November 3, 2025

Stat 440 Handout on Survey Regression

Topic: 'Model-Assisted' estimation of y-totals t_y by

survey-weighted regression (least-squares linear model fitting),

(y_i, x_i), $i \in \mathcal{S}$ a survey sample from population $\mathcal{U} = \{1, 2, \ldots, N\}$

Estimation target t_y with total t_x known

'model' does **not** relate individual y_i's to individual x_i's
but estimates coefficients (a, b) minimizing $\sum_{i=1}^{N} (y_i - a - bx_i)^2$

and then defines **regression estimator** $\widehat{t}_y^{\,\mathrm{reg}} = \widehat{a} + \widehat{b}\, t_x$
from the sample-based estimators $\widehat{a}, \widehat{b}$

Introduced this last Thursday, for $\mathcal{S}$ a SRS of size n; sketched here, deriving large-sample variance formula for $\hat{t}_y^{\text{reg}}$

The argument is given 3 times

- briefly for simple linear regression from SRS samples
- generalized to survey design with inclusion probabilities π_i
- still more generally, it can be done with only notational changes usng survey weights when the predictive variables $\mathbf{x}_i$ with known survey totals are vectors; this is done in the least two slides of this handout.

Knowledge of t_x is crucial, and (a, b) are descriptive population characteristics: the 'best line fitted to y_i in terms of x_i'

Basic Equations in Unweighted or SRS Case

Population: $\min_{a,b} \sum_{i=1}^{N} (y_i - a - bx_i)^2$ implies

$$\sum_{i=1}^{N} \begin{pmatrix} 1 \\ x_i \end{pmatrix} (y_i - a - b\,x_i) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$a = \bar{y}_U - b\bar{x}_U \quad , \quad b = \sum_{i=1}^{N} (x_i - \bar{x}_U) y_i \Big/ \sum_{i=1}^{N} (x_i - \bar{x}_U)^2$$

$\hat{a}, \hat{b}$ defined as minimizers of $\sum_{i \in \mathcal{S}} (y_i - a_0 - b_0 x_i)^2$ in (a_0, b_0)

$$\hat{a} = \bar{y}_{\mathcal{S}} - \hat{b}\bar{x}_{\mathcal{S}} \quad , \quad \hat{b} = \sum_{i \in \mathcal{S}} (x_i - \bar{x}_{\mathcal{S}}) (y_i - \bar{y}_{\mathcal{S}}) \Big/ \sum_{i \in \mathcal{S}} (x_i - \bar{x}_{\mathcal{S}})^2$$

Using known t_x, $\quad \hat{t}_y^{\text{reg}} = N\,(\hat{a} + \hat{b}\,\bar{x}_U)$

$$\frac{1}{N}\,\hat{t}_y^{\text{reg}} = \hat{a} + \hat{b}\,\bar{x}_U + \frac{1}{n}\sum_{\mathcal{S}}\{y_i - \hat{a} - \hat{b}x_i\} = \bar{y}_{\mathcal{S}} + \hat{b}(\bar{x}_U - \bar{x}_{\mathcal{S}})$$

$$= \bar{y}_{\mathcal{S}} - \bar{y}_U + \bar{y}_U - b\,(\bar{x}_{\mathcal{S}} - \bar{x}_U) + (b - \hat{b})\,(\bar{x}_{\mathcal{S}} - \bar{x}_U)$$

$$= (\bar{y}_{\mathcal{S}} - b\bar{x}_{\mathcal{S}} - a) + \bar{y}_U + (b - \hat{b})\,(\bar{x}_{\mathcal{S}} - \bar{x}_U)$$

For large N, n, 3rd term is remainder of order $\frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{n}}$

so $\hat{t}_y^{\text{reg}} - t_y \approx \frac{N}{n} \sum_{\mathcal{S}} (y_i - a - bx_i)$ has large-sample Variance

$\approx N^2\,(\frac{1}{n} - \frac{1}{N})\,s^2_{y-a-bx,U}$ estimated by $N^2\,(\frac{1}{n} - \frac{1}{N})\,s^2_{y-\hat{a}-\hat{b}x,\mathcal{S}}$

Now repeat similar steps in survey-weighted case:

Basic Equations in $1/\pi_i$ Weighted Case

Population a, b as before, $\pi_i =$ inclusion pr. $P(i \in \mathcal{S})$. Denote

$$\widehat{N} = \sum_{\mathcal{S}} \frac{1}{\pi_i}, \quad \bar{y}_{\mathcal{S},\pi} = \sum_{\mathcal{S}} \frac{y_i}{\pi_i}/\widehat{N}, \quad \bar{x}_{S,\pi} = \sum_{\mathcal{S}} \frac{x_i}{\pi_i}/\widehat{N}$$

