Fitting Regression Models to Survey Data - Discussion Points

Notation

- Draw a sample S of n units from a finite population of N units.
- Observations $\{(y_i, x_i); i \in S\}$.
- Associated weights $\{w_i; i \in S\}$,
  - (Broadly) weights indicate number of population units represented by the i^{th} sample.
  - In some cases, $w_i = \pi_i^{-1}$, where π_i is the probability of selecting the i^{th} unit.
- R_i: Indicator that unit i in the population was sampled.
- E_π: expectation over finite population sampling
- E_p: expectation over the model
- β_0: true parameter value in the superpopulation model
- $\tilde{\beta}_N$: maximum quasi-likelihood estimator of β_0 that would be obtained from the full population data.
- β^*: the limit in probability of $\tilde{\beta}_N$
- $\hat{\beta}_n$ the maximum pseudo-likelihood estimator obtained from sample

Question: why are $\tilde{\beta}_N$ and β_0 not the same?

Question: what is the superpopulation?

Assumptions

- Sample population is a realization of the superpopulation probability model with density $f(Y|X;\beta)$
- $g(E[Y|X=x]) = g(\mu) = x'\beta \iff \mu = g^{-1}(x'\beta)$
- $var[Y|X=x] = \sigma^2 V(\mu)$

Interesting Note: the paper claims 'when the primary interest is in the marginal regressions there does not seem to be any important loss of generality in treating the population as an i.i.d. sample'.

Pseudo-Likelihood Estimation

- Approach underlying regression modeling in all major statistical software packages for survey analysis
- Solve the following estimating (score) equations:
  - $\sum_{i=1}^N R_i w_i U_i(\beta) = \sum_{i=1}^N R_i w_i x_i \frac{1}{g'(\mu_i)V(\mu_i)}(y_i - \mu_i(\beta)) = 0$
  - unbiased estimating equations for $\tilde{\beta}_N$ if $E_\pi[w_i R_i] = 1$
  - $\hat{\beta}_n$ asymptotically normal and consistent for β_0 when the superpopulation model is correctly specified
- $Var(\hat{\beta}_n)$ is the sum of two components
  - Finite population sampling variance of $\hat{\beta}_n$ around $\tilde{\beta}_N$ of order n^{-1}.
  - Model based sampling variance of $\tilde{\beta}_N$ around β_0 of order N^{-1}.
  - Model based sampling variance often ignored when N is much larger than n.
  - Usual sandwich estimator: $A^{-1}BA^{-1}$
    - $A = \sum_{i=1}^N w_i R_i \frac{\partial U_i(\beta)}{\partial \beta}|_{\beta=\hat{\beta}_n}$
    - $B = \hat{v}ar_\pi(\sum_{i=1}^N w_i R_i U_i(\beta))$
- Wald and score tests possible

Use of weights

- Need for using weights depends on type of sampling
  - exogenous: R is independent of Y conditional on predictor variables X
  - endogenous: R is not conditionally independent of Y
- If sampling is exogenous, the non-sampled fraction of the population is MAR.
  - Use of sampling weights may not be needed
  - See paper for additional details
- For exogenous sampling, the non-sampled fraction is NMAR and the weights (apparently) should be used
  - Not much discussion of this in the paper

Working Likelihood/Maximum Likelihood

- In general, no straightforward likelihood function for survey data
- Standard analytic techniques (above) use estimating equations and do not rely on likelihood based approaches
- For hypothesis testing, still possible to construct an analogue of the likelihood ratio test based on the pseudo-likelihood
  - $\hat{l}(\beta) = \sum_{i=1}^{N} w_i R_i log(f(y|x;\beta))$
  - $\Lambda = 2(\hat{l}(\hat{\beta}) - \hat{l}(\hat{\beta}_0))$
- Case-control studies are notable exceptions where full likelihood methods are still possible.