“Hard” versus “Soft” Predictions from Unit-level Models for Small Area Estimation of Proportions

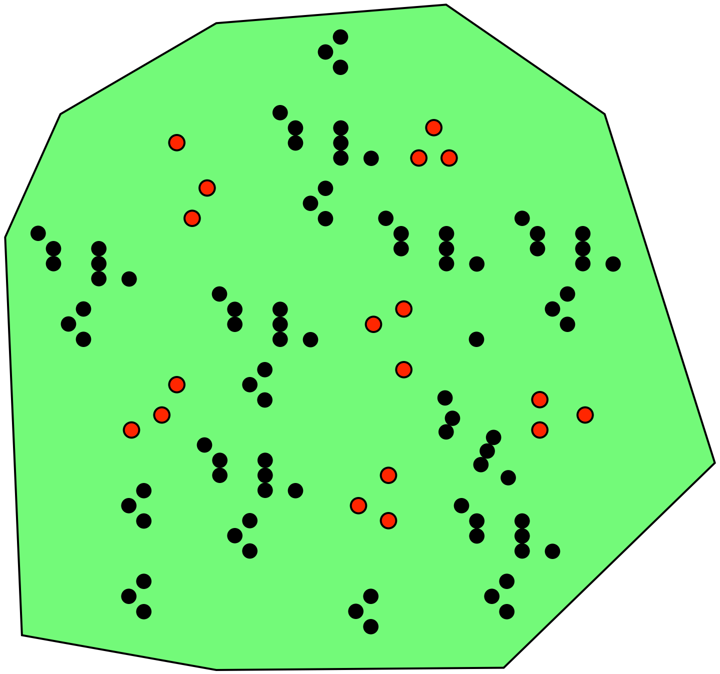
Geoff Jones, Penny Bilton

Massey University, NZ

ELL Method for Poverty Mapping

Survey
Y, X

Expenditure pp
Kcal pae
Height-for-age
Weight-for-age



Census
X

Regression

$$Y_{ij} = X_{ij}\beta + h_i + e_{ij}$$

Elbers C., Lanjouw J. and Lanjouw P. (2003) Micro-level estimation of poverty and inequality, *Econometrica* 71, 355-364.

Missing Data Problem?

ic	ucode	tcode	div	lnexp	urban	num_hh	f_hhh	electric	aglnd
1	40905	409	1	7.1994	1	8	0	1	0
2	40905	409	1		1	4	0	1	0
3	40905	409	1		1	3	1		
4	40905	409	1		1	4	0	1	0
5	40905	409	1		1	4	0	0	0
6	40905	409	1		1	4	0	0	0
7	40905	409	1	6.6678	1	4	0	0	0
8	40905	409	1		1	5	0	0	0
9	40905	409	1		1	3	0	0	0
10	40905	409	1		1			0	0
11	40905	409	1	6.0834	1	7	0	1	1
12	40905	409	1		1	7	1	1	0
13	40905	409	1		1	3	0	0	1
14	40905	409	1		1	6	0	0	1
15	40905	409	1		1	1	0	1	1
16	40905	409	1						
17	40905	409	1		1	5	0	1	0
18	40905	409	1		1	3	0	1	0
19	40905	409	1	6.2621	1	4	0	1	0
...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...
97	40905	409	1	6.2838	1	5	0	1	0
98	40905	409	1		1	3	1	0	0
99	40909	409	1		1	5	0	0	1
100	40909	409	1	6.2901	1	7	0	0	1

