Efficient baseline adjustment in a randomized trial

Brice Ozenne

November 6, 2020

Disclaimer: this note is a compilation of section 5.4 of Tsiatis (2007), Zhang and Gilbert (2010) and a note by Torben Martinussen.

1 Motivation, objective, and notations

We consider a randomized trial with a single binary or continuous outcome (Y), two treatment arms: placebo ($A = 0$) and active ($A = 1$), and some baseline variables (Z). There are in total $n = n_0 + n_1$ patients, n_0 in the placebo arm and n_1 in the treatment arm. The observed data is therefore $\chi = (\chi_i)_{i\in\{1,\dots,n\}} = (Y_i, A_i, Z_i)_{i\in\{1,\dots,n\}}$.

Our parameter of interest is the average difference in outcome:

$$\psi = \mathbb{E}\left[Y|A=1\right] - \mathbb{E}\left[Y|A=0\right] = \mu_1 - \mu_0$$

which we would like to estimate as efficiently as possible by making use of the baseline variables. We denote $\pi = \mathbb{P}\left[A=1\right]$ which is known.

2 Naive estimator

A possible estimator for ψ is:

$$\hat{\psi}_n = \frac{\sum_{i=1}^n A_i Y_i}{\sum_{i=1}^n A_i} - \frac{\sum_{i=1}^n (1-A_i) Y_i}{\sum_{i=1}^n (1-A_i)}$$

which satisfies the following decomposition:

$$\begin{aligned}
\sqrt{n}\left(\hat{\psi}_n - \psi\right) &= \sqrt{n}\left(\frac{\sum_{i=1}^n A_i Y_i}{\sum_{i=1}^n A_i} - \mu_1\right) - \sqrt{n}\left(\frac{\sum_{i=1}^n (1-A_i) Y_i}{\sum_{i=1}^n (1-A_i)} - \mu_0\right) \\
&= \sqrt{n}\frac{\sum_{i=1}^n A_i (Y_i - \mu_1)}{\sum_{i=1}^n A_i} - \sqrt{n}\frac{\sum_{i=1}^n (1-A_i)(Y_i - \mu_0)}{\sum_{i=1}^n (1-A_i)} \\
&= \frac{1}{\sqrt{n}}\frac{\sum_{i=1}^n A_i (Y_i - \mu_1)}{\frac{1}{n}\sum_{i=1}^n A_i} - \frac{1}{\sqrt{n}}\frac{\sum_{i=1}^n (1-A_i)(Y_i - \mu_0)}{\frac{1}{n}\sum_{i=1}^n (1-A_i)} \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{A_i}{\pi}\left(Y_i - \mu_1\right) - \frac{(1-A_i)}{1-\pi}\left(Y_i - \mu_0\right) + o_p(1) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathcal{IF}_{\hat{\mu}_1}(\chi_i) - \mathcal{IF}_{\hat{\mu}_0}(\chi_i) + o_p(1) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathcal{IF}_{\hat{\psi}}(\chi_i) + o_p(1)
\end{aligned}$$

where $\mathcal{IF}_x$ denotes the influence function associated with the estimator x.

3 Derivation of the semi-parametric efficient estimator

3.1 Geometry of the set of all influence function

The log-likelihood can be decomposed as:

$$\log(f(Y, A, Z)) = \log(f(Y|A, Z)) + \log(f(A|Z)) + \log(f(Z))$$

While f denotes the true density, we will denote by f_θ a parametric model for this density with parameter θ, where for a specific parameter value (denoted θ_0), the modeled density equal the true density (i.e. $f_{\theta_0} = f$). For instance $Z \sim \mathcal{N}(0, 1)$ and $f_\theta(Z)$ could be the density of a Gaussian distribution; in this case θ would be a vector composed of the mean and variance parameters and $\theta_0 = (0, 1)$. We will also denote by $\mathcal{S}_\theta(Y|A, Z) = \frac{\partial \log(f_\theta(Y|A,Z))}{\partial \theta}$ the associated score function, and by $\{B\mathcal{S}_\theta(Y|A, Z), \forall B\}$ its nuisance tangent space, i.e. the space of all linear combinations of the score function.