$\widehat{a}, \widehat{b}$ defined as minimizers of $\sum_{\in \mathcal{S}} \frac{1}{\pi_i}(y_i - a_0 - b_0\, x_i)^2$ in (a_0, b_0)

giving $\sum_{\mathcal{S}} \frac{1}{\pi_i} \binom{1}{x_i} (y_i - a_0 - b_0\, x_i) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ implying

$$\widehat{a} = \bar{y}_{\mathcal{S},\pi} - \widehat{b}\bar{x}_{\mathcal{S},\pi}, \quad \widehat{b} = \sum_{\mathcal{S}} \frac{y_i}{\pi_i}(x_i - \bar{x}_{\mathcal{S},\pi}) \Big/ \sum_{\mathcal{S}} \frac{1}{\pi_i}(x_i - \bar{x}_{\mathcal{S},\pi})^2$$

$\bar{x}_U = \frac{t_x}{N}$ known, $\quad \widehat{t}_y^{\,\mathrm{reg}} = \widehat{N}\,(\widehat{a} + \widehat{b}\,\bar{x}_U) = \widehat{N}\,(\bar{y}_{\mathcal{S},\pi} - \widehat{b}(\bar{x}_{\mathcal{S},\pi} - \bar{x}_U))$

$$\frac{1}{\widehat{N}}\,\hat{t}_y^{\,\mathrm{reg}} \;=\; \bar{y}_{\mathcal{S},\pi} - \bar{y}_U + \bar{y}_U - b\,(\bar{x}_{\mathcal{S},\pi} - \bar{x}_U) + (b - \hat{b})\,(\bar{x}_{\mathcal{S},\pi} - \bar{x}_U)$$

$$= \;(\bar{y}_{\mathcal{S},\pi} - b\bar{x}_{\mathcal{S},\pi} - a) \;+\; \bar{y}_U \;+\; (b - \hat{b})\,(\bar{x}_{\mathcal{S},\pi} - \bar{x}_U)$$

Same conclusion as before, 3rd term is product of two small differences, and for large N, n

$\hat{t}_y^{\,\mathrm{reg}} - \widehat{N}\,\bar{y}_U \;\approx\; \sum_{\mathcal{S}} \frac{1}{\pi_i}\,(y_i - a - bx_i)$ has large-sample Variance

$$\approx\; \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\pi_{i,j} - \pi_i\,\pi_j}{\pi_i\,\pi_j}\,(y_i - a - bx_i)\,(y_j - a - bx_j)$$

generally of order N^2/n as in SRS case, and estimated by

$$\sum_{i\in\mathcal{S}} \sum_{j\in\mathcal{S}} \frac{\pi_{i,j} - \pi_i\,\pi_j}{\pi_{ij}\,\pi_i\,\pi_j}\,(y_i - \hat{a} - \hat{b}x_i)(y_j - \hat{a} - \hat{b}x_j)$$

Do two further steps:

(i) Compare for SRS the regression variance to that of $N\bar{y}_{\mathcal{S}}$

(ii) Give example of $\hat{t}_y^{\mathrm{reg}}$ formulas in Poisson sampling case.

Next extend in the last two slides to multivariable survey-weighted regression, called `GREG`. The argument is almost exactly the same, showing an approximate form for the regression estimator and its variance based on multivariate weighted least squares.

Regression versus Horvitz-Thompson SRS Variance Comparison

Let $\widehat{\rho} = \widehat{Cor}(x,y) = \frac{1}{n-1}\sum_{\mathcal{S}}(x_i - \bar{x}_{\mathcal{S}})(y_i - \bar{y}_{\mathcal{S}})\Big) \Big/ \Big[s_{x,\mathcal{S}}\, s_{y,\mathcal{S}}\Big]$

and recall $\widehat{b} = \frac{1}{n-1}\sum_{\mathcal{S}}(x_i - \bar{x}_{\mathcal{S}})(y_i - \bar{y}_{\mathcal{S}})\Big)/s^2_{x,\mathcal{S}} = \widehat{\rho}\, s_{y,\mathcal{S}}/s_{x,\mathcal{S}}$

$$\widehat{\mathsf{Var}}(\hat{t}_y^{\,\mathtt{reg}}) = \frac{N(N-n)}{n}\cdot\frac{1}{n-1}\sum_{\mathcal{S}}\Big[y_i - \bar{y}_{\mathcal{S}} - \widehat{b}(x_i - \bar{x}_S)\Big]^2$$

$$= \frac{N(N-n)}{n}\cdot\Big[s^2_{y,\mathcal{S}} - 2\,\widehat{b}\,\widehat{\rho}\,s_{x,\mathcal{S}}\,s_{y,\mathcal{S}} + \widehat{b}^2\,s^2_{x,\mathcal{S}}\Big]$$

$$= \frac{N(N-n)}{n}\,s^2_{y,\mathcal{S}}\Big[1 - 2\,\widehat{\rho}^2 + \widehat{\rho}^2\Big] = \widehat{\mathsf{Var}}(N\,\bar{y}_{\mathcal{S}})\cdot(1-\widehat{\rho}^2)$$

