Inverse Weighting and Double Robustness

December 10, 2018

Suppose the “full data” consists of (Δ, Y, Z), but we observe only $X = (\Delta, \Delta Y, Z)$, where $\Delta \in \{0,1\}$. Think of an outcome Y, an indicator Δ that indicates whether Y is observed or missing and a covariate Z.

We assume that Y is “missing at random”: the variables Y and Δ are conditionally independent given Z. This is the equivalent of the “conditional exchangeability” assumption if Y were a counterfactual. Thus the conditional probability

$$\pi(Z) := \mathrm{P}(\Delta = 1 |\, Z, Y)$$

depends indeed only on Z and we can complete the description of the distribution of the full data by specifying a conditional distribution of Y given Z and a marginal distribution of Z.

Suppose that we are interested in estimating

$$\mu = \mathrm{E}Y = \mathrm{E}b(Z),$$

if $b(Z) = \mathrm{E}(Y |\, Z)$ is the conditional expectation of Y given Z.

If π were known, then one estimator based on a sample $X_1, \ldots, X_n$ of observed data is

$$T = \mathbb{P}_n\Bigl(\frac{\Delta Y}{\pi(Z)}\Bigr) = \frac{1}{n}\sum_{i=1}^n \frac{\Delta_i Y_i}{\pi(Z_i)}.$$

Here $\mathbb{P}_n$ denotes the empirical distribution, and we shall similarly use short notation for expectations, such as

$$\mathrm{E}T = P\Bigl(\frac{\Delta Y}{\pi(Z)}\Bigr) = \mathrm{E}\Bigl(\frac{\Delta Y}{\pi(Z)}\Bigr) = \mathrm{E}\Bigl(\frac{\mathrm{E}(\Delta |\, Z) Y}{\pi(Z)}\Bigr) = \mathrm{E}Y = \mu.$$

This equation shows that T is unbiased for μ. By the central limit theorem $\sqrt{n}(T - \mu)$ tends to a mean-zero normal distribution, so the order of magnitude of the estimation error is $1/\sqrt{n}$, which is good.

In practice π will not be known and hence it will have to be replaced by an estimator $\hat\pi$. We shall now show that even if π were known, the estimator will become better by replacing π by $\hat\pi$, in the sense that the resulting estimator will be asymptotically normal with smaller variance (under mild conditions on $\hat\pi$).

The explanation of this seemingly surprising fact is that T itself is a bad estimator. The reweighting gets the bias equal to zero, but it does not optimize the variance in any way.

For a proof it helps to consider first, for a given function c,

$$T_c = \mathbb{P}_n\Bigl(\frac{\Delta Y}{\pi(Z)} - \frac{\Delta - \pi(Z)}{\pi(Z)}c(Z)\Bigr).$$

If π were known, then this would also be an unbiased estimator for μ, as the added term has mean zero. It is not difficult to find c so that its variance is smaller. For instance, fix some "direction" d and determine a scalar λ so that the newly added term with $c = \lambda d$ is the orthogonal projection of $\Delta Y/\pi(Z)$ onto the one-dimensional linear space spanned by the variable $(\Delta - \pi(Z))d(Z)/\pi(Z)$. Since projection decreases variance, the resulting estimator $T_{\lambda d}$ will be better than T. The appropriate value of λ (setting the covariance between T_c and the correction term zero) is given by

$$\lambda = \frac{\mathrm{cov}\Bigl(\frac{\Delta Y}{\pi(Z)}, \frac{\Delta - \pi(Z)}{\pi(Z)}d(Z)\Bigr)}{\mathrm{var}\Bigl(\frac{\Delta - \pi(Z)}{\pi(Z)}d(Z)\Bigr)} = \frac{P\Bigl(\frac{bd(1-\pi)}{\pi}\Bigr)}{P\Bigl(\frac{d^2(1-\pi)}{\pi}\Bigr)}.$$

In the last two expressions the P means taking the expectation relative to the distribution of Z. The expression on the right now also depends on another unknown $b(Z) = \mathrm{E}(Y|\,Z)$, which would have to be estimated. (Estimating the expectation in which b appears is enough.) We shall turn to this later on, but this is irrelevant for the following.

We shall argue that the effect of estimating π in the original estimator T is that the resulting estimator behaves like T_c, with a projected $c = \lambda d$, at least asymptotically. Thus estimating π gives a projection, and hence variance reduction, without having to estimate b.

