כנס "מאין באנו ולאן הגענו"

סקר ארוך טווח של הלמ"ס

31 בינואר 2018

חשיבות השימוש במקדמי ניפוח במחקר מבוסס סקרים

פרופ' דני פפרמן

הסטטיסטיקן הלאומי

Basic Concepts and Notation

U - population = ***finite*** set of units (elements)

e.g., all individuals aged 16+ in Israel.

U = {1,....,***N***} ; ***N*** = **population size**

Variables Y, X,… taking values $y_i, x_i, \ldots$

Finite (Summary) population parameters

e.g., Total: $Y = \sum_U y_i$; **mean:** $\bar{Y} = N^{-1}\sum_U y_i$

Regression coefficient, $B = \dfrac{\sum_U (x_i - \bar{x}) y_i}{\sum_U (x_i - \bar{x})^2}$,…

'Superpopulation' model,

e.g., $y_i \sim N(\mu, \sigma^2)$; Parameters of interest, μ, σ^2

Estimation of *Y*, $\bar{Y}$, *B* → **Descriptive**

Estimation of μ, σ^2 → **Analytic**

Basic Concepts and notation (cont.)

S = **sample** – subset of population for which we collect data.

sampling design – method of selecting the sample.

sample size – *n* (may be fixed or random).

data - $y_i, x_i, \ldots; i \in s$ (**if** no **non-response**)

estimator – function of data

e.g., $\bar{y} = n^{-1} \sum_s y_i$,

estimates $\bar{Y} = N^{-1} \sum_U y_i$ and $\mu = E(y_i)$.

If sample selected with **equal** probabilities.

Examples of Sampling Designs

1- Simple random sampling without replacement (SRSWOR)

Each subset of size ***n*** has equal probability of forming the sample ($\Rightarrow$ each subset of size $m \leq n$ has equal probability of being in the sample).

2- Stratified sampling

Population partitioned into **strata**, $h=1,...,L$, **e.g.,** regions of country.

Select samples independently within each stratum (for example by **SRSWOR**)

- improves precision if Y is homogeneous within strata,
- Permits **over-sampling** domains of interest,
- Permits estimation of strata parameters,
- Permits efficient sample allocations.

3- (Single stage) Cluster Sampling

Population partitioned into (**natural**) clusters, **e.g. *areas*, households ,…**

Select sample of clusters, select all units within the selected clusters.

- **Loss** in efficiency, **big saves** in costs

4. Multistage Cluster Sampling

Primary sampling units (**PSUs**)

Secondary sampling units (**SSUs**)

Select sample of PSUs**;** select sample of SSUs within each selected PSU

- Often necessary if frame is hierarchical,
- Interview workloads reduced if, for example, PSUs are **areas**,
- Does not require list of SSUs.

4- Probability Proportional to Size (PPS) Sampling

Suppose there exist ***M*** PSU's. Measure of size of PSU ***g*** is M_g. Select PSUs such that,

$$\Pr(g \in s) \propto \mathrm{M_g}.$$

e.g. Household surveys

PSUs defined by household.

Pr(select PSU ***g***) = nM_g / M. (***n*** **number of selected households).**

SSUs defined by **members within households**

$\boldsymbol{M}_{\mathrm{g}}$ = number of ***members*** in PSU **g** .

Pr (select ***member i*** in selected PSU g) = $\frac{m_g}{M_g}$.

$$\Downarrow$$

$$\Pr[\text{select member } (i,g)] = \frac{nM_g}{M} \times \frac{m_g}{M_g} = \frac{nm_g}{M}$$

Constant if $m_g = m$.

General Estimation Theory (Design-Based)

Parameter θ (function of $\mathbf{y_1, ..., y_N}$), **e.g.,** $\theta = \bar{Y}$

Estimator $\hat{\theta}$ (function of y_i, $\mathbf{i \in s}$); $\hat{\theta} = \hat{\theta}(s)$

e.g., $\hat{\theta} = \bar{y}_s$.

s = source of **random variation** (**different samples**$\Rightarrow$**different estimates**. Population values considered as **fixed** numbers).

p(s) = probability of selecting **s**

S = set of all possible samples **s**

$E_D(\hat{\theta}) = \sum_s \hat{\theta}(s)\, p(s)$; **D-bias** = $[E_D(\hat{\theta}) - \theta]$.

