Eric Slud Feb. 3, 2014

Topics and Papers for Spring '14 RIT

The general topic of the RIT is inference for parameters of interest, such as population means or nonlinear-regression coefficients, in the presence of missing data some of which is designed (as in sample surveys) and some part of which is observational (nonresponse, withdrawal, etc.) and must be modeled.

Methods based on weighting, which adjust the terms of estimating equations by an amount related to the (estimated) probability with which these terms appear, are central to this topic and particularly interesting when methods for sample surveys and mainstream biostatistics cross-fertilise each other.

Simplest example: data $(X_i,\, I_i,\, I_i Y_i)$ for $i \in \mathcal{U}$ where I_i is the *inclusion indicator* in a sample or study with *frame* $\mathcal{U}$ of size N, where $\pi_i = E(I_i)$ is known.

Horvitz-Thompson Estimator $\dfrac{1}{N} \sum\limits_{i \in \mathcal{U}} I_i\, Y_i/\pi_i$

for the population average μ_Y is classical in survey sampling, but may be considered as the solution of an estimating equation

$$\sum_{i \in \mathcal{U}} \frac{1}{\pi_i}\, I_i\, (Y_i - \mu_Y)$$

in an independent-data setting.

More generally:

we are interested in estimating μ_Y and other parameters in settings with predictor-variables $\mathbf{X}_i = (\mathbf{X}_i^{(1)},\ \mathbf{X}_i^{(2)},\ \ldots)$ and outcome-variables Y_i observable only when successive-stage non-missing indicators $I_i,\ A_i,\ r_i,\ \ldots$ are 1.

Typically there are constituent models

$$\texttt{outcome model} \qquad E(Y \mid \mathbf{X}) \ = \ \mu(\mathbf{X}, \beta)$$

$$\texttt{propensity model} \qquad E(r \mid \mathbf{X},\ Y) \ = \ q(\mathbf{X}, \gamma)$$

which may be given in multiple stages for indicators $I,\ A,$ etc. and may depend on different predictor-subvectors $\mathbf{X}$.

The formulas and topics sketched here restrict to the *Missing at Random (MAR)* or 'noninformative' case where propensities conditioned on predictors and outcomes. But many authors attempt to treat the more general `NMAR` case in spite of its identifiability problems.

An example of the kind of estimating equations we arrive at in estimating β 'efficiently' when the indicators I_i have known expectations is the weighted-regression type estimating equation

$$\sum_{i \in \mathcal{U}} \ I_i \ \frac{1}{\pi_i} \ A(\mathbf{X}_i) \ (Y_i \ - \ \mu(\mathbf{X}_i, \beta)) \ = \ 0$$

Overall Questions are:

• How to estimate parameters like $E(Y)$ and β as efficiently as possible from estimating equations in the presence of parametric or `semi`parametric models.

• How to present the estimators so that they carry over to the sample-survey setting, where inclusion indicators I_i need not be independent and almost no conditions other than a few restrictions on large-sample averages of $\mathbf{X}$'s and Y's can be made, but where propensity models are *unavoidable*.

Methods & Background

Estimators coming from estimating equations are the `Regular Asymptotically Linear` estimators such that as sample-size $n = \sum_i I_i A_i$ gets large

$$\sqrt{n}\,(\hat{\theta} - \theta) \stackrel{P}{\approx} \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{U}} \varphi(Y_i, \mathbf{X}_i, I_i, r_i, A_i)$$

In iid setting, `semiparametric` techniques for providing classes of such functions φ in which to search for efficient estimators even though *nuisance parameters* are infinite-dimensional are described in detail in the book

Tsiatis, A. (2006), "Semiparametric Theory and Missing Data", Springer.

We read parts of that book last year and (after recap) should probably do a couple more chapters this semester (especially Chapters 4 and 7).

Some papers I want to get to are those of Zhiqiang Tan (see the link to his web-page from the RIT web-page).

URL: `http://www.math.umd.edu/ evs/RITF13.html`

contains a reading list from last semester. My overall goal is to understand as much as possible about optimality of weighted-estimating equation estimators in the setting of iid data with missingness both designed and observational, with the ultimate idea of formulating good new estimators as well as suitable optimality criteria in the survey-sampling case.

Material to read from original papers includes:

- the original 1994 JASA paper of Robins, Rotnitzky and Zhao (on RIT web-page) which introduced `Augmented Inverse Probability Weighted Estimating Equations`.

- (maybe) the 2012 paper (Ma and Zhu, JASA) where semiparametric techniques were brought into the area of 'sufficient dimension reduction'.

- (maybe) one of the papers on Composite Likelihoods (mentioned in the announcement email, to be posted to the web-page), which is a fruitful technique for deriving new estimating equations in difficult problems.

- *Empirical Likelihood* is an important method of writing estimating equations which readily incorporate constraints, and which are efficient in some simple problems and also useful in some very hard ones.

• Survey-weighted *Calibration* estimators were discussed last term as a way to embody constraints from known survey totals. Weight adjustment subject to penalty functions is another strand in survey research. Papers in these areas can also help to motivate techniques in statistics with independent data.

• Another set of methods which carries over between independent and survey data are those based on `Imputation`, which broadly means methods by which missing data are filled in (once or maybe many times) by some simulation or resampling scheme and the filled-in data are then analyzed (once or more) by complete-data estimation algorithms. Material in this direction could be found in the book of Rubin (and various papers of Little and Rubin and many other authors) on *Multiple Imputation*, later papers (such as a Biometrika paper of Robins and Wang around the year 2000) discussing conditions under which related methods are consistent, and also techniques from survey literature on hot (or cold or random) deck imputation. These methods are also closely related to the EM algorithm.

• Papers on `two-phase sampling` in biostatistics (in which a vector of outcomes Y_i and inexpensive predictors $\mathbf{X}_i^{(1)}$ are observed on a large cohort of subjects and a more accurate and expensive vector $\mathbf{X}_i^{(2)}$ of predictors is observed on a subsample of the cohort, with the binary indicator r_i showing which subjects are subsampled) form an important source of examples for the semiparametric theory in an iid setting, and Tsiatis' (2006) book refers

to a stream of papers by Breslow and co-authors on this topic.

Pointers to papers on these topics are highlighted with red asterisks * in the reading list on the RIT web-page.