ELL: $$Y_{ij}^{b} = X_{ij}\beta^{b} + h_{i}^{b} + e_{ij}^{b}$$

Multiple Imputation and Aggregation

$$SAE_R^b = \frac{1}{|R|} \sum_{ij \in R} \left(Y_{ij}^b < z \right)$$

then

$$SAE_R = \underset{b}{\text{mean}} \left(SAE_R^b \right)$$

$$\text{se}\left(SAE_R \right) = \underset{b}{\text{sd}} \left(SAE_R^b \right)$$

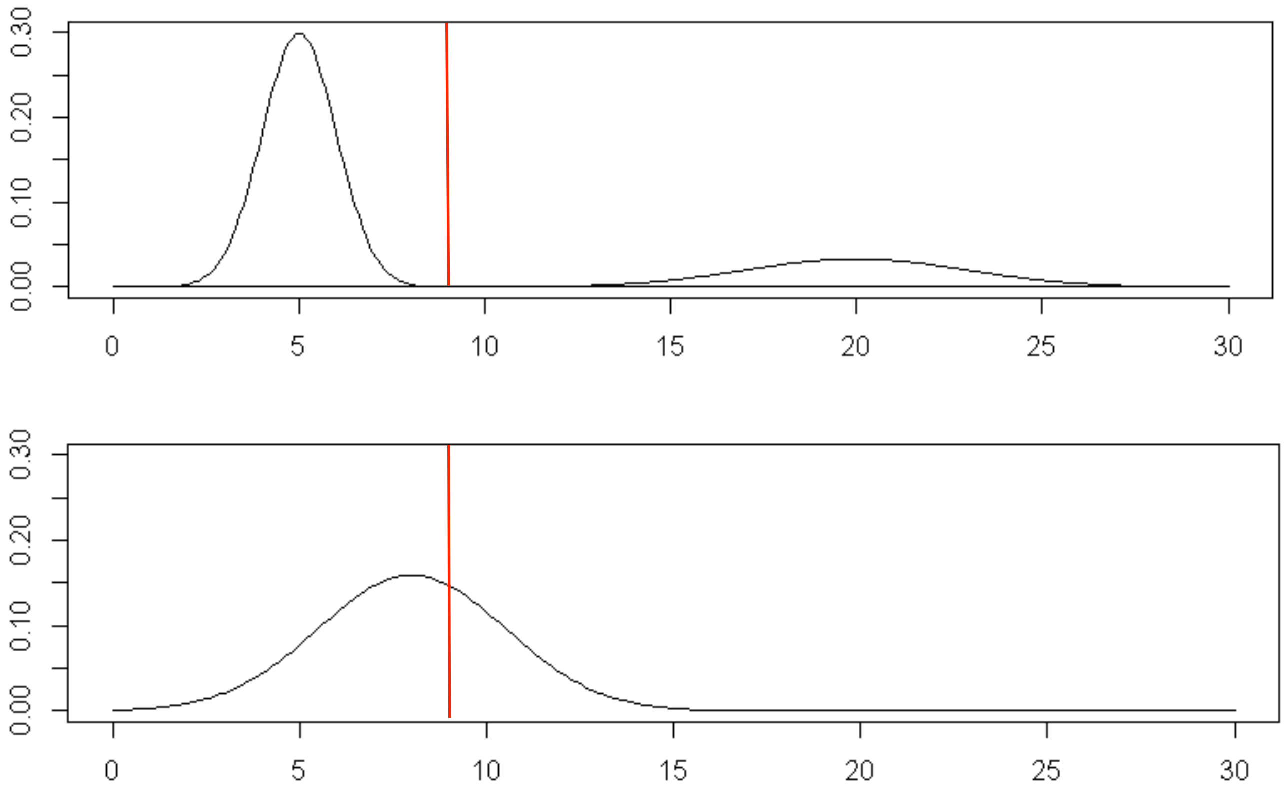
Population / Superpopulation?

Estimate / Predict / Impute?