$Var_D(\hat{\theta}) = \sum_s p(s)[\hat{\theta}(s) - E_D(\hat{\theta})]^2$.

Standard error = s.e.$(\hat{\theta})$ = $\sqrt{Var_D(\hat{\theta})}$.

Central limit theorem,

$$\hat{\theta}(s) \underset{n\to\infty}{\sim} \boldsymbol{N}\,[E_D(\hat{\theta}),\, Var_D(\hat{\theta})].$$

Bias from Ignoring Sampling Design

Suppose we ignore the sampling design and wish to estimate population **expectation** $\mu = E(y_i)$ by sample mean $\bar{y}_s = n^{-1}\sum_s y_i$.

Suppose ***n*** is **fixed** under the sampling design.

Let $\mathbf{I_i = 1}$ if unit ***i*** sampled and $\mathbf{I_i = 0}$ if not.

Pr(unit ***i*** sampled) = π_i = **inclusion probability.**

It follows that,

$$E(\bar{y}_s) = \sum_U \pi_i E(y_i \mid I_i = 1) / \sum_U \pi_i .$$

$= \mu$ **if** $E(y_i \mid I_i = 1)$ (**unbiased**).

But Biased if $E(y_i \mid I_i = 1) \neq E(y_i) = \mu$.

Informative sampling - π_i **related to** y_i**.**

Example – Stratified Sampling (2 strata)

SRS of size n_h within stratum h of size N_h

$\pi_i = \frac{n_h}{N_h}$ for every ***i*** in stratum ***h***.

Population mean $\mu = \frac{N_1\mu_1 + N_2\mu_2}{N_1 + N_2}$.

$$\bar{y}_s = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2} \Rightarrow E(\bar{y}_s) = \frac{n_1\mu_1 + n_2\mu_2}{n_1 + n_2}$$

Generally **biased** unless $\mu_1 = \mu_2 = \mu$, **or**

$n_1 / N_1 = n_2 / N_2$ **(EPSEM).**

Proportional sample allocation.

Weighting for unequal selection probabilities

Define **sample weight** for unit ***i*** as,

$w_i = \pi_i^{-1}$ **(expansion weight, base weight).**

Horvitz-Thompson estimator of $t = \sum_U y_i$ is

$\hat{t}_{HT} = \sum_s w_i y_i$.

For given population values,

$$E_D(\hat{t}_{HT}) = \sum_U \frac{1}{\pi_i} y_i \pi_i = \sum_U y_i \Rightarrow \textbf{D-unbiased}$$

Weighting for Nonresponse

Respondents = **R**= subset of sample **s**

$\pi_{r|i}$ = Pr (unit ***i*** **responds** | ***i*** sampled)

We would like to set,

$w_{r|i} = \pi_{r|i}^{-1}$ (**conditional response weight**)

$w_i = \pi_i^{-1}$ (**base weight**)

Combined response weight:

$$\tilde{w}_i = w_{r|i} w_i = \frac{1}{\Pr(i \text{ observed})} \rightarrow$$

→ **"2-stage sampling".**

But $\pi_{r|i}$ usually **unknown.**

Sample-based nonresponse weights

Suppose both respondents and nonrespondents can be divided into **H weighting classes, e.g.,** areas or socio demographic groups.

For unit ***i*** in class ***h***,

$$\hat{\pi}_{i|h} = \frac{\text{number of respondents in class h}}{\text{number of sampled units in class h}}$$

Bias eliminated if response probability is indeed the **same** for all the units within the same weighting class (**noninformative nonresponse** within classes).

➢ Modelling the response probabilities **very difficult in practice**. Depends on what is known about the **response mechanism**.

Calibration

When population **totals** $X = X_1,...,X_K$ are **known** for measured values $x_i = x_{1i}...x_{Ki}$, we may want to change the sampling weights such that the estimates are **calibrated** to these totals, **i.e., change** w_i to w_i^* such that $\sum_{i\in s} w_i^* x_i = X$.

