On the weighted least squares estimator for the ATE and its relation to matching by propensity score

Gonzalo Mena

April 1, 2016

1 Main result

The following proposition allow us to link the matching by propensity score and weighted least squares methods.

Proposition: Consider the following estimator for the ATE

$$\hat{\tau} = \hat{\beta}, \quad (\hat{\alpha}, \hat{\beta}) = \arg\min_{(\alpha,\beta)} \sum_{i=1}^{N} (Y_i - \alpha - \beta D_i)^2 \left(\frac{D_i}{p_i} + \frac{1 - D_i}{1 - p_i} \right),$$

where p_i are the treatment probabilities or propensity scores (assumed known). Then,

$$\hat{\tau} = \frac{\sum_{i=1}^{n} D_i Y_i / p_i}{\sum_{i=1}^{n} D_i / p_i} - \frac{\sum_{i=1}^{n} (1 - D_i) Y_i / (1 - p_i)}{\sum_{i=1}^{n} (1 - D_i)/(1 - p_i)}.$$

Proof: By elementary calculus, to solve the above program we have to set the derivatives of the objetive (name it f) equal to zero. Upon calling $\gamma_i = \frac{D_i}{p_i} + \frac{(1-D)}{1-p_i}$ we obtain: (using the relations $D_i D_i = D_i$ and $D_i(1 - D_i) = 0$)

$$\frac{\partial f}{\partial \alpha} = -2 \left(\sum_{i=1}^{n} Y_i \gamma_i - \alpha \sum_{i=1}^{n} \gamma_i - \beta \sum_{i=1}^{n} \frac{D_i}{p_i} \right) = 0, \tag{1}$$

$$\frac{\partial f}{\partial \beta} = -2 \left(\sum_{i=1}^{n} Y_i \frac{D_i}{p_i} - \alpha \sum_{i=1}^{n} \frac{D_i}{p_i} - \beta \sum_{i=1}^{n} \frac{D_i}{p_i} \right) = 0. \tag{2}$$

Multiplying (1) by $\sum_{i=1}^{n} \frac{D_i}{p_i}$, (2) by $\sum_{i=1}^{n} \gamma_i$ and taking the difference between the two resulting equations we obtain

$$0 = \sum_{i=1}^{n} \frac{D_i}{p_i} \left(\sum_{i=1}^{n} Y_i \frac{D_i}{p_i} + \sum_{i=1}^{n} Y_i \frac{1 - D_i}{1 - p_i} \right) - \left(\sum_{i=1}^{n} \gamma_i \right) \sum_{i=1}^{n} Y_i \frac{D_i}{p_i} - \beta \left(\left(\sum_{i=1}^{n} \frac{D_i}{p_i} \right)^2 - \sum_{i=1}^{n} \frac{D_i}{p_i} \sum_{i=1}^{n} \gamma_i \right), \tag{3}$$

which after rearrangement of the terms reduces to

$$\hat{\beta}\left(\sum_{i=1}^{n}\frac{1-D_i}{1-p_i}\right)\left(\sum_{i=1}^{n}\frac{D_i}{p_i}\right)=\sum_{i=1}^{n}\frac{1-D_i}{1-p_i}\left(\sum_{i=1}^{n}Y_i\frac{D_i}{p_i}\right)-\sum_{i=1}^{n}\frac{D_i}{p_i}\left(\sum_{i=1}^{n}Y_i\frac{1-D_i}{1-p_i}\right), \quad (4)$$

from which the result follows.

This estimator is similar to the classical "matching by propensity" score one for the ATE (also known as the Horvitz-Thompson),

$$\hat{\tau_{HT}}=\frac{\sum_{i=1}^{n}D_iY_i/p_i}{n}-\frac{\sum_{i=1}^{n}(1-D_i)Y_i/(1-p_i)}{n}.$$

The difference is that Horvitz-Thompson is based only on relations of the type $E(Y^1)=E(Y^1D/p(X))$ while $\hat{\tau}$ also 'uses' the fact that $E(D/p(x))=1$, leading to gains in variance for finite samples at the expense of some bias [AS13] (more on this below).

2 Commentary

At this point one may ask: in there any way to express a causal relation (a 'generative model') so that $\hat{\tau}$ is obtained as the WLS solution to that relation in some sense? This question comes from the observation that the naive estimator $E_n(Y_i|D=1)-E_n(Y_i|D=0)$ is the OLS solution to the set of equations defined by

$$Y=\alpha+\beta D+\epsilon. \quad (5)$$

In the context of a randomized experiment (no possibility of a omitted variable bias) the above estimator will be unbiased for the ATE, in virtue of the Gauss-Markov theorem.

Unfortunately, the above reasoning cannot be easily extended to account for a more general, non-randomized situation. Indeed, in the more general case one may try to state a (tautological) relation as the following,

$$Y=\mu_0+D(\mu_1-\mu_0)+\epsilon_D, \quad \epsilon_D=D\nu^1+(1-D)\nu^0, \quad \mu_i=E(Y^i), \quad \nu^i=Y^i-\mu_i \quad (6)$$

And then appeal to the Gauss-Markov theorem or its generalization, the Aitken's theorem [Ait36] to show that the corresponding WLS estimator is unbiased for $\mu_1-\mu_0$, the ATE. However, this is impossible as $\hat{\tau}$, also known in the literature as the Hajek ratio estimator, although consistent, is biased (by the tightness of Jensen's inequality).
Then, a 'trickier' decomposition would be needed. If that decomposition was possible the term that goes with D in equation (4) must be the expectation of $\hat{\mu}$, and the residual error ϵ_D should be one such that its expected variance given D is equal to y γ_i^{-1}.
That does not look trivial.

References

[Ait36] Alexander C Aitken. Iv.on least squares and linear combination of observations. *Proceedings of the Royal Society of Edinburgh*, 55:42–48, 1936.

[AS13] Peter M Aronow and Cyrus Samii. Estimating average causal effects under interference between units. *arXiv preprint arXiv:1305.6156*, 2013.