Linearity and Nonlinearity

$$\hat{Y}_{ij} = X_{ij}\hat{\beta}$$

$$SAE_R = \frac{1}{|R|}\sum_{ij \in R}\left(\hat{Y}_{ij} < z\right)$$



General Framework

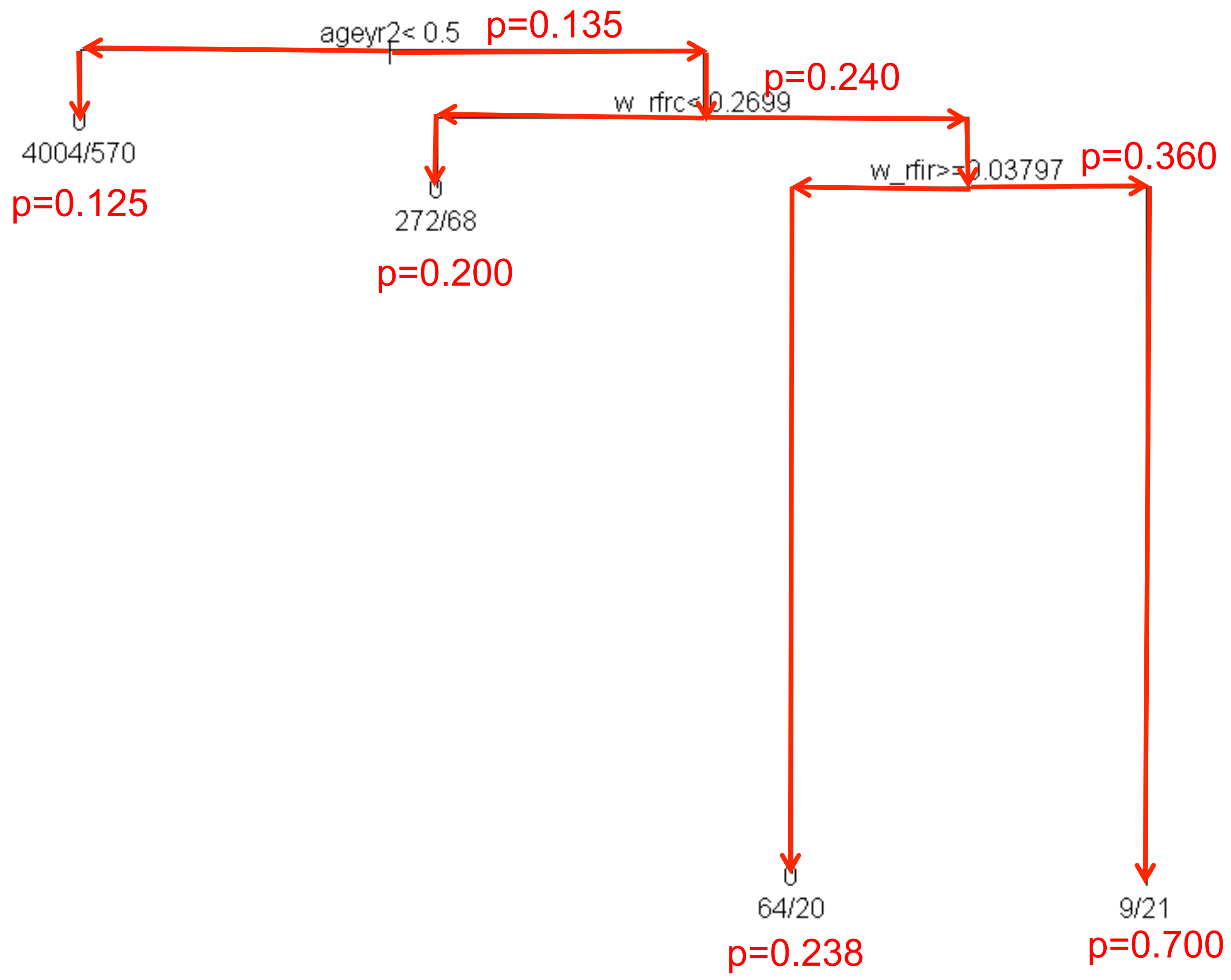
- Denote the full census data by $\mathbf{C}$;
- Denote the area-level quantities of interest by $\varphi_a(\mathbf{C})$;
- φ_a operates on rows of census data (households or children) to produce values that are then aggregated to area level;
- Part of the census data is unobserved; write $\mathbf{C} = \mathbf{C}_o + \mathbf{C}_u$;
- Assume that $\mathbf{C}_u$ is "like" $\mathbf{C}_o$ in some sense;
- This "likeness" (a "model") is used to infer $\mathbf{C}_u^*$;

  - in sae, this model is usually explicit: $\mathbf{C}_u^* = E[\mathbf{C}_u \mid X, \mathbf{C}_o]$

