An Overview of the Pros and Cons of Linearization versus Replication in Establishment Surveys

Richard Valliant

University of Michigan
and
Joint Program in Survey Methodology
University of Maryland

Introduction

- Nonlinear estimators are rule—not exception—in survey estimation
- Means: ratios of estimated totals
- Totals: nonlinear due to nonresponse adjustments, poststratification, other calibration estimation

More Complex Examples

- Price indexes

  Long-term index = product of short-term indexes across time periods

  Each short-term index may be ratio of long-term indexes
- Regression parameter estimates from X-sectional survey
- Autoregressive parameter estimates from panel survey
- Time series models with trend, seasonal, irregular terms

Options for Variance Estimation

- Linearization

  - Standard linearization

  - Jackknife linearization

- Replication

  - Jackknife

  - BRR

  - Bootstrap

Examples of Establishment Survey Designs

- Stratified, single-stage (often equal probability within strata)
  - Current Employment Statistics (US)
  - Occupational Employment Statistics (US)
  - Business Payrolls Survey (Canada)
  - Survey of Manufacturing (Canada)

- Stratified, two-stage
  - Consumer Price Index (US); geographic PSUs
  - National Compensation Survey (US); geographic PSUs
  - Occupational Safety and Health Statistics (US); establishments are PSUs/injury cases sampled within

JPSM The Joint Program in Survey Methodology

Goals of Variance Estimation

- Construct confidence intervals to make inference about pop parameters
- Estimate variance components for survey design
- Desiderata
  - Design consistent under a design close to what was actually used
  - Model consistent under model that motivates the point estimator
  - Easy application to derived estimates (differences or ratios in domain means, interquartile ranges)

JPSM The Joint Program in Survey Methodology

Example

- Ratio estimator in srs

$$\hat{T}_R = N\bar{X}\hat{\beta},\ \hat{\beta} = \frac{\bar{y}_s}{\bar{x}_s}.$$ Motivated by model

$$E_M\left(y_k\right) = \beta x_k;\ \text{var}_M\left(y_k\right) = \sigma^2 x_k$$

- Design-consistent but not model-consistent estimator:

$$v_L = \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\frac{\sum_s r_k^2}{n-1},\ r_k = y_k - x_k\hat{\beta}$$

- Both design-consistent and model-consistent:

$$v_2 = \left(\frac{\bar{X}}{\bar{x}_s}\right)^2 v_L$$

JPSM The Joint Program in Survey Methodology

- Frequentist approach
  - Objections: piecemeal, every problem needs a new solution
  - Bootstrap is more general, frequentist solution for some problems—generate entire distribution of statistic

    Generate pseudo-population:

    Booth, Butler, & Hall, *JASA* (1994)

    Canty and Davison, *The Statistician* (1999)

    More specialized bootstraps:

    Rao & Wu *JASA* (1988)

    Langlet, Faucher & Lesage, *Proc JSM* (2003)

- (One) Bayesian solution
  - Generate entire posterior of population parameter; use to estimate mean, intervals for parameter, etc
  - Polya posterior: Ghosh & Meeden, *Bayesian Methods for Finite Pop Sampling* (1997)
  - Unknown in practice but theoretically interesting; not available for clustered pops

Practical Issues/Work-arounds

- How to reflect weight adjustments in variance estimates
  - Unknown eligibility
  - Nonresponse
  - Use of auxiliary data (calibration)
- Linearization: some steps often ignored (like NR adjustment)
- Replication: Units combined into groups for jackknife, other methods
  - Loss of degrees of freedom; poor combinations can inject bias
- Item Imputation: special procedures needed

More Practical Issues / Work-arounds

- Design compromise: assume 1st stage units selected with replacement
  - Without replacement theory possible but not always practical
  - Joint selection probabilities not tracked or uncomputable in many (most?) designs
- Adaptive procedures
  - Cell collapsing in PS, NR
  - Weight censoring

JPSM The Joint Program in Survey Methodology

- Some designs do not permit design-unbiased or consistent variance estimates
  - Systematic sampling from an ordered list
  - Standard practice is PISE (pretend it's something else)
- Replication estimators are often both model consistent (assuming independent 1st stage units) and design consistent (assuming with-replacement sampling of 1st stage units)