Estimate total Y as $\hat{Y}_{Cal} = \sum_{i\in s} w_i^* y_i$.

Perfect if $y_i \cong A_1 x_{1i} + ... + A_k x_{Ki}$ (linear).

The new weights w_i^* are chosen such that they are close to the original weights w_i by some distance function, **e.g., minimize,**

$$\sum_{i\in s} \frac{(w_i^* - w_i)^2}{w_i} \text{ subject to } \sum_{i\in s} w_i^* x_i = X.$$

➢ Calibration often used to adjust nonresponse.

Weighting in Regression

Denote $y_i \rightarrow$ value of the dependent variable,

$x_i' \rightarrow$ values of explanatory variables,

OLS estimator:

$$\hat{\beta}_{OLS} = (X'X)^{-1} X'y = \left(\sum_s x_i x_i'\right)^{-1} \sum_s x_i y_i .$$

Sample-weighted (*probability weighted*) est.:

$$\hat{\beta}_w = \left(\sum_s w_i x_i x_i'\right)^{-1} \sum_s w_i x_i y_i$$

w_i = sampling weight.

$$E_D\left(\hat{\beta}_w\right) \approx \left(\sum_U x_i x_i^t\right)^{-1} \sum_U x_i y_i = B \text{ (census est.)}$$

The expectation is over repeated sampling.
(True under **any** population model).

If **population model** is,

$$y_i = x_i'\beta + e_i, E_M\left(e_i \mid x_i\right) = 0$$

where E_M denotes expectation under model,

$\Rightarrow E_M E_D\left(\hat{\beta}_w\right) \approx \beta,$

- $\hat{\beta}_w$ **consistent** for β (under **correct** model).

Why weight?

Robustness against Model Misspecification

$\hat{\beta}_w$ is **design consistent** for B **(census estimate)**

B is well-defined ***population parameter*** even if the model is **misspecified.**

B minimises $\sum_U \left(y_i - x_i^t \beta\right)^2$,

Defines the **best** fitting **linear** regression function in finite population.

$\hat{\beta}_w$ minimizes $\sum_{i \in s} w_i \left(y_i - x_i^t \beta\right)^2$, which is design unbiased for $\sum_U \left(y_i - x_i^t \beta\right)^2$.

- $E_D[\sum_{i \in s} w_i \left(y_i - x_i^t \beta\right)^2] = \sum_U \left(y_i - x_i^t \beta\right)^2$.

Example

Suppose that the wage **pay-off** to **education** is lower for women than for men and that the sampling design **oversamples women**.

Correct model

$$y_i = \beta_0 + \beta_1 Ed_i + \beta_2 S_i + \beta_3 [Ed_i \times S_i] + e_i$$

where $\mathbf{S_i = 1}$ *if* $i = woman$; $\mathbf{S_i = 0}$ *if* $i = man$.

Note: $\beta_3 [Ed_i \times S_i]$ accounts for **difference** in coefficient of education

Misspecified model, $y_i = \alpha_0 + \alpha_1 Ed_i + \alpha_2 S_i + e_i^*$

The ***unweighted*** **est.** $\hat{\alpha}_{1,OLS}$ **underestimates** the strength of relationship between wages and education in the population. (**women are oversampled**).

The ***weighted*** **est.** $\hat{\alpha}_{1,W}$ estimates the **average increase in hourly wage per year of education** in the population **under study**.

- The misspecified model **tells nothing** about the **different wage-education relationships** in the two groups or about **other populations.**

Weighting: Pros and Cons

Pros

- Simple adjustment for informative sampling.
- Possible advantage of robustness.
- Guarantees use of correct variance estimator.
- Consistent with general practice.

Cons

- Complicates standard model fitting.
- Does not permit conditioning on selected sample (**observed x's**).
- Distribution of weighted estimators **generally unknown**.
- May inflate variance, **e.g.,**

$y_i = x_i^t \beta + e_i, var(e_i) = \sigma_i^2$, $\hat{\beta}_w$ **Best** if $\pi_i \propto \sigma_i^2$ **(GLS).**

If not, loss of efficiency under ignorable sampling.