If there was no restriction (i.e no randomization) the terms of the log-likelihood would be variationnally independent and the entire Hilbert space [1] could therefore be partitionned in three orthogonal spaces (theorem 4.5 in Tsiatis (2007)):

$$\mathcal{H} = \mathcal{T}_1 \oplus \mathcal{T}_2 \oplus \mathcal{T}_3$$

[1]Here, when Z has dimension 1, the Hilbert space is the space of 3-dimensional mean-zero finite-variance measurable functions, equipped with the covariance inner product.

where $\mathcal{T}_1$ (resp $\mathcal{T}_2$ and $\mathcal{T}_3$) is the mean-square closure of parametric submodel tangent spaces for $f(Y|A,Z)$ (resp. $f(A|Z)$ and $f(Z)$). More precisely, $\mathcal{T}_1$ is the space of functions $h(Y|A,Z) \in \mathcal{H}$ such that there exists, for a sequence of parametric submodel indexed by $j \in \mathbb{N}$, $\{B_j \mathcal{S}_{\theta,j}(Y|A,Z)\}_{j\in\mathbb{N}}$ such that:

$$||h(Y|A,Z) - B_j \mathcal{S}_{\theta,j}(Y|A,Z)||^2 \xrightarrow{j\to\infty} 0$$

Since the corresponding score should have conditional expectation 0, we get that $\mathcal{T}_1$ is the space of functions of Y, A, Z with finite variance and null expectation conditional to A and Z. A similar result holds for the other spaces which is summarized as:

$$\begin{aligned}
\mathcal{T}_1 &= \{\alpha_1(Y,A,Z), \mathbb{E}\left[\alpha_1(Y,A,Z)|A,Z\right] = 0\} \\
\mathcal{T}_2 &= \{\alpha_2(A,Z), \mathbb{E}\left[\alpha_2(A,Z)|Z\right] = 0\} \\
\mathcal{T}_3 &= \{\alpha_3(Z), \mathbb{E}\left[\alpha_3(Z)\right] = 0\}
\end{aligned}$$

In our application, because of randomization $f(A|Z) = f(A) = \pi^A(1-\pi)^{1-A}$ is known. In that case the tangent space is equal to:

$$\mathcal{T} = \mathcal{T}_1 \oplus \mathcal{T}_3$$

so the orthogonal of the tangent space, $\mathcal{T}^\perp$, is $\mathcal{T}_2$. We first introduce an alternative representation of the element of $\mathcal{T}_2$:

$$\mathcal{T}_2 = \{\alpha_2(A,Z) - \mathbb{E}\left[\alpha_2(A,Z)|Z\right]\}$$

Moreover since A is binary we can write without loss of generality $\alpha_2(A,Z) = Af(Z) + g(Z)$. So:

$$\begin{aligned}
\mathcal{T}_2 &= \{Af(Z) + g(Z) - \mathbb{E}\left[Ag(Z) + g(Z)|Z\right]\} \\
&= \{(A-\pi)g(Z)\}
\end{aligned}$$

From the semi-parametric theory we know that the set of all influence function is spanned by the orthogonal to the tangent space:

$$\begin{aligned}
\{\mathcal{IF}_{\hat{\psi}} + \mathcal{T}_2\} &= \left\{\mathcal{IF}_{\hat{\psi}} + (A-\pi)g(Z)\right\} \\
&= \left\{\frac{A}{\pi}(Y-\mu_1) - \frac{(1-A)}{1-\pi}(Y-\mu_0) + (A-\pi)g(Z)\right\}
\end{aligned}$$

where g is an arbitrary function.

