Weight watchers: How to optimize your weight

2015 Nordic and Baltic Stata Users Group meeting

Michele Santacatterina, Matteo Bottai

Unit of Biostatistics, Karolinska Institutet

September 4th 2015

Outline

```
. *mean of Y in f0
. mean y

Mean estimation                     Number of obs    =       500

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
           y |   181.7257   1.892362      178.0077    185.4437
--------------------------------------------------------------

. *weighted mean of Y using w
. qui svyset id [pw=w]
. svy: mean y
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =         1          Number of obs    =       500
Number of PSUs   =       500          Population size  =       500
                                      Design df        =       499

--------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
           y |   157.3507   21.34233      115.4188    199.2826
--------------------------------------------------------------
```

```
. *check weights
. summ w, detail
```

w

Percentiles		Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	500
25%	0	0	Sum of Wgt.	500
50%	0		Mean	1
		Largest	Std. Dev.	12.14537
75%	0	27.62159		
90%	0	65.71575	Variance	147.51
95%	4.22e-21	163.1522	Skewness	14.48649
99%	19.60755	204.0528	Kurtosis	222.3649

Objective

1. We have ideal probability weights, w_{id}.
2. The weighted estimate has large variance.
3. We estimate the weights closest to w_{id} within a variance constraint.

We propose a general method to estimate optimal weights based on the solution of a nonlinear constrained optimization problem.

Introduction

In statistics, probability-weighted (PW) methods are commonly used to

- compensate for nonresponse,
- control for disproportional sampling fractions,
- balance covariate patterns, etc.

However, these methods behave poorly when PWs are highly variable, causing biased estimates and high standard errors.
In survey literature, several techniques have been proposed to address this issue, commonly known as

- weight trimming (truncation) [4] and
- weight smoothing techniques ([5], [1]).

Additionally, within the class of doubly-robust methods [6], new estimators have been proposed to address highly variable weights and models misspecifications yielding to an improved performance (e.g. [7]).

Nonetheless, these methods lack of objective criteria.

The general nonlinear constrained optimization problem (NLOP)

Let $\mathbf{p}$ be a vector of parameters in $\mathbb{R}^k$, then we consider (NLOP) in the form

$$\begin{aligned} \underset{\mathbf{p}\in\mathbb{R}^k}{\text{minimize}} \quad & f(\mathbf{p}) \\ \text{subject to} \quad & g_i(\mathbf{p}) \leq 0, \quad i \in \mathcal{I}, \quad \mathcal{I} \cup \varepsilon = \{1, \ldots, m\} \\ & h_i(\mathbf{p}) = 0, \quad i \in \varepsilon, \quad \mathcal{I} \cap \varepsilon = \emptyset \\ & \mathbf{p} \geq 0 \end{aligned} \tag{NLOP}$$

where $f : \mathbb{R}^k \mapsto \mathbb{R}$, $g : \mathbb{R}^k \mapsto \mathbb{R}^m$ and $h : \mathbb{R}^k \mapsto \mathbb{R}^n$, and where $\mathbf{p} \geq 0$ are k bound constraints. To solve (NLOP) we used the method of sequential quadratic programming (SQP). To fix ideas, SQP method essentially find an optimal solution $\mathbf{p}_*$ to (NLOP) solving a sequence of sub-quadratic problems that are local approximations to (NLOP) [3].

Find optimal PWs solving (NLOP)

Let $f(P, Y) = \theta_w$ be the estimator we are interested in (e.g. the weighted mean or the weighted total) and define $Var(\theta_w) = \sigma_w^2$ its variance. Let $\hat{\sigma}_w^2$ be an unbiased estimator of σ_w^2, and let $\mathbf{nk} \in \mathbb{R}^K$ be a vector containing the size of each covariate pattern. We formulate our (NLOP) in order to minimize the distance (or divergence) between the ideal PWs and the vector of parameters $\mathbf{p}$ subject to

- a inequality constraint of $\hat{\sigma}_w^2$;
- an equality constraint on the sum of the PWs and
- k bound constraints to ensure positivity of the PWs.

