Fundament of Inference - Michael Elliott

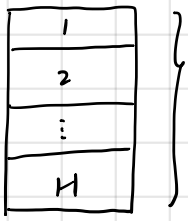
Different adjustment estimators. E.g.,
post-stratification;
model-assisted estimator;
model-based estimator

2018-2019 Spring

Auxiliary information

Point I : what's poststratification

Poststratification (often based on auxiliary information), developing weights so that the weighted mean of the sample matches that of population (external information)



Population N has H strata, and N_h are known.

For an unit i that belongs to stratum h : $w_{hi} = \frac{N_h}{n_h}$

Poststratified estimator of population total :

$$\hat{T}_{ps} = \sum^{n} w_{hi} y_{hi} = \sum^{H} \sum^{n_h} w_{hi} y_{hi} = \sum^{H} \sum^{n_h} \frac{N_h}{n_h} \cdot y_{hi} = \sum^{H} N_h \sum^{n_h} \frac{y_{hi}}{n_h}$$

$$= \sum^{H} N_h \cdot \bar{y}_h$$

Poststratified estimator of population mean :

$$\hat{\bar{Y}}_{ps} = \frac{\hat{T}_{ps}}{N} = \frac{\sum^{H} N_h \cdot \bar{y}_h}{N} = \sum^{H} \frac{N_h}{N} \bar{y}_h$$

$$= \sum^{H} P_h \cdot \bar{y}_h$$ ← note that this is also $\bar{y}_{ps}$ (poststratified sample mean)

Point II.

Poststratification usefulness 1: Poststratified estimator has reduced variance

Poststratification is useful bc it reduces variance of the estimator

poststratified estimator's variance:

$$V(\bar{y}_{ps}) = E(V(\bar{y}_{ps}|n_1, n_2 \cdots n_h)) + V(E(\bar{y}_{ps}|n_1, n_2 \cdots n_h))$$

(dataset is a random variable, this means conditioning on realization of the dataset (i.e. current data))

in which $E(\bar{y}_{ps}|n_1, n_2 \cdots n_h) = E(\sum_h^H P_h \cdot \bar{y}_h) = \sum_h^H P_h E(\bar{y}_h) = \sum_h^H P_h \bar{Y}_h = \bar{Y}$

$$V(\bar{y}_{ps}|n_1, n_2 \cdots n_h) = V(\sum_h^H P_h \bar{y}_h) = \sum_h^H P_h^2 V(\bar{y}_h) = \sum_h^H P_h^2 \frac{S_h^2}{n_h}$$

$$\therefore V(\bar{y}_{ps}) = E(\sum_h^H P_h^2 \frac{S_h^2}{n_h}) + \cancel{V(\bar{Y})^0} \approx \sum_h^H P_h^2 \frac{S_h^2}{E(n_h)} = \sum_h^H P_h^2 \frac{S_h^2}{n \cdot P_h}$$

$\therefore$ approximately $= \sum_h^H P_h S_h^2 / n$

$E(\frac{1}{n_h}) > \frac{1}{E(n_h)} \because (\frac{1}{3} + \frac{1}{2})/2 > \frac{1}{(3+2)/2}$

while we know that $E(\frac{1}{n_h}) > \frac{1}{E(n_h)}$, there is no close equation to calculate by how much.

$\therefore$ Approximate a term to compensate

$$V(\bar{y}_{ps}) \approx \frac{1}{n}\left[\sum_h^H P_h S_h^2 + \sum_h^G P_h(1-P_h)\frac{S_h^2}{n_h}\right]$$

SRS's variance: ignoring finite population correction

$$V(\bar{y}) = \frac{S^2}{n}$$

where S^2 is population variance $= \frac{1}{N-1}\sum_h^H \sum_i^{N_h} (Y_{hi} - \bar{Y})^2$

$$= \frac{1}{N-1}\sum_h^H \sum_i^{N_h} (Y_{hi} - \bar{Y}_h + \bar{Y}_h - \bar{Y})^2$$

$$= \frac{1}{N-1}\sum_h^H \sum_i^{N_h} \left[(Y_{hi} - \bar{Y}_h)^2 + (\bar{Y}_h - \bar{Y})^2 + 2(Y_{hi} - \bar{Y}_h)(\bar{Y}_h - \bar{Y})\right]$$

$$= \frac{1}{N-1}\left[\sum_h^H \sum_i^{N_h} (Y_{hi} - \bar{Y}_h)^2 + \sum_h^H \sum_i^{N_h} (\bar{Y}_h - \bar{Y})^2 + 2\sum_h^H \sum_i^{N_h} (Y_{hi} - \bar{Y}_h)(\bar{Y}_h - \bar{Y})\right]$$

$$= \frac{1}{N-1}\left[\sum_h^H \sum_i^{N_h} (Y_{hi} - \bar{Y}_h)^2 + \sum_h^H \sum_i^{N_h} (\bar{Y}_h - \bar{Y})^2 + 2\sum_h^H (\bar{Y}_h - \bar{Y}) \cancel{\sum_i^{N_h} (Y_{hi} - \bar{Y}_h)}^0\right]$$

$$= \frac{1}{N-1}\left[\sum_h^H (N_h - 1) S_h^2 + \sum_h^H N_h (\bar{Y}_h - \bar{Y})^2\right]$$

$$\approx \frac{1}{N}\left[\sum_h^H N_h \cdot S_h^2 + \sum_h^H N_h (\bar{Y}_h - \bar{Y})^2\right]$$

$$= \sum_h^H P_h \cdot S_h^2 + \sum_h^H P_h (\bar{Y}_h - \bar{Y})^2$$

$$\therefore V(\bar{y}) = \frac{1}{n}\sum_h^H P_h \cdot S_h^2 + \frac{1}{n}\sum_h^H P_h (\bar{Y}_h - \bar{Y})^2$$

$$= V(\bar{y}_{ps}) + \frac{1}{n}\sum_h^H P_h (\bar{Y}_h - \bar{Y})^2$$

$\Rightarrow V(\bar{y})$ under SRS is $\geq V(\bar{y}_{ps})$

Point III.