Assume that we use a model $\theta \mapsto \pi_\theta$ smoothly indexed by a (for simplicity) one-dimensional parameter θ, which we estimate by maximum likelihood:

$$\hat\theta = \operatorname*{argmax}_\theta \prod_{i=1}^n \pi_\theta(Z_i)^{\Delta_i}(1 - \pi_\theta(Z_i))^{1-\Delta_i}.$$

Suppose that π_θ for a given parameter value θ is the correct conditional probability function. Then by standard theory $\hat\theta - \theta$ will be asymptotically normal (under mild conditions) and asymptotically linear with influence function equal to the score function of the model divided by the Fisher information, i.e.

$$\hat\theta - \theta \simeq \frac{1}{i_\theta}\mathbb{P}_n\Bigl(\frac{\Delta - \pi_\theta(Z)}{\pi_\theta(1-\pi_\theta)(Z)}\dot\pi_\theta(Z)\Bigr),$$

where i_θ is the Fisher information

$$i_\theta = \mathrm{var}\Bigl(\frac{\Delta - \pi_\theta(Z)}{\pi_\theta(1-\pi_\theta)(Z)}\dot\pi_\theta(Z)\Bigr) = P\Bigl(\frac{\dot\pi_\theta^2}{\pi_\theta(1-\pi_\theta)}\Bigr).$$

(The error in $\simeq$ is $o_P(1/\sqrt{n})$ and hence is negligible if we are interested in the limit law of $\sqrt{n}(\hat\theta - \theta)$.) Let $\hat T$ be the estimator obtained by plugging in $\pi_{\hat\theta}$ for

π in T. By first order Taylor expansion

$$\begin{aligned}\hat{T} - T &\simeq (\hat{\theta} - \theta)\frac{\partial}{\partial\theta}\mathbb{P}_n\Big(\frac{\Delta Y}{\pi_\theta(Z)}\Big)\\ &\simeq \frac{1}{i_\theta}\mathbb{P}_n\Big(\frac{\Delta - \pi_\theta(Z)}{\pi_\theta(1-\pi_\theta)(Z)}\dot{\pi}_\theta(Z)\Big)P\Big(-\frac{\pi_\theta b}{\pi_\theta^2}\dot{\pi}_\theta\Big).\end{aligned}$$

The expression on the right side has the form of the correction term added into T_c, for $c = \lambda d$, and $d = \dot{\pi}_\theta/(1-\pi_\theta)$ and $\lambda = i_\theta^{-1}P(b\dot{\pi}_\theta/\pi_\theta)$. It can be checked that λ is indeed the scalar corresponding to the projection, as in the preceding. Thus $\hat{T}$ is asymptotically equivalent to $T_{\lambda d}$ and hence has smaller variance than T. The extra randomness introduced by estimating θ has lead from the bad estimator T to a better estimator with a projected influence function.

This works for any smooth parametric model $\theta \mapsto \pi_\theta$ provided the true π belongs to this model. In fact, the preceding argument does not so much rely on the model, but on the asymptotic linearity of the resulting estimator, with an influence function of the type $(\Delta - \pi(Z))c(Z)/\pi(Z)$, which is equal to the correction term. Estimating the parameter has the effect of changing the original influence function $\Delta Y/\pi(Z)$ to the influence function of T_c, which is a projection.

In the preceding we have verified this projection property by explicit calculations, and particularly the value of λ may come as a happy surprise. Of course, there is no coincidence. An intuitive explanation for the projection property comes from the 'calculus of score functions', as in semiparametric theory. The function $Y\Delta/\pi(Z)$ is an influence function of the functional μ in the model in which π is known. Because the functional μ does not depend on π, in a model where π is not known, an influence function for μ must be orthogonal to scores for π, and hence the new influence influence after estimating π must differ from the old one by a function orthogonal to the score for π.

The argument (calculation or intuition) may be extended to multi-parameter models for π, which will give multiple influence functions, and hence projections on a space of dimension more than one. The bigger model for π you fit, the more you cut the variance!

There is a caveat to this, that in the preceding we use approximations that are correct asymptotically. The main finite-sample error is probably in the bias terms, which are order $o(1/\sqrt{n})$ and negligible in the asymptotics. While T is exactly unbiased, this is not true for $\hat{T}$.

Adding some bias and cutting some variance is definitely wise, but it is not very clear where this stops. We can gain some insight from the extreme case, where we optimize over an infinite-dimensional model, as follows.