Find optimal PWs solving (NLOP)

Formally, we define (NLOP1) in the form

$$\begin{aligned} \underset{\mathbf{p}\in\mathbb{R}^k}{\text{minimize}} \quad & \mathbf{d}(\mathbf{p}, \mathbf{w_{id}}) \\ \text{subject to} \quad & \hat{\sigma}_w^2 - \alpha \leq 0 \\ & \mathbf{n}\mathbf{k}^T\mathbf{p} - n = 0 \\ & \mathbf{p} \geq 0 \end{aligned} \tag{NLOP1}$$

where, $\mathbf{d}(\mathbf{p}, \mathbf{w_{id}}) = \mathbf{d}(\mathbf{w_{id}}, \mathbf{p}) = \sum_k^K (p_k - w_{id,k})^2$ is the Euclidian distance squared. (NLOP1) is a convex problem.

Simulation

To examine the performance of our proposed method we simulated $n = 200$ data from two Normal distributions, one for the random variable $X \sim \mathcal{N}(40, 2)$, and one for the outcome $Y \sim \mathcal{N}(4x+20, 100)$ and we compared the weighted estimator at a specific value x of the random variable X derived by using

- the optimal PWs obtained by solving (NLOP1) for different levels of α;
- the trimmed PWs at the 90^{th} percentile and
- the ideal PWs.

We chose the weighted mean μ_w as the estimator we are interested in. The α levels were set equal to

- $\alpha = Var(\mu_0)$ where the variance of the weighted mean equals the variance of the sampled mean;
- $\alpha = Var(\mu_{trim})$, where μ_{trim} is the trimmed mean, and
- $\alpha = Var(\mu_{id})$ where μ_{id} is the ideal weighted mean.

Finally, we chose $x = 35$, and $f_1(X) \sim \mathcal{N}(35, 2)$.