Comparing Weighted and Unweighted Est.

- May detect **model misspecification**
- permits detection of **sample selection bias** (**informative sampling**)

If no model misspecification, and sampling scheme is ignorable then

$H_0 : E(\hat{\beta}) = E(\hat{\beta}_w)$ holds

Testing of H_0 can be implemented by a variety of methods.

- Treat as diagnostic test – if H_0 **rejected** may either signifies model **misspecification** or **informative** sampling.

Example: National Maternal and Infant Health (NMIH) Survey, U.S.A.

Disproportionate stratified random sample of **Vital Records**. Strata defined by mother's ***race*** (‘***black***’, ‘***white***’) and child's ***birth weight;***

$$y < 1500,\ 1500 \le y < 2500,\ y \ge 2500$$

Simulation study: Consider the sample data as ‘**population**’, select independent samples with probabilities proportional to original selection probabilities. For each sample estimate the regression of ***birth weight*** (measured in **grams**) on **1st**, **2nd** and **3rd** powers of ***gestational age*** (measured in weeks).

The ‘***population model***’ (fitted by **OLS**) is,

$$y_i = 17866 - 1827.7x_i + 61.2x_i^2 - 0.61x_i^3 + e_i.$$

All the coefficients are highly **significant,**

$R^2 = 0.61$, $\sigma_e^2 = 603.2$. $\boldsymbol{Corr(w_i, \hat{e}_i)} = \mathbf{0.30}$.

$\Downarrow$

Informative sampling.

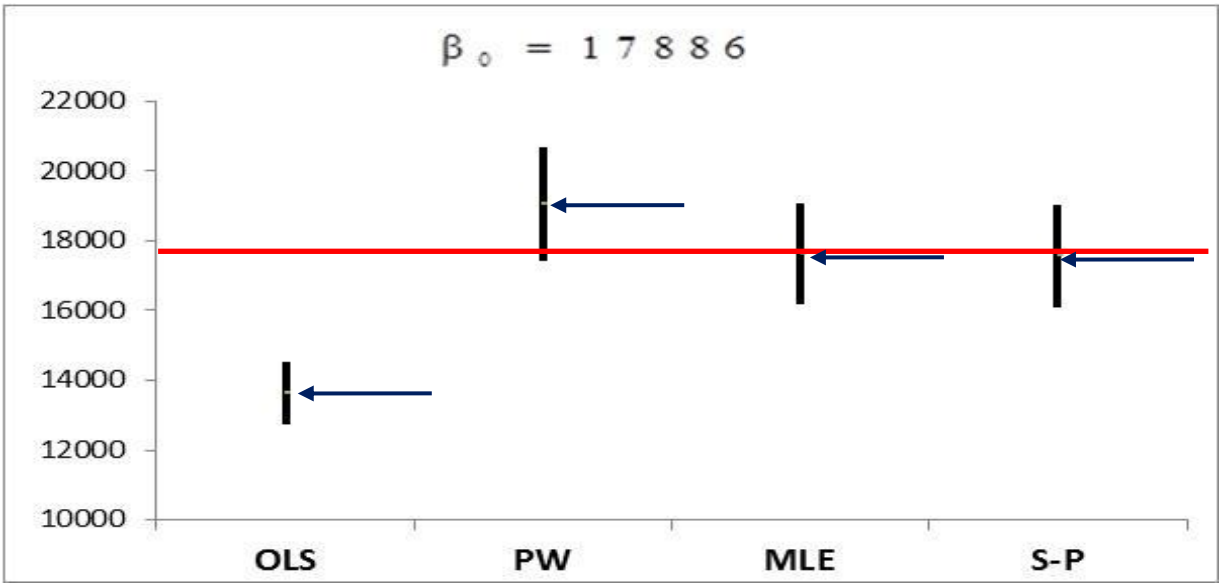
Results: Means and Standard Deviations (SD) of Means of Regression Estimates over 100 samples Selected from NMIH Data.