Poststratification usefulness 2: Correct for non-response bias

Poststratification is originally designed for reducing estimator's variance, but the post-stratified estimator also correct bias in the setting of non-response.

First, a little bit about non-response error

Deterministic approach: population consists of respondents ($\bar{Y}_R$) non-respondents ($\bar{Y}_{NR}$)

Then Bias = $\underbrace{(1-R)}_{\text{proportion of respondents}}(\bar{Y}_R - \bar{Y}_{NR})$

Stochastic approach: Each element in the population has a response propensity P_i

$$\text{Bias} = \frac{C(Y, P)}{\bar{P}}$$

$C(Y,P)$ ← covariance between Y and P

$\bar{P}$ ← average P

Assuming deterministic approach:

$$\bar{Y} = \sum^H P_h \cdot \bar{Y}_h = \sum^H P_h \left(\frac{N_R}{N} \bar{Y}_{hR} + \frac{N-N_R}{N} \bar{Y}_{h,NR} \right)$$

$$= \sum^H P_h \left[R_h \bar{Y}_{hR} + (1-R_h) \bar{Y}_{hNR} \right]$$

$$= \sum^H P_h \cdot R_h \cdot \bar{Y}_{hR} + \sum^H P_h \cdot \bar{Y}_{hNR} - \sum^H P_h \cdot R_h \bar{Y}_{hNR}$$

$$= \sum^H P_h \cdot \bar{Y}_{hR} - \sum^H P_h \bar{Y}_{hR} + \sum^H P_h \cdot R_h \bar{Y}_{hR} + \sum^H P_h \cdot \bar{Y}_{hNR} - \sum^H P_h \cdot R_h \bar{Y}_{hNR}$$

$$= \sum^H P_h \cdot \bar{Y}_{hR} - \sum^H (1-R_h) P_h \cdot \bar{Y}_{hR} + \sum^H (1-R_h) P_h \bar{Y}_{hNR}$$

$E(\bar{y}_R)$ ↘

$$= \sum^H P_h \cdot \bar{Y}_{hR} - \left[\sum^H (1-R_h) P_h (\bar{Y}_{hR} - \bar{Y}_{hNR}) \right]$$

$$\text{Bias} = \bar{Y}_R - \bar{Y} = \frac{\sum^H \sum^{N_{hR}} Y_{hRi}}{N_R} - \bar{Y} = \sum^H \frac{1}{N_R} \sum^{N_{hR}} Y_{hRi} - \bar{Y}$$

$$= \sum^H \frac{N_{hR} \cdot \bar{Y}_{hR}}{N_R} - \bar{Y} = \sum^H \frac{N_h R_h}{N \cdot R} \bar{Y}_{hR} - \bar{Y}$$

$$= \sum^H \frac{R_h}{R} P_h \cdot \bar{Y}_{hR} - \sum^H P_h \cdot \bar{Y}_{hR} + \left[\sum^H (1-R_h) P_h (\bar{Y}_{hR} - \bar{Y}_{hNR}) \right]$$

$$= \sum^H \left(\frac{R_h}{R} - 1 \right) P_h \cdot \bar{Y}_{hR} + \sum^H (1-R_h) P_h (\bar{Y}_{hR} - \bar{Y}_{hNR})$$

$$= \sum^H \frac{R_h - R}{R} P_h \cdot \bar{Y}_{hR} - \underbrace{\sum^H \frac{R_h - R}{R} P_h \cdot \bar{Y}_R}_{\text{this is 0}} + \sum^H (1-R_h) P_h (\bar{Y}_{hR} - \bar{Y}_{hNR})$$

$$= \underbrace{\sum^H \frac{R_h - R}{R} P_h (\bar{Y}_{hR} - \bar{Y}_R)}_{\text{This part can be estimated based on the observed data.}} + \underbrace{\sum^H (1-R_h) P_h (\bar{Y}_{hR} - \bar{Y}_{hNR})}_{\sum^H P_h (1-R_h)(\bar{Y}_{hR} - \bar{Y}_{hNR}) \text{ should be smaller than } (1-R)(\bar{Y}_R - \bar{Y}_{NR})}$$

Continuing the previous page:

$$\text{Bias} = \bar{Y}_R - \bar{Y}$$

$$= \underbrace{\sum^{H} \frac{R_h - R}{R} P_h (\bar{Y}_{hR} - \bar{Y}_R)}_{A} + \underbrace{\sum^{H} P_h (1-R_h)(\bar{Y}_{hR} - \bar{Y}_{hNR})}_{B}$$

Non-response missingness can be divided into 3 situations:

1. MCAR (missing completely at random).

   Then $R_h = R \quad \therefore A = 0$

   $\bar{Y}_{hR} = \bar{Y}_{hNR} \quad \therefore B = 0.$ $\Rightarrow$ no bias

2. MAR (missing at random)

   Conditioning on stratum, $\bar{Y}_{hR} = \bar{Y}_{hNR} \quad \therefore B = 0$

   Bias = A

3. MNAR (missing not at random)

   Bias = A + B

Finally get to the main point, what can poststratification do?