Comparison of weighted means

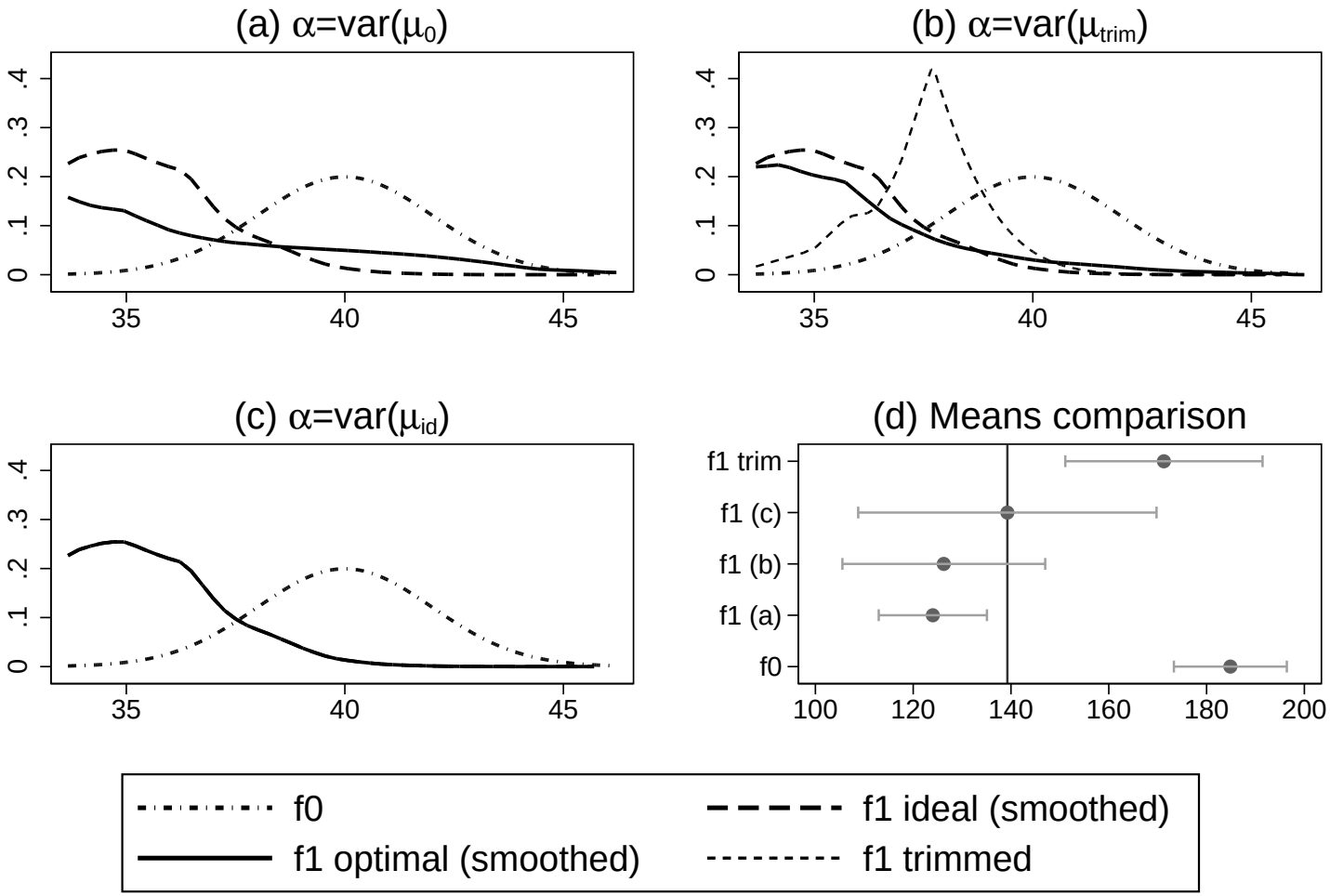


Figure: Comparison of weighted means: (a) $\alpha = Var(\mu_0)$; (b) $\alpha = Var(\mu_{trim})$; (c) $\alpha = Var(\mu_{id})$; (d) weighted means with relative confidence intervals.

Real data application

We applied our method on the evaluation of the heterogeneity of the mean CD4 cell count after 24 months of study across percentiles of

- age at baseline,
- CD4 cell count at baseline.