3.2 Identification of the efficient influence function

The efficient influence function, $\mathcal{IF}^{eff}_{\hat{\psi}}$, is orthogonal to the nuisance tangence space (here orthogonal to $\mathcal{T}$). So we just need to remove the composant of the naive influence function that lies in the nuisance tangent space:

$$\begin{aligned}\mathcal{IF}^{eff}_{\hat{\psi}} &= IF_{\hat{\psi}} - \Pi(IF_{\hat{\psi}}|\mathcal{T}^{\perp}) \\ &= IF_{\hat{\psi}} - \Pi(IF_{\hat{\psi}}|\mathcal{T}_2)\end{aligned}$$

where $\Pi(.|x)$ denotes the projection of . onto x. We first note that any element h of the Hilbert space can be decomposed as:

$$\begin{aligned}h(Y, A, Z) &= h_1(Y, A, Z) + h_2(Y, A, Z) + h_3(Y, A, Z) \\ h_1 &= \mathbb{E}\left[h(Y, A, Z)|Z\right] \\ h_2 &= \mathbb{E}\left[h(Y, A, Z)|Z\right] - \mathbb{E}\left[h(Y, A, Z)|A, Z\right] \\ h_3 &= \mathbb{E}\left[h(Y, A, Z)|A, Z\right] - h(Y, A, Z)\end{aligned}$$

Theorem 4.5 in Tsiatis (2007) shows that for any $j \in \{1, 2, 3\}$, $h_j = \Pi(h|\mathcal{T}_j)$. So:

$$\begin{aligned}\Pi(IF_{\hat{\psi}}|\mathcal{T}_2) =& \mathbb{E}\left[IF_{\hat{\psi}}|Z\right] - \mathbb{E}\left[IF_{\hat{\psi}}|A, Z\right] \\ =& \mathbb{E}\left[\mathbb{E}\left[\frac{A}{\pi}(Y-\mu_1) - \frac{(1-A)}{1-\pi}(Y-\mu_0)\middle|A, Z\right]\middle|Z\right] \\ &- \mathbb{E}\left[\frac{A}{\pi}(Y-\mu_1) - \frac{(1-A)}{1-\pi}(Y-\mu_0)\middle|A, Z\right] \\ =& \frac{\mathbb{E}\left[A\right]}{\pi}\left(\mathbb{E}\left[Y|A=1, Z\right] - \mu_1\right) - \frac{\mathbb{E}\left[1-A\right]}{1-\pi}\left(\mathbb{E}\left[Y|A=0, Z\right] - \mu_0\right) \\ &- \left(\frac{A}{\pi}\left(\mathbb{E}\left[Y=1|A, Z\right] - \mu_1\right) - \frac{(1-A)}{1-\pi}\left(\mathbb{E}\left[Y|A=0, Z\right] - \mu_0\right)\right) \\ =& \frac{\pi - A}{\pi}\left(\mathbb{E}\left[Y|A=1, Z\right] - \mu_1\right) - \frac{(1-p)-(1-A)}{1-\pi}\left(\mathbb{E}\left[Y|A=0, Z\right] - \mu_0\right)\end{aligned}$$

which lead to the following expression for the efficient influence function:

$$\begin{aligned}\mathcal{IF}^{eff}_{\hat{\psi}} =& \frac{A}{\pi}(Y-\mu_1) + \frac{\pi - A}{\pi}\left(\mathbb{E}\left[Y|A=1, Z\right] - \mu_1\right) \\ &- \frac{(1-A)}{1-\pi}(Y-\mu_0) - \frac{(1-p)-(1-A)}{1-\pi}\left(\mathbb{E}\left[Y|A=0, Z\right] - \mu_0\right) \\ =& \mathcal{IF}^{eff}_{\hat{\mu}_1} - \mathcal{IF}^{eff}_{\hat{\mu}_0}\end{aligned}$$

Solving $\sum_{i=1}^{n} \mathcal{IF}_{\hat{\mu}_1}^{eff} = 0$ in μ_1 gives:

$$\begin{aligned}
\sum_{i=1}^{n} \frac{A_i + \pi - A_i}{\pi} \tilde{\mu}_1 &= \sum_{i=1}^{n} \left(\frac{A_i Y_i}{\pi} + \frac{\pi - A_i}{\pi} \mathbb{E}\left[Y|A=1, Z\right] \right) \\
\tilde{\mu}_1 &= \frac{1}{n_1} \sum_{i=1}^{n} \left(A_i Y_i + (\pi - A_i) \mathbb{E}\left[Y|A=1, Z\right] \right) \\
&= \hat{\mu}_1 + \frac{1}{n_1} \sum_{i=1}^{n} (\pi - A_i) \mathbb{E}\left[Y|A=1, Z\right]
\end{aligned}$$