Basic Linearization Method

- Write linear approx to statistic; compute (design or model) variance of approx

$$\hat{\theta}-\theta = g\left(\hat{t}_1,\hat{t}_2,\ldots,\hat{t}_p\right)-\theta$$

$$\cong \sum_{j=1}^{p}\left(\frac{\partial g}{\partial \hat{t}_j}\Big|_{\hat{\mathbf{t}}=\mathbf{t}}\right)\left(\hat{t}_j - t_j\right)$$

Writing $\hat{t}_j = \sum_h \sum_{i \in s_h} \hat{t}_{jhi}$ and reversing PSU (i) and variable (j) sums and noting t_j is constant:

$$\mathrm{var}\left(\hat{\theta}-\theta\right) \cong \mathrm{var}\left[\sum_{h,i\in s_h}\sum_{j=1}^{p}\left(\frac{\partial g}{\partial \hat{t}_j}\Big|_{\hat{\mathbf{t}}=\mathbf{t}}\right)\hat{t}_{jhi}\right]$$

JPSM The Joint Program in Survey Methodology

- The variance can be w.r.t. a model or design
- Assuming units in different strata are independent (model) or sampling is independent from stratum to stratum (design):

$$\text{var}\left(\hat{\theta}-\theta\right) \cong \sum_h \text{var}\left(\sum_{i\in s_h} u_i\right)$$

$$u_i = \sum_{j=1}^{p} \left(\frac{\partial g}{\partial \hat{t}_j} \bigg|_{\hat{\mathbf{t}}=\mathbf{t}} \right) \hat{t}_{jhi}$$

(***linear substitute*** method)

Variance is computed under whatever design or model is appropriate. Assumption of *with-replacement* selection of PSUs not necessary but often used for design-based calculation.

- Issue of evaluating partial derivatives
  - when to substitute estimates for unknown quantities?
- Binder, *Surv Meth* (1996)
  - Take total differential of statistic
  - Evaluate derivatives at sample estimates where needed
  - Leads to variance estimators with better conditional (model) properties

JPSM The Joint Program in Survey Methodology

Example

- More general ratio estimator: $\hat{t}_R = \frac{t_x}{\hat{t}_x}\hat{t}_y$

  Evaluating partials at pop values gives "standard" linearization:

$$\hat{t}_R - t \cong \hat{t}_y - \frac{t_y}{t_x}\hat{t}_x = \sum_{k\in s} w_k r_k$$

w_k = survey base weight

$r_k = y_k - \frac{t_y}{t_x}x_k$; $\left(t_x/\hat{t}_x\right)\Big|_{\hat{\mathbf{t}}=\mathbf{t}} = 1$ in partials

  Binder recipe:

$$\hat{t}_R - t \cong \frac{t_x}{\hat{t}_x}\sum_{k\in s} w_k r_k$$

Retains $t_x/\hat{t}_x$ in variance estimate

- In case of srs without replacement

Standard linearization: $v_L = \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\frac{\sum_s r_k^2}{n-1}$

Royal-Cumberland/Binder:

$$v_2 = \left(\frac{\bar{X}}{\bar{x}_s}\right)^2 v_L$$

Design consistent (under srswor) and approximately model-unbiased (under model that motivates ratio estimator); special case of "*sandwich*" estimator

Problems in Panel Surveys

- Estimating change over time—involves data from 2 or more time periods
  - Linear substitute useful in multi-stage design if PSUs same in all periods (true in US CPI)
  - Not clean solution in single-stage sample with rotation of PSUs (establishments)—need to worry about non-overlap, births, deaths when computing variance of change

Price Indexes in Panel Surveys—Hard to Use Linearization

- Jevons (geometric mean) price index of change from time 0 to time t is product of 1-period price changes. Each 1-period change is estimated by a geomean:

$$\hat{P}_J\left(\mathbf{p}_t,\mathbf{p}_0\right)=\prod_{u=0}^{t-1}\hat{P}_J\left(\mathbf{p}_{u+1},\mathbf{p}_u,s_{u+1}\right)$$ where

$$P_J\left(\mathbf{p}_{u+1},\mathbf{p}_u,s_{u+1}\right)=\prod_{k\in s_{u+1}}\left(p_k^{u+1}\big/p_k^{u}\right)^{w_k E_k^a}$$

E_k^a = proportion of expenditure due to item k at a reference period a. s_{u+1} = set of sample items at $u+1$.