Poststratified estimation of a sample

$$\bar{y}_{ps} = \sum^{H} P_h \cdot \bar{y}_{Rh} \quad \leftarrow \text{ignore the response rate } (\bar{Y}_R = \sum^{H} \frac{R_h}{R} P_h \cdot \bar{Y}_{Rh})$$

and use respondents' mean in a stratum to represent the stratum mean

$$= \frac{\sum^{H} N_h \cdot \bar{y}_{Rh}}{N}$$

$$= \frac{\sum^{H} \sum^{n_{Rh}} N_h \cdot \bar{y}_{Rhi}/n_{Rh}}{\sum^{H} N_h} = \frac{\sum^{H} \sum^{n_{Rh}} \frac{n_R}{N} \frac{N_h y_{Rhi}}{n_{Rh}}}{\sum^{H} \frac{n_R}{N} N_h}$$

$$= \frac{\sum^{H} \sum^{n_{Rh}} w_{hi} y_{Rhi}}{\sum^{H} \sum^{n_{Rh}} w_{hi}} \quad \text{where } w_{hi} = \frac{N_h/N}{n_{Rh}/n_R}$$

$$\text{Bias}_{ps} = E(\bar{y}_{ps}) - \bar{Y} = E(\sum^{H} P_h \bar{y}_{Rh}) - \left[\sum^{H} P_h \cdot \bar{Y}_{hR} - \sum^{H} (1-R_h) P_h (\bar{Y}_{hR} - \bar{Y}_{hNR})\right]$$

$$= \sum^{H} P_h E(\bar{y}_{Rh}) - \sum^{H} P_h \cdot \bar{Y}_{Rh} + \sum^{H} (1-R_h) P_h (\bar{Y}_{hR} - \bar{Y}_{hNR})$$

$$\sum^{H} P_h E(\bar{y}_{Rh}) - \sum^{H} P_h \cdot \bar{Y}_{Rh} = 0$$

$$= \sum^{H} (1-R_h) P_h (\bar{Y}_{hR} - \bar{Y}_{hNR})$$

Comparing Bias_{ps} to Bias. A part is gone.

Poststratification is a reasonable practice under MAR assumption.

$\Rightarrow$ Both A and B are gone. $\therefore$ with MAR could have unbiased estimator with poststratification.

Following Point II

For large enough sample size, variance of poststratified mean will be less than or equal to SRS mean.

But this does NOT hold if the sample size shrinks:

$$V(\bar{y}_{ps}) \approx \frac{1}{n}\sum^{H} P_h S_h^2 \left(1 + \frac{1-P_h}{nP_h}\right)$$

$$\therefore V(\bar{y}_{ps}) - V(\bar{y}) \approx \frac{1}{n^2}\sum^{H} S_h^2 (1-P_h) - \frac{1}{n}\sum^{H} P_h (\bar{Y}_h - \bar{Y})^2$$

This could be positive or negative, so $V(\bar{y}_{ps})$ could be larger or smaller than $V(\bar{y})$

Following Point III

For bias indued by non-response, post stratified estimator always improve over SRS estimator if there is a relationship between

a) the probability of response and strata

$R_h \neq R$

b) the respondents' mean in the strata and the overall mean

$\bar{Y}_{hR} \neq \bar{Y}_h$

$\Rightarrow$ There is an A part in bias for the poststratification to remove

If MAR, B part is gone

$\therefore |B(\bar{y}_{r,ps}) - B(\bar{y}_r)| = |A|$ ← difference between the 2 biases is $|A|$

Point IV. Summing up Point II and Point III.

Mean Square Error = Bias^2 + Variance

Difference between the 2 MSE:

$$MSE(\bar{y}_{r,ps}) - MSE(\bar{y}_r) = Bias^2(\bar{y}_{r,ps}) - Bias^2(\bar{y}_r) + V(\bar{y}_{r,ps}) - V(\bar{y}_r)$$

$$= \underbrace{\frac{1}{n^2}\sum^{H} S_h^2 (1-P_h) - \frac{1}{n}\sum^{H} P_h (\bar{Y}_h - \bar{Y})^2}_{\text{above}} - \left[\underbrace{\sum^{H} \frac{R_h - R}{R} P_h (\bar{Y}_{hR} - \bar{Y}_R)}_{A}\right]^2$$

Point V. Raking (in contrast to all discussion on poststratification above)

Iteratively match with marginal distributions of auxiliary variables one at a time.

Better to see an example. Don't summarize here.

Point VI. poststratification and raking under GREG (generalized regression) framework

Both poststratification and raking fit within a framework — calibration estimation.

The idea is to recalibrate the survey weights d_i to new values w_i, where w_i satisfy the constraint $\sum w_i \vec{x}_i = \vec{X}$ where $\vec{X}$ is a vector of known population totals for auxiliary variable x;

(Poststratification can be think of one example of w_i, that works on top of the original weight $d_i = 1$ for all individuals)

if the original weights are d_i (E.g., based on selection probability)

$$w_i = d_i \left[1 + (\vec{X} - \hat{\vec{X}})' \vec{T}^{-1} \vec{x}_i \right]$$

where $\vec{T} = \sum d_i \vec{x}_i \vec{x}_i^T$ and $\hat{\vec{X}} = \sum d_i \vec{x}_i$

To break down the formula above:

- $\vec{X}$ observed auxiliary variables to be matched.
- $\hat{\vec{X}}$ estimated values of the auxiliary variables based on d_i
- $\vec{X} - \hat{\vec{X}}$ difference between actual X and d_i-based estimated X. if this is small, the d_i is regarded as already calibrated.
- $\vec{T}$ weighted cross product
- $\vec{T} \vec{x}_i$ can shrink of expand d_i depending on the sign of $\vec{X} - \hat{\vec{X}}$

Point VI.

