Overview of Stat MAPS-REU Summer 2014 Project

Statistical Estimation via Calibration

Eric V. Slud, Univ. of Maryland, Mathematics Dept.

June 9, 2014

Data Setting

Data are $\{\mathbf{X}_i, Y_i \,:\, i \in S\}$, $\mathbf{X}_i \in \mathbf{1} \times \mathbf{R}^{p-1}$, $Y_i \in \mathbf{R}$

$S \subset U$ is a probability sample drawn from pop'n list U with known inclusion probabilities π_i, and *design weights* $w_i^o = 1/\pi_i$

Objective: to estimate unknown total $t_Y = \sum_{i \in U} Y_i$ by an estimator $\sum_{i \in S} w_i Y_i$ which is unbiased (nearly, in large samples), with variance as small as possible

where *final weights* w_i are modified from w_i^o using relationships between $\mathbf{X}_i$ for $i \in S$ and known population totals $t^*_{\mathbf{X}} = \sum_{i \in U} \mathbf{X}_i$

Idea of Survey Calibration

To improve on *Horvitz-Thompson Estimator* $\hat{t}_Y^{HT} = \sum_{i \in S} w_i^o \, Y_i$

Estimate $t_Y = \sum_{i \in U} Y_i$ by $\hat{t}_Y = \sum_{i \in S} w_i \, Y_i$

where $\{w_i : \; R_i = 1\}$ minimizes Loss $= \sum_{i \in S} (w_i - w_i^o)^2 \, / w_i^o$
subject to **calibration constraints** $\sum_{i \in S} w_i \, \mathbf{X}_i = t^*_{\mathbf{X}}$.

Equivalent to Generalized Regression Estimators for Y on **X**.

Survey (Weighted Least Squares) Regression

Pop'n-Level Least Squares Equation $\sum_U \mathbf{X}_i \, (Y_i - \mathbf{X}_i' \beta) \;=\; \mathbf{0}$

Estimated and solved in sample as: $\sum_{i \in S} w_i^o \, \mathbf{X}_i \, (Y_i - \mathbf{X}_i' \beta) \;=\; \mathbf{0}$

Regression survey estimator

$$\hat{t}_Y \;=\; (\textstyle\sum_{i \in U} \mathbf{X}_i)' \, \widehat{\beta} + \sum_{i \in S} w_i^o \, (Y_i - X_i' \widehat{\beta}) \;=\; t_{\mathbf{X}}^{*\,\prime} \, \widehat{\beta}$$

requires knowledge of population totals $t_{\mathbf{X}}^*$.

Here $\widehat{\beta} \;=\; \left(\sum_{i \in S} w_i^o \, \mathbf{X}_i \, \mathbf{X}_i' \right)^{-1} \sum_{i \in S} w_i^o \, \mathbf{X}_i \, Y_i$

Folklore: (Fuller 2009) Var$(\hat{t}_Y)$ *is large when* dim$(\mathbf{X}_i)$ *is.*

Why Not Linear Regression Using all p Predictors?

- p is often large, and corresponding weights w_i from using all will vary too much: an **overfitting** problem.

- not all of the elements $t^*_{\mathbf{X}}$ are known to high accuracy

- the practical requirement to equate $\sum_{i \in S} w_i \mathbf{X}_i = t^*_{\mathbf{X}}$ is not strong in all entries, only in a relative few.

Variable Selection in Regression

In ordinary least squares regression this is a classic problem when p is large but $\ll n = |S|$. Some approaches, translated to survey notation:

Ridge Regression: $\min_{\beta \in \mathbf{R}^p} \left\{ \sum_{i \in S} w_i^o \, (Y_i - \mathbf{X}_i' \, \beta)^2 + \lambda \|\beta\|_2^2 \right\}$

Mallows C_p **:** $\min_{dim(\tilde{X}_i) = q \leq p} \left\{ \frac{\sum_S w_i^o (Y_i - \tilde{\beta}' \tilde{X}_i)^2}{\sum_S w_i^o (Y_i - \widehat{\beta}' X_i)^2} - n + 2q \right\}$

LASSO : $\min_{\beta \in \mathbf{R}^p} \left\{ \sum_{i \in S} w_i^o \, (Y_i - \mathbf{X}_i' \, \beta)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\}$

Sequential forward and backward variable selection (greedy) algorithms often used to simply the combinatorics of search.

Quadratic Programming Optimization Approach

Explicitly subdivide the $n \times p$ design matrix $\mathbf{X}$ with rows $X_i, i \in S$ into $n \times p_k$ blocks, $\mathbf{X}^{(k)}$, $k = 0, \ldots, K$ for which different calibration accuracies are appropriate, with $\sum_{k=0}^{K} p_k = p$, and

$$\min_{\mathbf{w}} \left\{ \sum_{i \in S} \frac{(w_i - w_i^o)^2}{2 w_i^o} + \sum_{k=1}^{K} a_k \, \| \sum_{S} w_i \mathbf{X}_i^{(k)} - t^*_{\mathbf{X}^{(k)}} \|_2^2 \right\}$$

subject to $\quad \sum_{i \in S} w_i \, \mathbf{X}_i^{(0)} = t^*_{\mathbf{X}^{(0)}} \quad$ and $\quad c_1 \leq w_i / w_i^o \leq c_2$

Research Problems for the Project

• What is the interplay between dimension of $q \le p$ of selected set of regressors required to include a specified set with $n \times p_0$ design matrix $\mathbf{X}^{(0)}$?

• Study the behavior of variance of survey estimators based on these variable selection ideas on simulated survey datasets (generated via pseudo-random generators from known distributions), obtaining formulas and bounds where possible.

• Consider and devise new variable selection strategies for the survey setting.

• Analyze real survey data from public-use survey data files using these ideas.