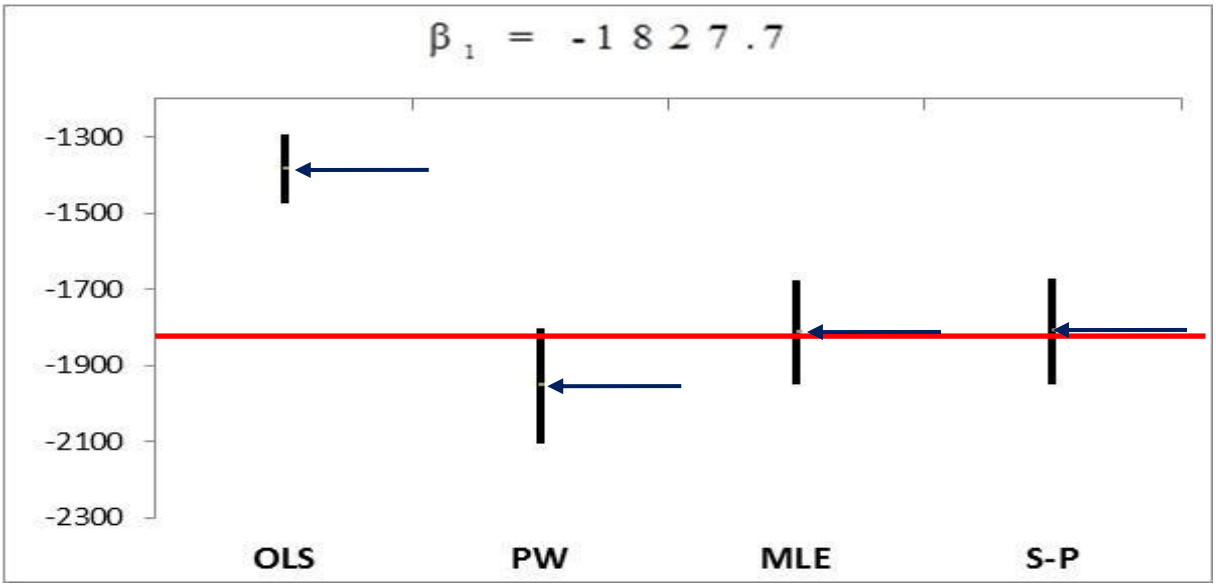
'Pop.' size =9447, Av. Sample size=233.6.

Method

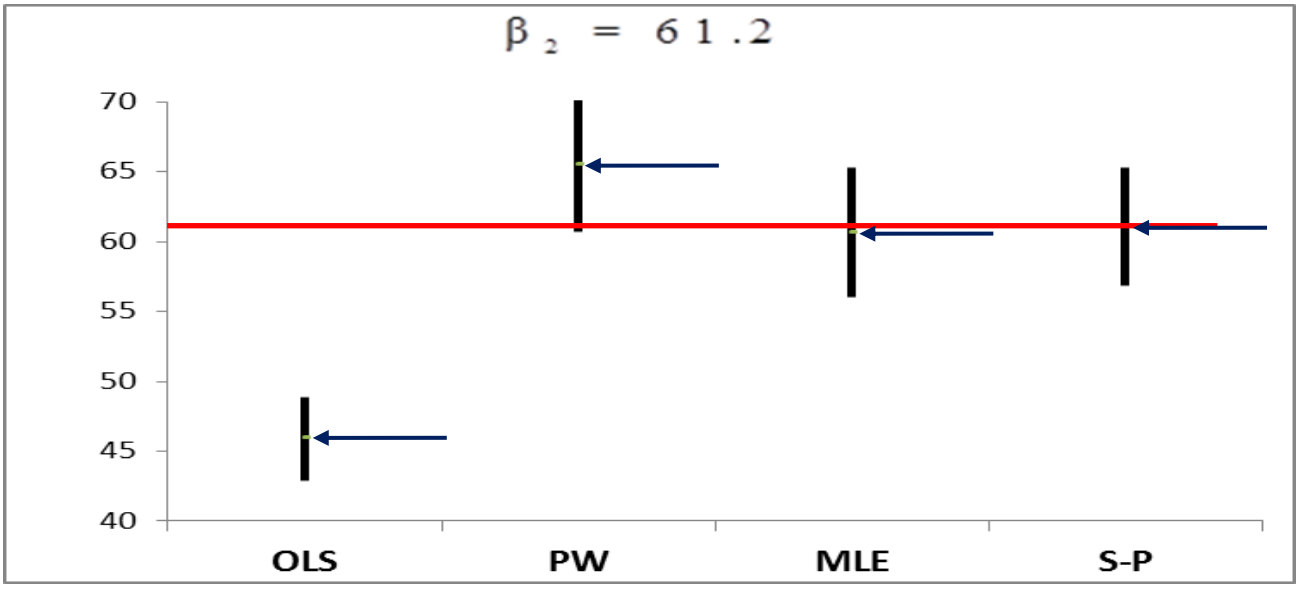
True Coeff. SD Bet. Est.	OLS	PW	MLE	S-P
$\beta_0 = 17886$	13625.2	19035.7	17630.8	17556.3
$SD(\hat{\beta}_0)$	453.9	810.6	721.7	733.3
$\beta_1 = -1827.7$	-1382.8	-1952.7	-1813.7	-1809.5
$SD(\hat{\beta}_1)$	45.8	76.4	68.6	69.0
$\beta_2 = 61.2$	45.90	65.50	60.21	61.07
$SD(\hat{\beta}_2)$	1.5	2.4	2.1	2.1
$\beta_3 = -0.61$	-0.45	-0.66	-0.60	-0.62
$SD(\hat{\beta}_3)$	0.02	0.02	0.02	0.02