The discussion above are under a design-based perspetive. Each unit i represents w_i units in the population.
Can also think of it as a model alternatively

First, in the above discussion $\bar{Y}$ is fixed — a population characteristics that the sample tries to capture.
But can also conceptualize $\bar{Y}$ as a random variable — Bayesian

$$\bar{Y} = \sum_h^H P_h \cdot \bar{Y}_h = \sum_h^H P_h \left(\frac{n_h}{N_h} \bar{y}_h + \frac{N_h - n_h}{N_h} \cdot \bar{Y}_{h,ns} \right)$$

$\bar{y}_h$: sample data/observation
$\bar{Y}_{h,ns}$: non-sampled. need to work out a posterior distribution

$y_{hi} \mid \mu_h, \sigma_h^2, z_i = h \sim N(\mu_h, \sigma_h^2)$ allow each stratum to have its own mean and variance

$P(\mu_h, \log \sigma_h^2) \propto 1$ flat prior

Bayes rule is used here to flip the probability to become the probability of the underlying conditions on the observed.

$\Rightarrow$ posterior distribution (deduction not shown here)

$$\mu_h \mid y, z \sim N\left(\bar{y}_h, \frac{\sigma_h^2}{n_h}\right)$$
$$\sigma_h^2 \mid y, z \sim \text{Inverse } \chi^2 (n_h - 1, s_h^2)$$

posterior distributions of μ_h and σ_h^2 are not enough. Since we are dealing with a finite population situation here, need to leverage μ_h and σ_h^2 and further work out a distribution of $\bar{Y}_{h,ns}$ conditioning on data.

we know that $\bar{Y}_{h,ns} \mid \mu_h, \sigma_h^2 \sim N(\mu_h, \frac{\sigma_h^2}{N_{h,ns}})$. observing data doesn't change this distribution $\bar{Y}_{h,ns} \mid \mu_h, \sigma_h^2, y, z \sim N(\mu_h, \frac{\sigma_h^2}{N_{h,ns}})$

Remove μ_h from condition

$$E(\bar{Y}_{h,ns} \mid \sigma_h^2, y, z) = E_{\mu_h}\left(E(\bar{Y}_{h,ns} \mid \sigma^2, y, z, \mu_h) \mid \sigma_h^2, y, z \right)$$
$$= E_{\mu_h}(\mu_h \mid \sigma_h^2 \, y \, z) = \bar{y}_h$$

$$V(\bar{Y}_{h,ns} \mid \sigma_h^2, y, z) = E\left(V(\bar{Y}_{h,ns} \mid \sigma_h^2 y z \mu_h) \mid \sigma_h^2 y z\right) + V\left(E(\bar{Y}_{h,ns} \mid \sigma_h^2 y z \mu_h) \mid \sigma_h^2 y, z\right)$$
$$= E\left(\frac{\sigma_h^2}{N_{h,ns}} \mid \sigma_h^2 \, y \, z\right) + V(\mu_h \mid \sigma_h^2 y z)$$
$$= \frac{\sigma_h^2}{N_{h,ns}} + \frac{\sigma_h^2}{n_h} = \frac{\sigma_h^2}{(1 - \frac{n_h}{N_h}) n_h}$$

Bayes rule is no long applied in this part.
μ_h and σ_h^2 are like a helper bridge.
With their help, work out a distribution of the rest of the finite population conditioned on the sample

continue to the next page

Remove σ_h^2 from condition

$$E(\bar{Y}_{h,ns} \mid y, z) = \underset{\sigma_h^2}{E}\left(E(\bar{Y}_{h,ns} \mid y, z, \sigma_h^2) \mid y, z \right) = \underset{\sigma_h^2}{E}(\bar{y}_h \mid y, z) = \bar{y}_h$$

$$V(\bar{Y}_{h,ns} \mid y, z) = E(V(\bar{Y}_{h,ns} \mid y, z, \sigma_h^2) \mid y, z) + V(E(\bar{Y}_{h,ns} \mid y, z, \sigma_h^2) \mid y, z)$$

$$= E\left(\frac{\sigma_h^2}{(1-\frac{n_h}{N_h})n_h} \mid y, z\right) + \cancel{V(\bar{y}_h \mid y, z)}^{\,0}$$

$\bar{y}_h$ doesn't have a variance conditioning on data.

$$= \frac{n_h - 1}{n_h - 3} \frac{S_h^2}{(1-\frac{n_h}{N_h})n_h} \leftarrow \text{sample variance}$$

$$E(\bar{Y}_h \mid y, z) = E\left(\frac{n_h}{N_h}\bar{y}_h + \frac{N_h - n_h}{N_h}\bar{Y}_{h,ns} \mid y, z\right) = \bar{y}_h$$

$V(\bar{Y}_h \mid y, z)$ only comes from $\bar{Y}_{h,ns}$

$$= V\left(\frac{n_h}{N_h}\bar{y}_h + \frac{N_h - n_h}{N_h}\bar{Y}_{h,ns} \mid y, z\right)$$

$$= \left(\frac{N_h - n_h}{N_h}\right)^2 \cdot \frac{n_h - 1}{n_h - 3} \frac{S_h^2}{(1-\frac{n_h}{N_h})n_h}$$

$$= \frac{N_h - n_h}{N_h} \frac{n_h - 1}{n_h - 3} \frac{S_h^2}{n_h}$$

$$E(\bar{Y} \mid y, z) = E\left(\sum^H P_h \bar{Y}_h \mid y, z\right) = \sum^H P_h \bar{y}_h$$

$$V(\bar{Y} \mid y, z) = V\left(\sum^H P_h \bar{Y}_h \mid y, z\right) = \sum^H P_h^2 \frac{n_h - 1}{N_h - 3} \frac{N_h - n_h}{N_h} \frac{S_h^2}{n_h}$$

+ covariance → 0

Extend the above discussion:

$$y_{hi} \mid \mu_h, \sigma_h^2 z_i = h \sim N(\mu_h, \sigma_h^2)$$

If the prior of μ_h is not flat $\mu_h \sim N(\lambda, \tau^2)$

still $p(\log \sigma_h^2) \propto 1$

$$E(\bar{Y}_h \mid y, z, \sigma_h^2, \tau^2)$$

Module 10

Theoretical differentiation

2 Situations:

1) model as a mean to an end. The goal is to estimate descriptive statistics of the population. Models have two usages.
   a. One is model-assisted inference:
      models are used to improve efficiency of design-based approach
   b. The other is model-based inference:
      take survey design into account. When use of model is required for inference bc need to borrow information across different elements of the data.