Regression Estimator Variance under Poisson Sampling

Poisson sampling has $\pi_{i,j} = \pi_i \cdot \pi_j$ if $i \neq j$. So

$$\widehat{\mathsf{Var}}(\hat{t}_y^{\mathsf{reg}}) = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}} \frac{\pi_{i,j} - \pi_i \, \pi_j}{\pi_{ij} \, \pi_i \, \pi_j} (y_i - \hat{a} - \hat{b} x_i)(y_j - \hat{a} - \hat{b} x_j)$$

$$= \sum_{i \in \mathcal{S}} \frac{1 - \pi_i}{\pi_i^2} (y_i - \hat{a} - \hat{b} x_i)^2$$

Survey-Weighted Multivariate Regression (GREG)

$\mathbf{x}_i \in \mathbb{R}^p$ non-constant predictors with known (vector) totals $t_{\mathbf{x}}$.

Pop'n coefficients a and $\mathbf{b} \in \mathbb{R}^p$: min $\sum_{i=1}^N (y_i - a - \mathbf{b}^{tr}\mathbf{x}_i)^2$

Population Equations: $a = \bar{y}_U - \mathbf{b}^{tr}\bar{\mathbf{x}}_U$

$$\mathbf{b} = \left(\sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}}_U)(\mathbf{x}_i - \bar{\mathbf{x}}_U)^{tr} \right)^{-1} \sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}}_U)(y_i - \bar{y}_U)$$

Sample Estimates: $\bar{y}_{\mathcal{S},\pi} = \widehat{N}^{-1} \sum_{i \in \mathcal{S}} \frac{y_i}{\pi_i}$, $\bar{\mathbf{x}}_{\mathcal{S},\pi} = \widehat{N}^{-1} \sum_{i \in \mathcal{S}} \frac{\mathbf{x}_i}{\pi_i}$

$\widehat{a} = \bar{y}_{\mathcal{S},\pi} - \mathbf{b}^{tr}\bar{\mathbf{x}}_{\mathcal{S},\pi}$

$$\widehat{\mathbf{b}} = \left(\sum_{i \in \mathcal{S}} (\mathbf{x}_i - \bar{\mathbf{x}}_{\mathcal{S},\pi})(\mathbf{x}_i - \bar{\mathbf{x}}_{\mathcal{S},\pi})^{tr} \right)^{-1} \sum_{i \in \mathcal{S}} (\mathbf{x}_i - \bar{\mathbf{x}}_{\mathcal{S},\pi})(y_i - \bar{y}_{\mathcal{S},\pi})$$

Regression Total-Estimator: $\widehat{N}^{-1}\,\hat{t}_y^{\,\mathrm{reg}} \;=\; \bar{y}_{\mathcal{S},\pi} + \hat{\mathbf{b}}^{tr}(\bar{\mathbf{x}}_U - \bar{\mathbf{x}}_{\mathcal{S},\pi})$

$$= \; \bar{y}_{\mathcal{S},\pi} - \bar{y}_U + \bar{y}_U - \mathbf{b}^{tr}(\bar{\mathbf{x}}_{\mathcal{S},\pi} - \bar{\mathbf{x}}_U) + (\mathbf{b} - \hat{\mathbf{b}})\,(\bar{\mathbf{x}}_{\mathcal{S},\pi} - \bar{\mathbf{x}}_U)$$

$$= \; (\bar{y}_{\mathcal{S},\pi} - \mathbf{b}^{tr}\bar{\mathbf{x}}_{\mathcal{S},\pi} - a) \;+\; \bar{y}_U \;+\; (\mathbf{b} - \hat{\mathbf{b}})\,(\bar{\mathbf{x}}_{\mathcal{S},\pi} - \bar{\mathbf{x}}_U)$$

Same conclusion as before, 3rd term is product of two small differences, and for large N, n

$\hat{t}_y^{\,\mathrm{reg}} - \widehat{N}\,\bar{y}_U \;\approx\; \sum_{\mathcal{S}} \frac{1}{\pi_i}\,(y_i - a - \mathbf{b}^{tr}\mathbf{x}_i)$ has large-sample Variance

$$\approx \; \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\pi_{i,j} - \pi_i\,\pi_j}{\pi_i\,\pi_j}\,(y_i - a - \mathbf{b}^{tr}\mathbf{x}_i)\,(y_j - a - \mathbf{b}^{tr}\mathbf{x}_j)$$

generally of order N^2/n as in SRS case, and estimated by

$$\sum_{i\in\mathcal{S}} \sum_{j\in\mathcal{S}} \frac{\pi_{i,j} - \pi_i\,\pi_j}{\pi_{ij}\,\pi_i\,\pi_j}\,(y_i - \hat{a} - \hat{b}x_i)(y_j - \hat{a} - \hat{b}x_j)$$