Similarly:

$$\begin{aligned}
\tilde{\mu}_0 &= \frac{1}{n_0} \sum_{i=1}^{n} \left((1 - A_i) Y_i + ((1 - \pi) - (1 - A_i)) \mathbb{E}\left[Y|A=0, Z\right] \right) \\
&= \hat{\mu}_0 + \frac{1}{n_0} \sum_{i=1}^{n} ((1 - \pi) - (1 - A_i)) \mathbb{E}\left[Y|A=0, Z\right]
\end{aligned}$$

and:

$$\begin{aligned}
\tilde{\psi} &= \tilde{\mu}_1 - \tilde{\mu}_0 \\
&= \hat{\psi} + \frac{1}{n_1} \sum_{i=1}^{n} (\pi - A_i) \mathbb{E}\left[Y|A=1, Z\right] - \frac{1}{n_0} \sum_{i=1}^{n} ((1 - \pi) - (1 - A_i)) \mathbb{E}\left[Y|A=0, Z\right]
\end{aligned}$$

4 Relationship to the G-formula computation

When performing a logistic regression including an intercept, A, and Z the score equation is:

$$\sum_{i=1}^{n} X_i \left(Y_i - \frac{1}{1 + exp(-X_i \theta)} \right) = 0$$

where $X_i = (1, A_i, Z_i)$ is the design matrix and $\theta = (\theta_1, \theta_A, \theta_Z)$ the set of model parameters. We can in fact reparametrize it as $X_i = (1 - A_i, A_i, Z_i)$ with $\theta = (\theta_{1-A}, \theta_A, \theta_Z)$. Then the logistic regression solves the following equations:

$$\begin{aligned}
\sum_{i=1}^{n} A_i \left(Y_i - \frac{1}{1 + exp(-X_i \theta)} \right) = 0 \\
\sum_{i=1}^{n} (1 - n A_i) \left(Y_i - \frac{1}{1 + exp(-X_i \theta)} \right) = 0
\end{aligned}$$

i.e.

$$\frac{1}{n}\sum_{i=1}^{n}\frac{A_i}{\pi}\left(Y_i - \frac{1}{1+exp(-\theta_A - Z_i\theta_Z)}\right) = 0$$
$$\frac{1}{n}\sum_{i=1}^{n}\frac{1-A_i}{1-\pi}\left(Y_i - \frac{1}{1+exp(-\theta_{1-A} - Z_i\theta_Z)}\right) = 0$$

So the G-formula estimator is asymptotically equivalent to the efficient estimator:

$$\begin{aligned}
\bar{\mu}_1 &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{1+exp(-\theta_A - Z_i\theta_Z)} \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[Y|A_i=1, Z_i\right] + o_p(1) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[Y|A_i=1, Z_i\right] + \frac{A_i}{\pi}\left(Y_i - \mathbb{E}\left[Y|A_i=1, Z_i\right]\right) + o_p(1) \\
&= \tilde{\mu}_1 + o_p(1)
\end{aligned}$$

Because

$$\begin{aligned}
\mathbb{E}\left[\frac{A}{\pi}\left(Y - \mathbb{E}\left[Y|A=1, Z\right]\right)\right] &= \mathbb{E}\left[\frac{A}{\pi}\left(Y - \mathbb{E}\left[Y|A, Z\right]\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{A}{\pi}\left(Y - \mathbb{E}\left[Y|A, Z\right]\right)\middle|A, Z\right]\right] \\
&= \mathbb{E}\left[\frac{\mathbb{E}\left[A\right]}{\pi}\left(\mathbb{E}\left[Y|A, Z\right] - \mathbb{E}\left[Y|A, Z\right]\right)\right] = 0
\end{aligned}$$

5 References

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

Zhang, M. and Gilbert, P. B. (2010). Increasing the efficiency of prevention trials by incorporating baseline covariates. *Statistical communications in infectious diseases*, 2(1).