JPSM The Joint Program in Survey Methodology

- With t time periods, this is function of 1-period geometric means

$$\log\left[\hat{P}_J\left(\mathbf{p}_t,\mathbf{p}_0\right)\right]=\sum_{u=0}^{t-1}\log\left[\hat{P}_J\left(\mathbf{p}_{u+1},\mathbf{p}_u,s_{u+1}\right)\right]$$
$$=\sum_{u=0}^{t-1}\sum_{k\in s_{u+1}} w_k E_k^a \log\left(p_k^{u+1}\big/ p_k^u\right)$$

- In case of US CPI, could expand sum over $k\in s_{u+1}$ in terms of strata and PSUs, reverse sum over time and samples. Then get linear substitute. Add to sum as time moves on.
- Method would work (between decennial censuses) because PSU sample is fixed. With single-stage establishment sample, may not be feasible.

Jackknife

- Delete one 1st-stage unit at a time; compute estimate from each replicate
- $v_J = \sum_h \frac{n_h - 1}{n_h} \sum_{i \in s_h} \left(\hat{\theta}_{(hi)} - \hat{\theta} \right)^2$
- Works for $\hat{\theta} = g\left(\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_p\right)$; smooth g, with-replacement sampling of 1st-stage units
- Example: GREG estimator of a total

$$\hat{t}_G = \hat{t}_\pi + \hat{\mathbf{B}}' \left(\mathbf{t}_x - \hat{\mathbf{t}}_x \right)$$

- Exact formula for jackknife for GREG is available (Valliant, *Surv Meth* 2004) for single-stage, unequal probability sampling
- An approximation is

$$v_J\left(\hat{t}_G\right) \cong \sum_s \left(\frac{w_k g_k r_k}{1-h_k}\right)^2$$

$r_k = y_k - \mathbf{x}_k' \hat{\mathbf{B}}$; $g_k = 1 + \mathbf{x}_k' \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\left(\mathbf{t}_x - \hat{\mathbf{t}}_x\right)$ = g-weight

h_k = weighted regression leverage

Jackknife Linearization Method

- Yung & Rao, (*Surv Meth* 1996, *JASA* 2000)
- General idea: get linear approximation to $\hat{\theta}_{(hi)} - \hat{\theta}$ and substitute in v_J
- For GREG $v_{JL} = \sum_h \frac{n_h}{n_h - 1} \sum_{i \in s_h} \left(r_{hi}^* - \overline{r}_h^* \right)^2$

  where $r_{hi}^* = \sum_{k \in s_{hi}} w_k g_k r_k$; $\overline{r}_h^* = \sum_{i \in s_h} r_{hi}^* \big/ n_h$

- In single-stage sampling

  $v_{JL} = \sum_h \frac{n_h}{n_h - 1} \sum_{k \in s_h} \left(w_k g_k r_k - \overline{(wgr)}_h \right)^2$ (missing leverage adjustment, but good large sample design/model properties)

More Advanced Linearization Techniques

- Deville, *Surv Meth* (1999)
  - Formulation in terms of influence functions
  - Goal: estimate some parameter $\theta = T(F)$, a function of the distribution function of y
- $\hat{\theta} = T(\hat{F})$ can be linearized near F
  - Compute variance of linear approx;
  - Deville (1999) gives many examples: correlation coefficient, implicit parameters (logistic $\boldsymbol{\beta}$), Gini coefficient, quantiles, principal components

- Demnati & Rao, *Surv Meth* (2004)
  - Extension of Deville—unique way to evaluate partials
  - Estimating equations
  - Two-phase sampling

JPSM The Joint Program in Survey Methodology

Accounting for Imputations

- If imputations made for missing items, variance of resulting estimates (totals, ratio means, etc) usually increased.

- Treating imputed values as if real can lead to severe underestimates of variances.