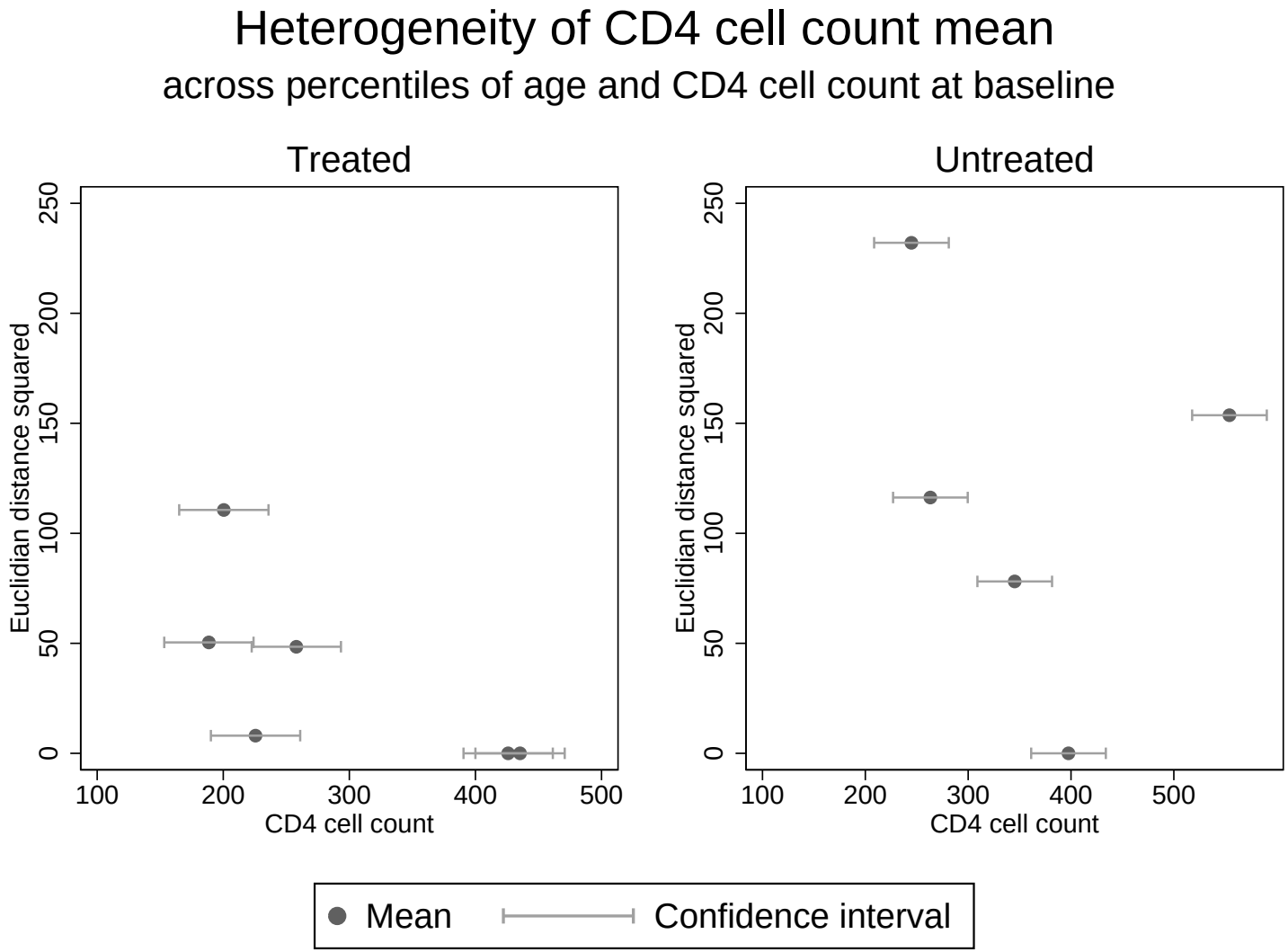
A randomized clinical trial about the impact of peer-support on virological failure and mortality in Vietnam [2] was used for the analyses.

```
/*****************************************************************************/
* f0
/*****************************************************************************/

/*kernel density estimation for each rv of interest*/

kdensity CD4_BL, nograph at(CD4_BL) gen(fc)
kdensity age_in_study, nograph at(age_in_study) gen(fa)

/*f0 becomes the product of the two (i am assuming them indp)*/
gen f0 = fc*fa

/*****************************************************************************/
* f1
/*****************************************************************************/
/*CD4 */
_pctile CD4_BL , p(30)
gen xc = CD4_BL-r(r1)
gen f1c = normalden(xc/8)/8

/*age_in_study*/
matrix res[‘k’,2] = ‘j’
gen xa = age_in_study-‘j’
gen f1a = normalden(xa/.5)/.5

gen f1 = f1c*f1a
```

```
/*****************************************************************************/
* weights
/*****************************************************************************/
/*ideal_weight*/
gen ideal_weight = f1 / f0
summ ideal_weight, detail
replace ideal_weight = ideal_weight / r(mean)

/*****************************************************************************/
* set up
/*****************************************************************************/

/*set up the dataset*/
*sum
bysort CD4_BL age_in_study: egen Sum = sum(CD4_24m)
*sum2
gen cd4_2 = CD4_24m^2
bysort CD4_BL age_in_study: egen Sum2 = sum(cd4_2)
*NK
bysort CD4_BL age_in_study: egen NK = count(CD4_24m)
*iterations
gen maxite = 1000
*init
gen init = 1/N
*alpha level
qui mean CD4_24m
matrix a = e(V)
gen alpha = a[1,1]
```

```
/*****************************************************************************/
* optimize
/*****************************************************************************/

*set the wrapper
program nlopt, plugin using("C:ado\personal\slsqp.dll")

*solve the problem
plugin call nlopt init ideal_weight Sum Sum2 NK alpha N maxite opt_w



---------- Iteration = 1
f(x) = 11011.19487
h(x) = -153
g(x) = 4.504653475e-06

...
...
...

---------- Iteration = 1000
f(x) = 48.43127702
h(x) = 1.705302566e-13
g(x) = 2.020914849e-07
nlopt ok!
```