Rather than optimizing $\operatorname{var} T_c$ over a given one-dimensional submodel, consider optimizing over all functions c. This leads to the function $c = b$. Indeed, one can verify that, for every c,

$$\operatorname{cov}\Big(\frac{\Delta Y}{\pi(Z)} - \frac{\Delta - \pi(Z)}{\pi(Z)}b(Z), \frac{\Delta - \pi(Z)}{\pi(Z)}c(Z)\Big) = 0.$$

This shows that $(\Delta-\pi(Z))b(Z)/\pi(Z)$ is the orthogonal projection onto the linear space of variables $(\Delta-\pi(Z))c(Z)/\pi(Z)$, when c varies.

Since we shall not know the function b, the "estimator" T_b is not useful, but the calculation suggests to use initial estimators $\hat\pi$ and $\hat b$, and then use the estimator

$$\hat T_{\hat b}=\mathbb{P}_n\Bigl(\frac{\Delta Y}{\hat\pi(Z)}-\frac{\Delta-\hat\pi(Z)}{\hat\pi(Z)}\hat b(Z)\Bigr).$$

This is the so-called "double-robust" estimator, which should work provided at least one of the two estimators $\hat\pi$ and $\hat b$ is accurate. Under relatively mild conditions on $\hat b$ and $\hat\pi$ the difference $T_{\hat b}-T_b$ is $o_P(1/\sqrt{n})$ and hence the sequence $\sqrt{n}(\hat T_{\hat b}-\mu)$ tends to a zero-mean Gaussian limit with variance

$$\mathrm{var}\Bigl(\frac{\Delta Y}{\pi(Z)}-\frac{\Delta-\pi(Z)}{\pi(Z)}b(Z)\Bigr).$$

This is known to be the smallest possible variance in the semi- or nonparametric sense for the model in which π and the distribution of Y given Z are completely unknown.

A proof that $\hat T_{\hat b}-T_b=o_P(1/\sqrt{n})$ can consist of establishing the following two results:

$$\mathbb{G}_n\Bigl[\Bigl(\frac{\Delta Y}{\hat\pi(Z)}-\frac{\Delta-\hat\pi(Z)}{\hat\pi(Z)}\hat b(Z)\Bigr)-\Bigl(\frac{\Delta Y}{\pi(Z)}-\frac{\Delta-\pi(Z)}{\pi(Z)}b(Z)\Bigr)\Bigr]\overset{P}{\to}0,$$
$$\sqrt{n}P\Bigl[\Bigl(\frac{\Delta Y}{\hat\pi(Z)}-\frac{\Delta-\hat\pi(Z)}{\hat\pi(Z)}\hat b(Z)\Bigr)-\Bigl(\frac{\Delta Y}{\pi(Z)}-\frac{\Delta-\pi(Z)}{\pi(Z)}b(Z)\Bigr)\Bigr]\to 0.$$

Here $\mathbb{G}_n=\sqrt{n}(\mathbb{P}_n-P)$ is the empirical process of $X_1,\ldots,X_n$, and the first condition would be true if the functions between the brackets fall in a Donsker class and $\hat\pi$ and $\hat b$ are consistent for π and b, in the sense that the second moments of the functions between square brackets tend to zero. This is a relatively mild condition, which only starts to hurt if the models for π and b are really large (not restricted by e.g. existence of a certain number of derivatives). The expression in the second line can be rewritten as

$$\sqrt{n}P\frac{1}{\hat\pi}\bigl(\pi b-(\pi-\hat\pi)\hat b-b\hat\pi\bigr)=\sqrt{n}P\frac{1}{\hat\pi}\bigl((\pi-\hat\pi)(b-\hat b)\bigr).$$

This tends to zero if the product $(\pi-\hat\pi)(b-\hat b)$ tends to zero sufficiently fast. In particular, it would suffice that one factor is $O(1/\sqrt{n})$ and the second one is $o(1)$. This we might get by fitting a correct parametric model to one parameter and a nonparametric model to the other, and is one form of "double robustness".

We might also estimate both parameters nonparametrically, and then the bias will be negligible at $o_P(1/\sqrt{n})$-rate provided the true parameters π and b are smooth enough, relative to the dimension of Z.

If the estimators $\hat\pi$ and $\hat b$ are not consistent for the true parameters π and b, then the estimators $\hat T_{\hat b}$ and T_b need not be asymptotically equivalent, but the same argument shows that we still retain a $O_P(1/\sqrt{n})$ estimation rate if one of the two estimators converges at $O_P(1/\sqrt{n})$ rate.

In all of the preceding we have π in the denominator, and we silently understand this not to cause trouble. If $\pi(z)$ or its estimator will be close to zero for some set of z, then clearly there is trouble already with the initial estimator T and it may get worse by estimating π.