2) model is the end. The model itself is the focus of inference. Now the question is wether to incorporate design into the inference about the model.

Model-assisted inference Point I. Where is model?

Recall GREG. Reweight the original weights d_i to w_i

$$w_i = d_i\left[1 + (\vec{X} - \hat{\vec{X}})' \vec{T}^{-1} \vec{x}_i\right]$$

There is a model behind this formula — why the calculation is this way

Consider using w_i for estimating population total.

$$\hat{T}_w = \sum^n w_i y_i = \sum^n d_i\left[1 + (\vec{X} - \hat{\vec{X}})' \vec{T}^{-1} \vec{x}_i\right] y_i = \sum^n d_i y_i + \sum^n d_i (\vec{X} - \hat{\vec{X}})' \vec{T}^{-1} \vec{x}_i \cdot y_i$$

$$= \sum^n d_i y_i + \sum^n d_i \vec{X}' \vec{T}^{-1} \vec{x}_i y_i - \sum^n d_i \hat{\vec{X}}' \vec{T}^{-1} \vec{x}_i y_i$$

$$= \sum^n d_i y_i + \vec{X}'\underbrace{\left(\sum^n d_i \vec{x}_i \vec{x}_i^T\right)^{-1}\left(\sum^n d_i \vec{x}_i y_i\right)}_{\text{by definition } \hat{\beta}_w} + \hat{\vec{X}}'\underbrace{\left(\sum^n d_i \vec{x}_i \vec{x}_i^T\right)^{-1}\left(\sum^n d_i \vec{x}_i y_i\right)}$$

$$= \underbrace{\sum^n d_i y_i}_{\text{standard design-based estimator.}} + (\vec{X}' - \hat{\vec{X}}')\underbrace{\hat{\beta}_w}_{\text{model}}$$

∴ By changing d_i to w_i, GREG essentially corrects the original estimator based on d_i by adding an extra part $(\vec{X} - \hat{\vec{X}})\vec{\beta}_w$. The model is between y (variable of interest) and x (auxiliary variables). The model part makes correction if the original weights d_i are not sufficient to bring x to the population and there is a relationship between x and y

Technically, we only need aggregate $\vec{X}$ and the sample to make the above GREG correction. But assuming that we have information of X on the entire population (which is not completely impossible), then we can rewrite the above population total estimate.

$$\hat{T}_w = \sum^n d_i y_i + (\vec{X} - \hat{\vec{X}})\vec{\beta}_w$$

$$= \sum^n d_i y_i + \left(\sum^N \vec{x}_i - \sum^n d_i \vec{x}_i\right)\vec{\beta}_w = \sum^n d_i y_i + \sum^N \vec{x}_i \cdot \vec{\beta}_w - \sum^n d_i \vec{x}_i \vec{\beta}_w$$

$$= \sum^N \vec{x}_i \vec{\beta}_w + \underbrace{\sum^n d_i (y_i - \vec{x}_i \vec{\beta}_w)}_{\hat{y}_i^w} = \sum^N y_i^w + \underbrace{\sum^n d_i (y_i - \hat{y}_i^w)}_{=0 \text{ by definition of regression.}}$$

$\hat{y}_i^w$ is based on the weighted linear regression

$$\therefore \hat{T}_w = \sum^n d_i y_i + (\vec{X}' - \hat{\vec{X}}')\vec{\beta}_w = \sum^N \vec{x}_i \vec{\beta}_w$$

Two ways of expressing the GREG-weighted estimated total.

Either use the model ($\vec{\beta}_w$) to adjust the discrepancy on X $(\vec{X}' - \hat{\vec{X}}')$

Or apply the model ($\vec{\beta}_w$) directly on $\vec{x}_i$ of the entire population units

Either way $\hat{T}_w$ can be viewed as a regression estimator (based on x and y)

A general form of regression estimator is

$$\hat{T}_R = \sum^N \hat{y}_i + \underbrace{\sum^n d_i (y_i - \hat{y}_i)}_{\text{model}}$$

Don't equal to 0 as above bc the model is unweighted so $\sum^n (y_i - \hat{y}_i) = 0$ and $\sum^n d_i (y_i - \hat{y}_i) \neq 0$

Comparing to above, weights are only used to adjust the difference between y_i and $\hat{y}_i$, but not used for constructing $\hat{y}_i$ in the first place

Advantages comparing to $\hat{T}_w$: ① $\hat{y}_i$ is more stable than $\hat{y}_i^w$ → reduce variance

② Can still correct for model misspecification $y_i - \hat{y}_i$ → but increase variance

$$T = \sum^N (\hat{y}_i + e_i) = \sum^N \hat{y}_i + \sum^N e_i$$

Is $\hat{T}_R$ biased (deviate from T)? ⇐ $\sum^N e_i \neq \sum^N d_i (y_i - \hat{y}_i)$

Model-assisted inference continue

The previous discussion talks about how GREG can be framed as a regression estimator — An estimator that is based on the relationship between auxiliary variable's $\vec{x}$ and variable of interest y ($\vec{\beta}_w$). Recall that in GREG: $w_i = d_i[1 + (\vec{X} - \hat{\vec{X}})' T^{-1} \vec{x}_i]$. One example to interpret w_i and d_i is to view d_i as selection weights and w_i as adjustment on top of the selection weight.

It seems straightforward to see w_i-based estimator as a model between y and x because w_i is achieved by using x_i.