Results


Heterogeneity of CD4 cell count mean
across percentiles of age and CD4 cell count at baseline
Treated
Untreated
Euclidian distance squared
0
50
100
150
200
250
100
200
300
400
500
CD4 cell count
Mean
Confidence interval

Results

f	age	μ_i	se(μ_i)	lb	ub	$d(\mathbf{p}, \mathbf{w_{id}})$
Treated						
f_0	32	425.96104	17.924854	390.54887	461.37321	0
f_{26}	26	435.40811	17.924855	399.99595	470.82028	.00062383
f_{32}	32	188.58567	17.924853	153.17351	223.99784	50.416523
f_{35}	35	225.58382	17.924853	190.17166	260.99599	8.0331154
f_{38}	38	257.941	17.924852	222.52884	293.35317	48.431286
Untreated						
f_0	31	397.57746	18.365707	361.26972	433.88521	0
f_{26}	26	220.89805	18.365708	184.5903	257.2058	198.80678
f_{31}	31	283.33135	18.365706	247.0236	319.63909	91.749512
f_{32}	32	286.54874	18.365707	250.24099	322.85648	.22424519
f_{33}	33	311.334	18.365708	275.02626	347.64175	149.49336
f_{34}	34	345.04854	18.365707	308.7408	381.35629	84.547058

Treatment effect among patients with CD4 cell count 31 cells/μL and age 26 years old is equal to $f_{26,tr} - f_{26,untr} = 435.40811 - 220.89805 = 214.51006$.

Treatment effect among patients with CD4 cell count 31 cells/μL and age 32 years old is equal to $f_{32,tr} - f_{32,untr} = 188.58567 - 286.54874 = -97.96307$.

Conclusions

- PWs are used in many settings;
- PWs can be highly variable;
- ad-hoc or "rule-of-thumb" methods used to solve the issue of high variability;
- we proposed a more rigorous method based on the solution of a NLOP that provides global, optimal solution with a straightforward practical interpretation.

Beaumont, J.-F.

A New Approach to Weighting and Inference in Sample Surveys.
Biometrika 95, 3 (Sept. 2008), 539–553.

Do Duy, C.

Antiretroviral therapy among HIV-infected persons in Northeastern Vietnam : Impact of peer support on virologic failure and mortality in a cluster randomized controlled trial.
Dept of Public Health Sciences, Sept. 2012.

Johnson, S.

The NLopt nonlinear-optimization package.

Kokic, P., and Bell, P.

Optimal winsorizing cutoffs for a stratified finite population estimator.
Journal of Official Statistics 10, 4 (1994), 419.

Potter, F.

A study of procedures to identify and trim extreme sampling weights.
In *Proceedings of the American Statistical Association, Section on Survey Research Methods*, vol. 225230. 1990.

Robins, J. M., Rotnitzky, A., and Zhao, L. P.

Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data.
Journal of the American Statistical Association 90, 429 (Mar. 1995), 106–121.

Rotnitzky, A., Lei, Q., Sued, M., and Robins, J. M.

Improved double-robust estimation in missing data and causal inference models.
Biometrika 99, 2 (June 2012), 439–456.