Smooth Post-Stratification for Capture-Recapture

Zachary Kurtz (zkurtz@stat.cmu.edu)
Department of Statistics, Carnegie Mellon University

Problem

Given several incomplete lists of population units, how can you estimate the number of units missed by all of the lists? Examples:

- Census coverage accuracy, given a post-enumeration survey + IRS data.
- The number of bird species in a region based on several annual surveys.
- The number of errors in a body of computer code, after several reviewers compile error lists.

Some intuition: Peterson estimates

		List L_1 yes	List L_1 no
List L_2	yes	c_{11}	c_{01}
	no	c_{10}	c_{00}

Under list independence, the odds ratio is one, giving the Petersen estimator.

$$\hat{c}_{00} = \frac{c_{10}\, c_{01}}{c_{11}}$$

Traditional log-linear model

With three lists, there are 8 possible capture patterns. A generalization of the Petersen estimator comes from assuming that any two layers of the 2x2x2 classification array have the same odds ratio:

$$1 = \frac{c_{111}\, c_{010}\, c_{100}\, c_{001}}{\hat{c}_{000}\, c_{110}\, c_{101}\, c_{011}}$$

Alternatively, this can be derived as the MLE for a saturated log-linear model.

Let $\mathbf{y}$ denote the capture pattern for an arbitrary unit. Hence $\mathbf{y}_j = 1$ if the unit is on list j, and 0 otherwise. Let $p(\mathbf{y}) := P(Y = y)$, the probability that a random unit has capture pattern $\mathbf{y}$. A saturated log-linear model is then

$$p(\mathbf{y}) = u + u_1\mathbf{y}_1 + u_2\mathbf{y}_2 + u_3\mathbf{y}_3 + u_{12}\mathbf{y}_1\mathbf{y}_2 + u_{13}\mathbf{y}_1\mathbf{y}_3 + u_{23}\mathbf{y}_2\mathbf{y}_3$$

Here the u-terms are parameters to be estimated (i.e., maximize the multinomial likelihood).

Big Idea

A **problematic assumption** is that of homogeneity:

(A1) Multinomial sampling probabilities are constant across units.

Heterogeneity (when A1 fails) has two facets:

- *A priori* capture probabilities may vary across units.
- List interactions may vary across units.

Heterogeneous capture behavior may be explained by a unit-level covariate x.

Post-stratification is a crude way to control for x, by dividing the observed population into S different post-strata according to some partition of the covariate space. Imputing the number of missing units on each post-stratum separately, followed by aggregation, can reduce heterogeneity-induced bias.

Big Idea: Find a smooth (unit-level) generalization of post-stratification.

Other basic assumptions:
(A2) Units can be cross-tabulated across lists.
(A3) Population units act independently of one another.
(A4) The missing data is missing at random (MAR).
(A5) The population is closed (no births, deaths, or migration)

Smooth post-stratification

Let $\pi(\mathbf{y}, x) = P(\mathbf{Y} = \mathbf{y}|x)/P(\mathbf{Y} \neq 0|x)$.

(1) Find $\pi(\mathbf{y}, x)$ for $\mathbf{y} \neq 0$ using a nonparametric conditional density estimator.

(2) Impute $\pi(\mathbf{0}, x_i)$ using a log-linear model.

(3) Apply Horvitz-Thompson (reformatted):

$$\hat{c}_{\mathbf{0}} = \sum_i \hat{\pi}(\mathbf{0}, x_i)$$

where $c_{\mathbf{0}}$ is the number of missing units and i ranges over all of the observed units.

Example: Prevalence of multiple sclerosis in the Lorraine region, France

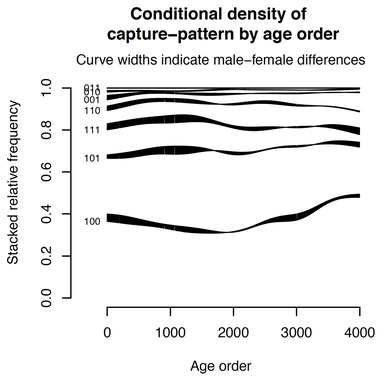
Three data sources (LR, RHIS, MRD) included a total of 4001 people diagnosed with multiple sclerosis in the Lorraine region of France:

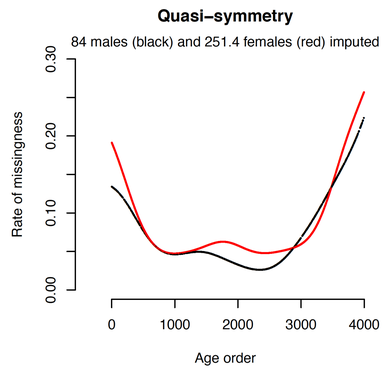
		In LR	Not in LR
In RHIS	In MRD	474	42
	Not in MRD	1342	199
Not in RHIS	In MRD	393	64
	Not in MRD	1486	c_0

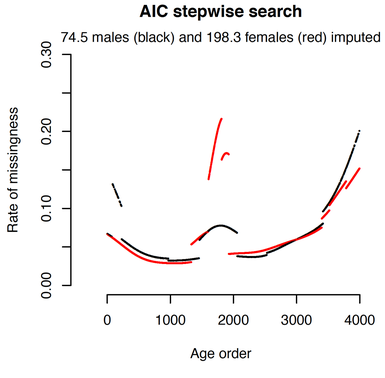
Smooth post-stratification on age and sex:
We compute nonparametric density estimates for the observable capture patterns as a function of age and sex, and fit log-linear models *locally*.

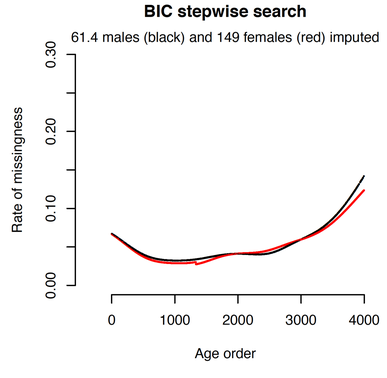
Define the *age order* from youngest to oldest, so age order = 1 for the youngest subject, and age order = 4001 for the oldest subject.

The npcdens command in the np package in R uses cross-validation for bandwidth selection in a nonparametric conditional density estimator.









Traditional analysis: Adssi et. al. (2012) selected a log-linear model by AIC:

Model	n (95% confidence interval)	AIC
Independent sources	4197.2 (4173.9–4223.8)	13.20
Interaction RHIS* MRD	4206.0 (4181.0–4234.5)	6.89
Interaction MRD* LR	4208.9 (4179.5–4243.3)	13.55
Interaction RHIS* LR	4214.2 (4166.0–4276.4)	14.69
Interaction (RHIS*MRD) and (MRD*LR)	4221.2 (4189.1–4258.7)	6.59
Interaction (RHIS*MRD) and (RHIS*LR)	4242.9 (4185.9–4317.7)	7.10
Interaction (MRD*LR) and (RHIS*LR)	4304.2 (4201.6–4459.4)	11.76
Interaction (RHIS*MRD), (MRD*LR) and (RHIS*LR)[a]	4405.7 (4261.5–4629.7)	0

Local log-linear modeling:
Let x denote the covariate vector (age order, sex). For each subject, we select a submodel of

$$\pi(\mathbf{y}, x) = u(x) + u_1(x)\mathbf{y}_1 + u_2(x)\mathbf{y}_2 + u_3(x)\mathbf{y}_3 + u_{12}(x)\mathbf{y}_1\mathbf{y}_2 + u_{23}(x)\mathbf{y}_2\mathbf{y}_3$$

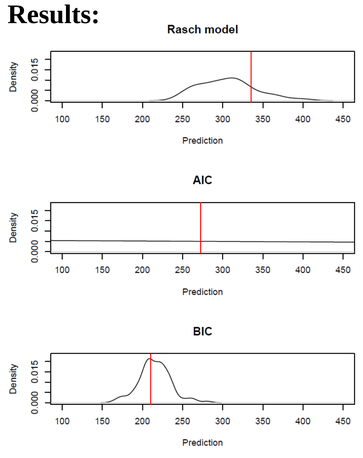
For the ith subject, the *local* data is the cross-classification array

$$C_i = \{\hat{\pi}(\mathbf{y}, x_i)\}_{\mathbf{y}\neq 0}$$

As an approximation, the scaled up and rounded array $\texttt{round}(\lambda C_i)$ may be treated as multinomial data, facilitating parameter estimation.

Suitable modifications of the AIC and BIC are relevant for model selection.

Results:



Acknowledgements

The author thanks William F. Eddy (adviser), Stephen Fienberg, Cosma Shalizi, and Rebecca Steorts for providing valuable suggestions. This work was supported in part by NSF grant SES1130706.