However, not only can w_i-based estimator be framed as a model, the d_i-based estimator can also be framed as a model, present below:

How to frame HT-estimator (d_i-based estimator) as a regression model:

with $d_i = \frac{1}{\pi_i}$ reverse of selection probability, HT Estimator $= \sum^n d_i y_i = \sum^n \frac{y_i}{\pi_i}$

Since π_i is only information, consider a regression that models the relationship between y_i and π_i $\quad E(Y_i) = \beta \pi_i$ $\quad V(Y_i) = \sigma^2 \pi_i^2$ > A weighted least square regression. (no intercept)

That is: $Y_i \sim N(\beta \pi_i, \sigma^2 \pi_i^2)$

Equivalent to: $Y_i = \beta \cdot \pi_i + \epsilon_i \quad \epsilon_i \sim N(0, \pi_i^2 \sigma^2)$

Assuming a helper variable $z_i = Y_i / \pi_i \quad \therefore E(z_i) = \beta \quad V(z_i) = \sigma^2$

given a sample of realization of z_i,

mean of this sample $= \hat{\beta}$

$$\sum^n \frac{z_i}{n} = \sum^n \frac{Y_i}{\pi_i n} = \frac{1}{n} \cdot \text{HT estimator} = \hat{\beta}$$

$\therefore$ HT-estimator $= n \cdot \hat{\beta}$ ← $\hat{\beta}$ is based on the model between y_i and π_i

Recall that if a regression model is used to assist the estimate, then the regression estimate (of population total) can be written as

$\hat{T}_R = \sum^N \hat{y}_i + \sum^n d_i (y_i - \hat{y}_i)$. Applying this formula here:

$$\hat{T}_R = \sum^N \beta \cdot \pi_i + \sum^n \frac{1}{\pi_i}(y_i - \beta \pi_i) = \sum^N \beta \cdot \pi_i + \sum^n \frac{y_i}{\pi_i} - \sum^n \beta = \sum^n \frac{y_i}{\pi_i} + \beta(\cancel{\sum^N \pi_i} \to 0 - n)$$

$$= \sum^n \frac{y_i}{\pi_i} = \text{HT-estimator}$$

This demonstrates how HT-estimator (the d_i-based estimator) can be written as a regression estimator: $Y_i \sim \pi_i$ regression estimator $\longleftrightarrow$ HT estimator

Thus, if the model captures the data well, the HT estimator works.

if the model doesn't match with the data at all, the regression estimator would be a bad estimator, so will HT estimator.

Model-based inference. Point II

The discussion above, whether it's GREG or regression estimator in general, or the model for HT estimator, mainly deals with weights.

Recall that weights are survey design are not the same thing! The same weights can result from many different designs.

Model-based inference is about incorporating survey-design (rather than just weights) into population inference.

Take stratification as an example: In the above discussion, the estimator $\hat{T}_W$ or $\hat{T}_R$ is an estimator for the population total T. There is only one overall T for the entire population. Following the perspetive of stratification, we could allow T_h for each stratum and then sum up different strata to get to the population estimate. (Rather than having one random variable, have H random variables)

2 **Stratification** — different strata have different mean and variance, and thus different models

(In the previous module) Under a design-based approach

$$\bar{Y} = \sum_{h}^{H} \frac{N_h}{N} \cdot \bar{Y}_h \quad \leftarrow \bar{Y}_h/\mu_h \text{ is presented as a true value, which } \bar{y}_h \text{ tries to capture.}$$

$$\hat{\bar{Y}} = \sum_{h}^{H} \frac{N_h}{N} \bar{y}_h$$

Under a Bayesian model-based approach: $\bar{Y}_h/\mu_h$ is framed as a random variable

(btw this is a new element comparing to GREG etc. discussion above. Model doesn't have to be Bayesian, but this is how it is discussed here)

$$y_{hi} \mid \mu_h, \sigma_h^2 \sim N(\mu_h, \sigma_h^2) \quad + \text{ flat prior } p(\mu_h, \sigma_h^2) \propto 1$$

posterior distribution $\mu_h \mid y_h \sim t_{n_h-1}\left(\bar{y}_h, \frac{s_h^2}{n_h}\right)$ where s_h^2 is the sample variance

[Where are "models"?
Recall that "model" essentially means assumption of distributions. So all these distributions, from prior, to y_{hi}, to posterior, are models.]

We already saw this in the previous module.

Utilizing the posterior distribution of $\mu_h \mid y_h$, can now calculate $E(\bar{Y} \mid y, z)$

1) given the posterior distribution $\mu_h \mid y_h \rightarrow \bar{Y}_{h,ns} \sim t_{n_h-1}\left(\bar{y}_h, (1-f_h)\frac{s_h^2}{n}\right)$ ★
   Draw $\hat{\bar{Y}}_{h,ns}$ values
2) compute $\hat{\bar{Y}}_h = \frac{n_h}{N_h}\bar{y}_h + \frac{N_h - n_h}{N_h}\hat{\bar{Y}}_{h,ns}$
3) repeat 1000 times $\Rightarrow E(\hat{\bar{Y}}_h)$
4) $\bar{Y} = \sum^{H} P_h \cdot \bar{Y}_h \approx \sum^{H} P_h E(\hat{\bar{Y}}_h)$

Essentially one model for each stratum. That's incorporating design (strata) into model-based inference.

