Eric Slud   February 17, 2014

Recap & Examples on Influence Functions

RECALL: an **RAL** estimator based on iid missing data from complete data $(R_i, Z_i) \equiv (R_i, X_i, Y_i)$ satisfies

$$\sqrt{n}\,(\hat{\beta} - \beta) \overset{P}{\approx} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi(R_i, Z_i)$$

The function $\varphi(R_i, Z_i)$ is called the **influence function** and for the Regular part of RAL must satisfy the properties

$$\varphi(R, Z) \perp \mathrm{cls}(S_\eta(R, Z)) , \quad E(\varphi(R, Z)S_\beta(R, Z)^{tr}) = I$$

where S_β, S_η are **scores** resp. for the parameters β of interest and the nuisance parameters η defined through

$$\frac{d}{dt} \{\log f(W, (\beta_0, \eta_0) + t\mathbf{v})\}$$

directional derivatives on finite-dim **submodels**, where W denotes the observable data.

  The use of influence functions is that

$$a.var(\hat{\beta}) = E(\varphi(R, Z)\,\varphi(R, Z)^{tr})$$

and that (apart from scaling by constant matrix) the function $\varphi(R, Z, \beta, \eta)$ provides an **estimating equation** for β when $\eta = \hat{\eta}$ is substituted.

Examples with Two-level Missingness

Thm. 10.1 in Tsiatis' book states that if the observable data are $Z = (X, Y) = (X^{(1)}, X^{(2)}, Y)$ when $R = 1$ and a subvector $Z^{(1)}$ when $R = 0$, and if *coarsening at random* condition $P(R = 1 \,|\, Z) = \pi(Z^{(1)})$ holds, then

optimal (smallest variance for fixed φ^F) influence functions are of the *Augmented IPWCC* form

$$\frac{R \, \varphi^F(Z)}{\pi(Z^{(1)})} - \frac{R - \pi(Z^{(1)})}{\pi(Z^{(1)})} \, E(\varphi^F(Z) \,|\, Z^{(1)})$$

In the special case (*restricted moment outcome model*) where $E(Y|X) = \mu(X, \beta)$, the optimal (unscaled) complete-data influence functions φ^F are

$$\varphi^F(Y, X) = A(X)\,(Y - \mu(X, \beta))$$

Optimal $A(X)$ in the case of no missing data is

$$A^*(X) = (-\nabla_\beta \, \mu(X, \beta)) \, (\mathrm{Var}(Y|X))^{-1})$$

Note: in coarsening-model notation `CAR = MAR`, and $\mathcal{C} = \infty \Leftrightarrow R = 1$, $\mathcal{C} = 1 \Leftrightarrow R = 0$.

AIPWCC Examples

(I). The optimal influence function the case of 2-level missingness

$$\frac{R\,\varphi^F(Z)}{\pi(Z^{(1)})} \;-\; \frac{R - \pi(Z^{(1)})}{\pi(Z^{(1)})}\, E(\varphi^F(Z)\,|\,Z^{(1)})$$

is simply $(R/\pi(Z^{(1)}))\,A(X)\,(Y - \mu(X,\beta)$ with no 'Augmented' term when $Z^{(1)} = g(X)$, because

$$E(Y - \mu(X,\beta)\,|\,X,\,R) \;=\; E(Y - \mu(X,\beta)\,|\,X) \;=\; 0$$

This can occur either when missingness is by design or when $\pi(\cdot) = \pi(\cdot,\gamma)$ must be estimated. applicable to surveys e.g. when $Z^{(1)} = X^{(1)}$.

(II). Tsiatis gives the different (biostatistical) example $Z = (Y, X)$, $Z^{(1)} = (Y, X^{(1)})$, and then the term

$$E(\,A(X)\,(Y - \mu(X,\beta)\,|\,Y,\,X^{(1)}\,) \;\neq\; 0$$

Optimal A is generally not be the same $A^*(X)$ as before. Finding it will involve an integral equation because the variance of the influence function involves $A(X)V(X)A(X)^{tr}$ minus another variance of conditional-expectation term involving $A(X)$, as in Robins, Rotnitzky and Zhao (1994).

Further Remarks

(A) The introduction of AIPWCC Estimators in Robins, Rotnitzky and Zhao (1994) was in NMAR-data settings, i.e. situations with **informative missingness**. That is not the situation on (II) above; but there also the *Augmented* estimating equation terms are important.

That will be the topic of the next presentation, on March 3, by Xia Li.

(B) Informative missingness models cannot be nonparametrically tested and identified, because they express joint relationships between variables that are not seen together, such as $(Y, X^{(2)})$ in the survey example in case $R = 0$.

(C). Although the Augmentation term need not and even *should not* be included in MAR settings if efficiency is the only concern, we saw last Fall that estimating equations with the term can be model-robust ('*doubly robust*') to misspecifications of the $\pi(Z^{(1)}, \gamma)$ or $\mu(X, \beta)$ models while the efficient influence function with only the IPWCC term is not. Papers of Z. Tan and later chapters of Tsiatis' book address the relationship of double robustness and augmented estimating equation terms.

Other Topics to present later:

Causal Inference (Rubin 1974) and **Propensity Score Matching** (Rosenbaum and Rubin 1983) are general and important statistical topics which have led to semiparametric statistical methods in econometrics and biostatistics. A paper where these topics enter, with all references in its bibliography, is Hirano, Imbens and Ridder (2003) now posted to the RIT web-page.

Sufficient Dimension Reduction is already mentioned on the RIT web-page.