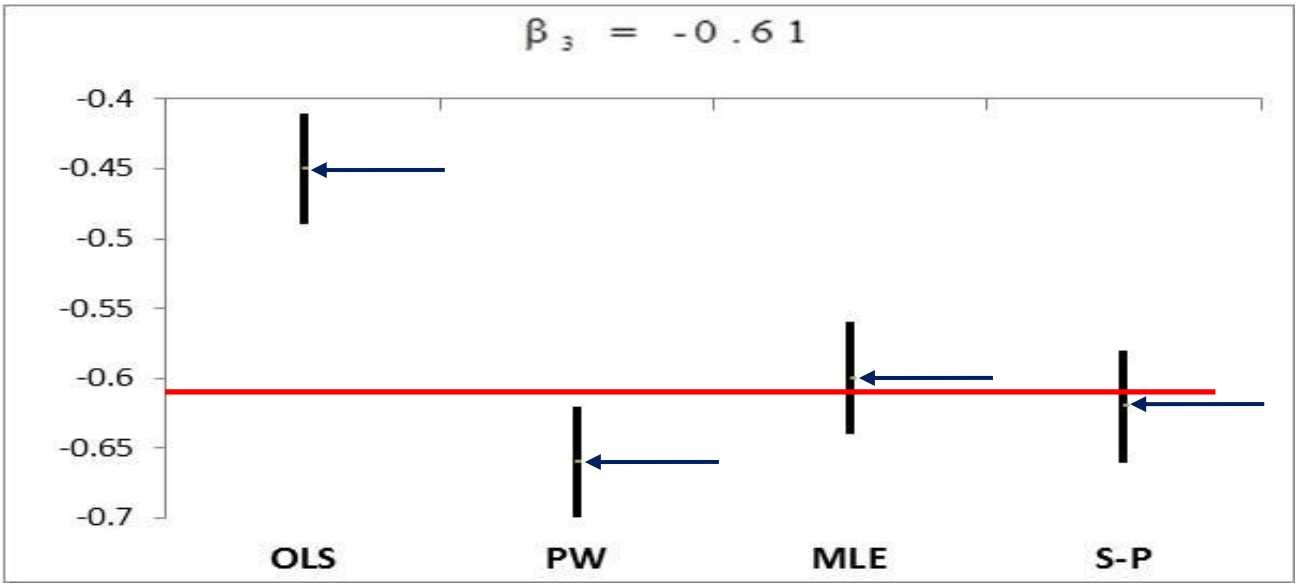
מקדמים אמתיים, אומדנים ורווחי סמך





מקדמים אמתיים, אומדנים ורווחי סמך (המשך)





תודה על ההקשבה. סליחה על הנוסחאות.