2 Clustering — random effect model

$$y_{ki} \mid \mu, \sigma_k^2, b_k \sim N(\mu + b_k, \sigma_k^2)$$

$$b_k \sim N(0, \tau^2)$$

Recall multilevel analysis:

$$y_{ki} = \beta_{0k} + \varepsilon_{ki} \qquad \varepsilon_{ki} \sim N(0, \sigma^2)$$

$$\beta_{0k} = \beta_0 + u_{0k} \qquad u_{0k} \sim N(0, \sigma_0^2)$$

Assuming that σ_k^2 and τ^2 are known

1) the posterior distribution of $\bar{Y}_{k,ns} \sim N(\tilde{y}_k, v_k^2)$

where for the sampled clusters:

$$\tilde{y}_k = \frac{n_k \cdot \tau^2}{n_k \tau^2 + \sigma_k^2} \cdot \bar{y}_k + \left(1 - \frac{n_k \tau^2}{n_k \tau^2 + \sigma_k^2}\right) \cdot \boxed{\frac{\sum_k^k \frac{n_k}{n_k \tau^2 + \sigma_k^2} \bar{y}_k}{\sum_k^k \frac{n_k}{n_k \tau^2 + \sigma_k^2}}} \quad \tilde{y}$$

← As for why it's like this, not shown here

$$v_k = \left(1 - \frac{n_k}{N_k}\right) \left(\sum^k \frac{n_k}{n_k \tau^2 + \sigma_k^2}\right)^{-1}$$

and for the non-sampled clusters:

$$\tilde{y}_k = \tilde{y} \quad \text{and} \quad v_k = \left(\sum^k \frac{n_k}{n_k \tau^2 + \sigma_k^2}\right)^{-1}$$

can draw $\bar{Y}_{k,ns}$ from $N(\tilde{y}_k, v_k^2)$

2) compute $\hat{\bar{Y}}$ from $\begin{cases} \text{for sampled cluster: } n_k \text{ sampled, mean } \bar{y}_k + (N_k - n_k) \text{ non-sampled, mean } \bar{Y}_{k,ns} \\ \text{for non-sampled cluster: all units } N_k \text{ non-sampled } \bar{Y}_{k,ns} \end{cases}$

$$\hat{\bar{Y}} = \frac{1}{N} \left[\sum^k \left[n_k \cdot \bar{y}_k + (N_k - n_k) \bar{Y}_{k,ns} \right] + \sum_{k=k+1}^{M} N_k \cdot \bar{Y}_{k,ns} \right]$$

3) Repeat $\Rightarrow E(\hat{\bar{Y}})$

2 Unequal probability of selection.

a. If categories of selection probabilities.

This can be viewed as a stratification problem: Each stratum refer to the elements of the population with equal probability of selection.

Intuitive example: if men and women have different selection probabilities, then $\frac{n_♀}{n} \neq \frac{N_♀}{N}$ and $\frac{n_♂}{n} \neq \frac{N_♂}{N}$. A way out of this would be to calculate $\bar{y}_♀$ and $\bar{y}_♂$ separately, then combine the two means proportional to their population portion ⇒ stratification.

$$\begin{aligned}
E(\bar{Y} \mid y, z) &= E\left(\sum^{H} P_h \cdot \bar{Y}_h \mid y, z\right) \\
&= \sum^{H} P_h E(\bar{Y}_h \mid y, z) \\
&\quad \text{(since } E(\bar{Y}_h \mid y, z) = \bar{y}_h\text{)} \\
&= \sum^{H} P_h \cdot \bar{y}_h \\
&= \sum^{H} \frac{N_h}{N} \cdot \frac{\sum^{n_h} y_i}{n_h} \\
&= \frac{1}{N} \sum^{H} \frac{N_h}{n_h} \sum^{n_h} y_i
\end{aligned}$$

What's "model" about this?

$E(\bar{Y}_h \mid y, z)$ exists bc $\bar{Y}_h \mid y, z$ follows a (assumed) distribution. This is model!

Especially if taken a Bayesian perspective, to know $\bar{Y}_h \mid y, z$, we need $\bar{Y}_{h,ns} \mid y, z$, which depends on $\mu_h \mid y, z$ (posterior distribution)

b. If rather than categories of selection probabilities, each individual has an unique selection probability:

Recall that HT-estimator can be expressed as a regression estimator based on the regression model between π_i and y_i

$$y_i \sim N(\pi_i \cdot \beta, \pi_i^2 \sigma^2)$$

It's possible to generalize this simple linear model between π_i and y_i to, e.g., $\pi_i \beta_1 + \beta_0$ or $\pi_i \beta_1 + \pi_i^2 \beta_2 + \beta_0$ (of course, the estimator would not be HT estimator)

In general: $y_i \sim N(g(\pi_i), \pi_i^k \sigma^2)$

How to predict $\bar{Y}$:

1) Given a sample of data on y_i and π_i, model relationship between y_i and π_i, for example, $y_i = \beta_0 + \pi_i \beta_1 + \pi_i^2 \beta_2$

↓ Based on the formula of linear regression $y_i \sim N(\vec{x}_i \cdot \vec{\beta}, \sigma^2)$

$$\hat{\beta} = (X^T X)^{-1} X^T \cdot y \quad \text{and} \quad SE(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

$\beta_1 \sim N(\beta_1, SE(\beta_1)^2)$

same for β_2 and β_0

2) draw $\hat{\beta}$ from these estimated distribution of βs

3) π_i is known for all units in the population.

for all units in population $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \pi_1 + \hat{\beta}_2 \cdot \pi_2^2$

4) Sum up all $\hat{Y}_i$ for unsampled units and y sample ⇒ $\bar{Y}$

- Comparing the situation a. and b. above. Notice how the information of π_i (selection probability) is used completely differently. In a., π_i is used for constructing strata, but π_i is not incorporated in the "models" directly. In contrast, in b., π_i are at the individual level and π_i is directly used for building the "model" for Y

Point III.

The above discussions are all about situation 1), in which (descriptive) statistics of the population is the goal. E.g., population mean or total — a scalar quantity
Now, consider situation 2), where the goal of inference is model.