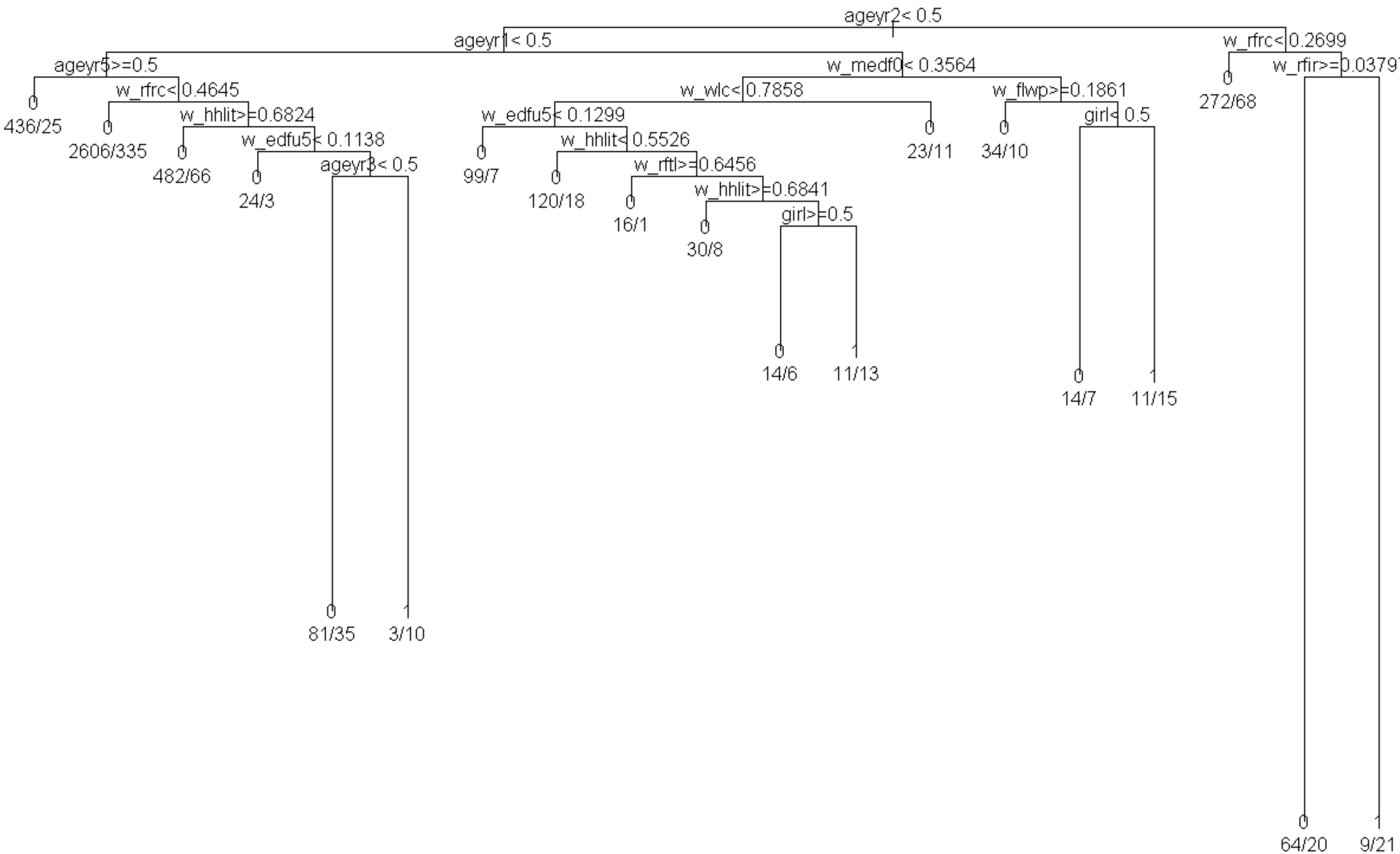
  The estimate of the area-level summary is then:

$$\hat{\varphi}_a = \varphi_a(\mathbf{C}_o + \mathbf{C}_u^*)$$

A Diarrhoea Tree (Nepal DHS 2011)



Another Diarrhoea Tree



Prediction on Census Data

Condition

Node i   0 / 1 p_i

N_i

For a given small area, N_i units emerge at node i

"Hard"

$$SAE_R^h = \frac{1}{N} \sum_i N_i \left(p_i > 0.5 \right)$$

"Soft"

$$SAE_R^s = \frac{1}{N} \sum_i N_i p_i$$

Or, Maybe ...

