu(1)$$

and $L(X, Y, Z)$ satisfies

$$\mathbb{E}[L(X, Y, Z)|Y(0), Y(1), X] = 0.$$

The function L is termed the *augmentation function*.

Inverse Probability Weighting

From above, one suitable h is

$$h(X, Y, Z) = \frac{ZY}{e(X)} - \mu(1)$$

as

$$\begin{aligned}
\mathbb{E}\left[\frac{ZY}{e(X)}\middle| Y(0), Y(1), X\right] &= \mathbb{E}\left[\frac{ZY(1)}{e(X)}\middle| Y(0), Y(1), X\right] \\
&= Y(1)\mathbb{E}\left[\frac{Z}{e(X)}\middle| Y(0), Y(1), X\right] \\
&= Y(1).
\end{aligned}$$

Inverse Probability Weighting

For $L(X, Y, Z)$ we can always write in the binary case

$$L(X, Y, Z) = (1 - Z)L_0(X, Y) + ZL_1(X, Y)$$

Therefore, taking expectations

$$\begin{aligned}\mathbb{E}[L(X, Y, Z)|Y(0), Y(1), X] = (1 - e(X))L_0(X, Y(0)) \\ + e(X)L_1(X, Y(1))\end{aligned}$$

Inverse Probability Weighting

Equating this to zero implies that

$$L_0(X, Y(0)) = -\frac{e(X)}{(1 - e(X))} L_1(X, Y(1)).$$

provided $0 < e(X) < 1$.

Further, as the left hand side is a function of $Y(0)$ and the right hand side is a function of $Y(1)$, this equation can only hold in general if L_0 and L_1 do not depend on Y at all.

Inverse Probability Weighting

Thus we can simplify

$$L_0(X) = -\frac{e(X)}{(1-e(X))} L_1(X).$$

so that

$$\begin{aligned} L(X,Y,Z) \equiv L(X,Z) &= \left(-(1-Z)\frac{e(X)}{(1-e(X))} + Z \right) L_1(X) \\ &= \left(\frac{Z-e(X)}{1-e(X)} \right) L_1(X). \end{aligned}$$

Inverse Probability Weighting

Combining terms in X, we see that the space of augmentation functions takes the form

$$\Lambda = \{(Z - e(X))g(X) : g(X) \text{ arbitrary}\}.$$

We find the optimal influence function by projecting

$$h(X, Y, Z) = \frac{ZY}{e(X)} - \mu(1)$$

onto Λ; this identifies a specific element of Λ

$$(Z - e(X))g_0(X)$$

say that obeys the usual orthogonality results.

Inverse Probability Weighting

That is, for arbitrary g, we need to solve for g_0 the condition

$$\mathbb{E}\left[\left(\frac{ZY}{e(X)}-\mu(1)-(Z-e(X))g_0(X)\right)(Z-e(X))g(X)\right]=0.$$

The expectation can be rewritten

$$\mathbb{E}_{X,Z}\left[\left(\frac{Z\mu(X,Z)}{e(X)}-\mu(1)-(Z-e(X))g_0(X)\right)(Z-e(X))g(X)\right]$$

Inverse Probability Weighting

Using iterated expectation, we have conditional on X that the interior expectation of Z is

$$(1-e(X))(-\mu(1)+e(X)g_0(X))(-e(X))g(X)$$

$$+e(X)\left(\frac{\mu(X,1)}{e(X)}-\mu(1)-(1-e(X))g_0(X)\right)(1-e(X))g(X).$$

To make this identically zero in expectation for any $g(X)$, we must have

$$e(X)(1-e(X))\left(\frac{\mu(X,1)}{e(X)}-g_0(X)\right)=0.$$

Inverse Probability Weighting

We find that

$$g_0(X) = \frac{\mu(X,1)}{e(X)} = \frac{1}{e(X)}\mathbb{E}[Y|X,Z=1]$$

so that

$$h(X,Y,Z)+L(X,Z) = \frac{ZY}{e(X)} - \frac{(Z-e(X))}{e(X)}\mathbb{E}[Y|X,Z=1] - \mu(1).$$

Inverse Probability Weighting

Therefore, the efficient influence function for $\mu(1)$ is

$$\varphi^{\text{eff}}(X,Y,Z) = \frac{ZY}{e(X)} - \frac{(Z-e(X))}{e(X)}\mu(X,1) - \mu(1)$$

or equivalently

$$\varphi^{\text{eff}}(X,Y,Z) = \frac{Z(Y-\mu(X,1))}{e(X)} + \mu(X,1) - \mu(1)$$

which is the basis of AIPW estimation.

Inverse Probability Weighting

The corresponding estimating equation is therefore

$$\sum_{i=1}^{n}\left(\frac{Z_i Y_i}{1-e(X_i)}-\frac{(Z_i-e(X_i))}{e(X_i)}\mu(X_i,1)-\mu(1)\right)=0.$$

Similarly for $\mu(0)$ we have

$$\sum_{i=1}^{n}\left(\frac{(1-Z_i)Y_i}{e(X_i)}+\frac{(Z_i-e(X_i))}{1-e(X_i)}\mu(X_i,0)-\mu(0)\right)=0.$$

as

$$(1-Z_i)-(1-e(X_i))=-(Z_i-e(X_i)).$$