Consider a simple example of no intercept linear model.

$$Y_i = B \cdot X_i + \varepsilon_i \quad \varepsilon_i \sim N(0, \sigma^2)$$

In terms of finite population inference, we can fit the entire population data to the linear model to obtain the population slope.

$$B = \frac{\sum^N Y_i X_i}{\sum^N X_i^2}$$

why Y_i is a random variable?

superpopulation β → finite population B

- Note that B would be an unbiased estimate of β if the modelled relationship between Y_i and X_i is the true relationship:

In superpopulation — $Y_i = \beta \cdot X_i + \varepsilon_i \quad \varepsilon_i \sim N(0, \sigma^2)$

$$E(Y_i) = \beta X_i$$

$$B = \frac{\sum^N Y_i X_i}{\sum^N X_i^2} \qquad E_{Y|X}(B) = E_{Y|X}\left(\frac{\sum^N Y_i X_i}{\sum^N X_i^2}\right) = \frac{\sum^N E(Y_i) X_i}{\sum^N X_i^2} = \frac{\sum^N X_i^2 \beta}{\sum^N X_i^2} = \beta$$

☆ why do we need a step of super population here?

- However, if the model is misspecified, say, in the superpopulation the true relationship is

$$Y_i = \beta X_i^2 + \varepsilon_i \quad \varepsilon_i \sim N(0, \sigma^2)$$

$$E(Y_i) = \beta X_i^2$$

$$E(B) = E\left(\frac{\sum^N Y_i X_i}{\sum^N X_i^2}\right) = \frac{\sum^N E(Y_i) X_i}{\sum^N X_i^2} = \frac{\sum^N X_i^3}{\sum^N X_i^2} \beta$$

Then B no longer estimates the true parameter governing the quadratic relationship. but the best linear approximation (be residuals will sum to 0 ← regression's definition)

Now we consider the realistic situation, Rather than dealing with how B infers β, we consider how to infer B from a sample. We have no information of population, but only a sample, but weights are available.

$$\sum^n w_i (y_i - \hat{B}_w x_i) x_i = 0 \quad \therefore \hat{B}_w = \frac{\sum^n w_i y_i x_i}{\sum^n w_i x_i^2}$$

☆ if w_i is independent from $Y|X$, then whether w_i is included here doesn't matter?

$\hat{B}_w$ is an unbiased estimation of B:

If the model is correct and sampling is independent of the conditional distribution of $Y|X$, then we do not necessarily need to consider w_i. The unweighted estimate will already be consistent for B:

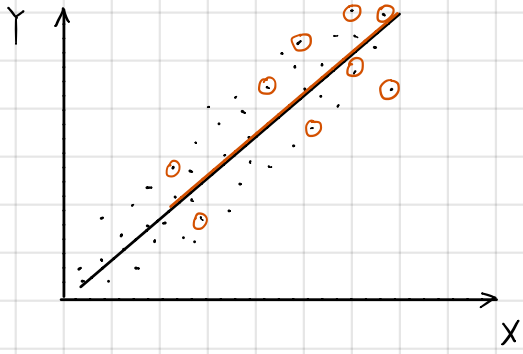
$$\hat{B}_{uw} = \frac{\sum^n y_i x_i}{\sum^n x_i^2} = \frac{\sum^N z_i Y_i X_i}{\sum^N z_i X_i^2}$$

$$E_Y(\hat{B}_{uw}) = \frac{\sum^N z_i E(Y_i) X_i}{\sum^N z_i X_i^2} = \frac{\sum^N z_i X_i B X_i}{\sum^N z_i X_i^2} = B \frac{\sum^N z_i X_i^2}{\sum^N z_i X_i^2}$$

☆ why need $E_Z\left(E_Y(\hat{B}_{uw})\right)$?

Continuing the previous page.

An intuitive example about when we need to consider w for $\hat{B}$

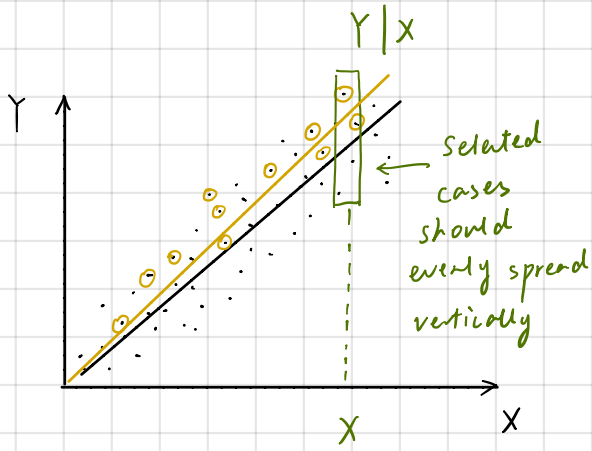


In the population, the true relationship between X and Y is like this (B captures β)

If the selection mechanism $I (\to w_i)$ is skewed, and that we ended up with these orange cases, that's okay!

→ the estimated $\hat{B}$ will be very similar to B.

That's why we can ignore the sampling mechanism (i.e., weights) when estimating coefficient.



The above discussion simplies the situation a little bit.

Can we unconditionally ignore the sampling mechanism? The answer is no, we can ignore sampling mechanism (weight) only if $I \perp Y|X$.

That is to say, it's okay if I is related to Y, like the example above, cases with higher valued Y are selected more. But it's not okay if I is related to $Y|X$, see example in the left:

If I is related to $Y|X$ (meaning that the selection doesn't spread evenly around the line in the vertical direction and we ended up with the yellow cases,

→ the estimated $\hat{B}$ will be biased.

(no intercept model)

Thus, 2 conditions need to be satisfied for us to ignore w_i in estimating B.

1) the model is correct (See previous page)
2) the sampling mechanism is independent of $Y|X$ (demonstrated above)