"Hard-ish"

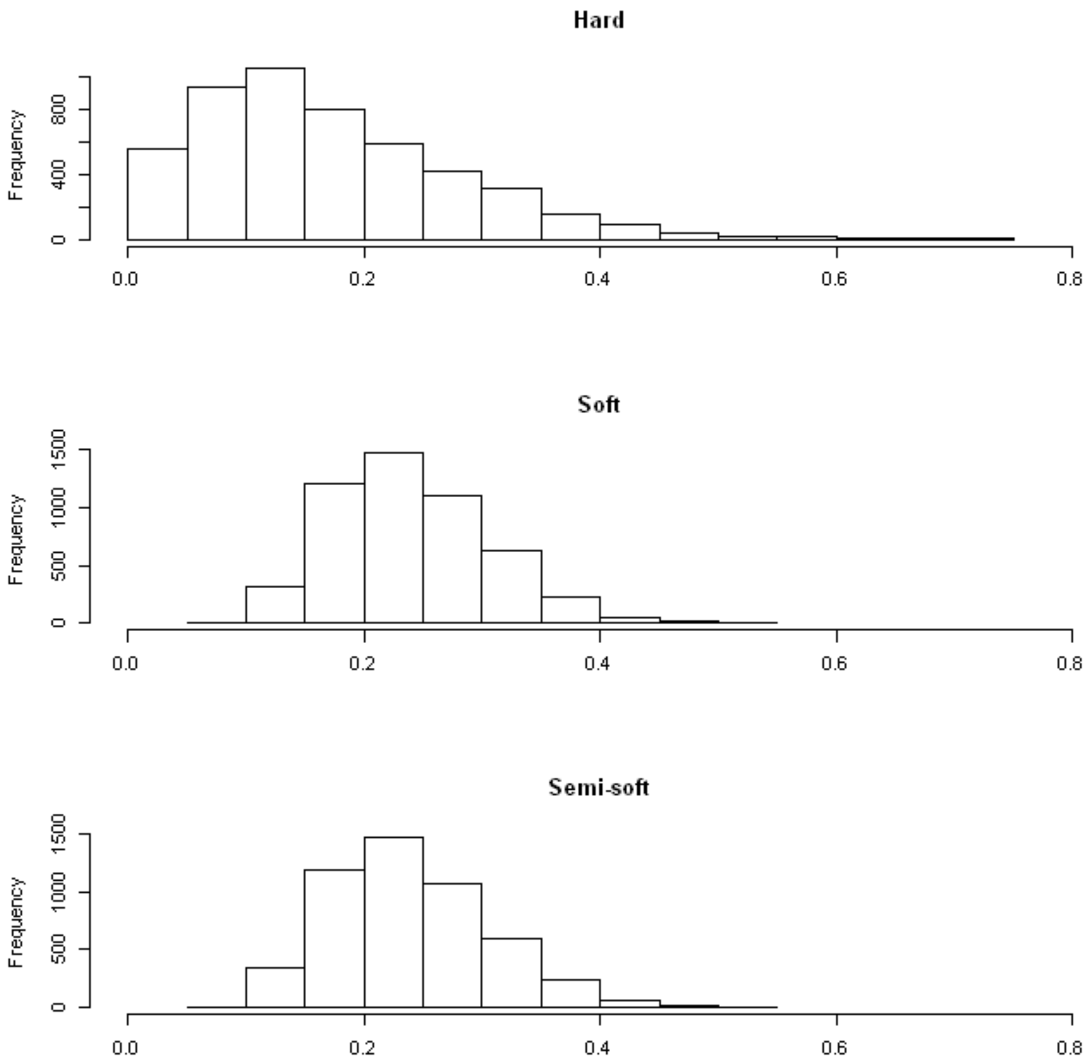
$$SAE_R^{hs} = \frac{1}{N}\sum_i \sum_j X_{ij} \quad \text{where } X_{ij} \sim \text{Bern}(p_i)$$

Note:

$$E\left[SAE_R^{hs} \,|\, p\right] = \frac{1}{N}\sum_i \sum_j p_i = SAE_R^s$$

$$V\left[SAE_R^{hs} \,|\, p\right] = \frac{1}{N^2}\sum_i N_i (1 - p_i) p_i$$

Bootstrapped Trees



Future Work

Adapt tree-fitting and standard error estimation for survey design:

- weights;
- clustering;
- strata.

Any questions?