- Special procedures needed to account for effect of imputing. Some choices:

  Multiple imputation MI (Rubin)

  Adjusted jackknife or BRR (Rao, Shao)

  Model-assisted MA (Särndal)

JPSM The Joint Program in Survey Methodology

To get theory for these methods, 4 different probabilistic distributions can be considered:

Superpopulation model	Sample design
Response mechanism	Imputation mechanism

- Assumptions needed for response mechanism, e.g. random response within certain groups and, in the cases of MI and MA, a superpopulation model that describes the analysis variable.

- How methods are implemented and what assumptions are needed for each depends, in part, on the imputation method used (hot deck, regression, etc).

Multiple Imputation

- MI uses a specialized form of replication. *M* imputed values created for a missing item $\Rightarrow$ must be a random element to how the imputations are created.

- $\hat{z}_{I(k)}$ = estimate based on the k-th completed data set

- $\hat{V}_{I(k)}$ = naïve variance estimator that treats imputed values as if they were observed. $\hat{V}_{I(k)}$ could be linearization, replication, or an exact formula.

- MI point estimator of θ is

$$\hat{z}_M = \frac{1}{M} \sum_{k=1}^{M} \hat{z}_{I(k)},$$

- Variance estimator is

$$\hat{V}_M = U_M + \left(1 + \frac{1}{M}\right) B_M \text{, where}$$

$$U_M = M^{-1} \sum_{k=1}^{M} \hat{V}_{I(k)} \text{ and}$$

$$B_M = (M-1)^{-1} \sum_{k=1}^{M} \left(\hat{z}_{I(k)} - \hat{z}_M\right)^2 .$$

For hot deck imputation, the MI method assumes a uniform response probability model and a common mean model within each hot deck cell.

- Overestimation in cluster samples—Kim, Brick, Fuller (*JRSS-B* 2006)

Adjusted jackknife

- Rao and Shao (1992) adjusted jackknife (AJ) variance estimator.

  In jackknife variance formula use

  $\hat{z}_I = g\left(\hat{y}_{I1},\ldots,\hat{y}_{Ip}\right)$ = full sample estimate including any imputed values

  $\hat{z}_{I(hi)} = g\left(\hat{y}_{I1(hi)},\ldots,\hat{y}_{Ip(hi)}\right)$ = an adjusted estimated with adjustment *dependent* on imputation method

JPSM The Joint Program in Survey Methodology

- Example: 1-stage stratified sample
  - Hot deck method: cells formed and donor selected with probability proportional to sampling weight
  - Adjusted $\hat{y}$ value, associated with deleting unit hi, is

$$\hat{y}_{I(hi)} = \sum_{g=1}^{G} \left\{ \sum_{j \in A_{Rg}} w_{j(hi)} y_j + \sum_{j \in A_{Mg}} w_{j(hi)} \left[y_j^* + e_{j(hi)} \right] \right\}$$

g = hot deck cell (which can cut across strata)

A_{Rg}, A_{Mg} = sets of responding and missing units in *g*

$w_{j(hi)}$ = adjusted weight for unit *j* when unit i in stratum *h* is omitted

y_j^* = hot deck imputed value for unit *j*

$e_{j(hi)} = \bar{y}_{Rg(hi)} - \bar{y}_{Rg}$, a residual specific to a replicate

- The AJ method assumes a uniform response probability model within each hot deck cell.

Software

- Options—Stata, SUDAAN, SPSS, WesVar, R survey package
- Off-the-shelf software may not cover what you need

  Likely omissions: NR adjustment, adaptive collapsing, specialized estimates (price indexes), item imputations

  $\Longrightarrow$ Write your own

JPSM The Joint Program in Survey Methodology

Summary

Linearization

Pros

- good large sample properties
- applies to complex forms of estimates
- can be computationally faster than replication
- maximizes degrees of freedom
- sandwich version is model-robust

Cons

- separate formula for each type of estimate
- special purpose programming
- hard to account for some sample adjustments, e.g., nonresponse, adaptive methods

Replication

Pros

- good large sample properties
- applies to complex forms of estimates
- sample adjustments easy to reflect in variance estimates
- applies to analytic subpopulations
- user does not need to know or understand sample design

Cons

- computationally intensive
- may be unclear how best to form replicates
- increased file sizes
- sometimes applied in ways that lose degrees of freedom

JPSM The Joint Program in Survey Methodology