Regression in complex sampling designs

Suppose that the population model is:

$$Y_j = \beta_0 + \beta_1 X_j + \varepsilon_j \quad , \; j = 1, 2, \ldots, N$$

Using least squares, the parameter values would be

$$\beta_1 = \frac{S_{xy}}{S_x^2} \quad , \quad \beta_0 = \bar{Y} - \beta_1 \bar{X} = \frac{t_y}{N} - \beta_1 \frac{t_x}{N}$$

Where $$S_x^2 = \frac{1}{N-1} \sum_{j=1}^{N} (X_j - \bar{X})^2$$

and $$S_{xy} = \frac{1}{N-1} \sum_{j=1}^{N} (X_j - \bar{X})(Y_j - \bar{Y})$$

2

Note: $$(N-1) S_x^2 = \sum_{j=1}^{N} X_j^2 - 2\bar{X} \underbrace{\sum_{j=1}^{N} X_i}_{N\bar{X}} + N\bar{X}^2$$

$$= \sum_{j=1}^{N} X_j^2 - N\bar{X}^2$$

$$= \sum_{j=1}^{N} X_j^2 - N\left(\frac{t_x}{N}\right)^2 = t_{x^2} - \frac{t_x^2}{N}$$

and
$$(N-1) S_{xy} = \sum_{j=1}^{N} X_j Y_j - \bar{Y} \sum_{j=1}^{N} X_j - \bar{X} \sum_{j=1}^{N} Y_j + N \bar{X}\bar{Y}$$

$$= \sum_{j=1}^{N} X_j Y_j - \bar{Y} \cdot N\bar{X} - \bar{X} \cdot N\bar{Y} + N\bar{X}\bar{Y}$$

$$= \sum_{j=1}^{N} X_j Y_j - N\bar{X}\bar{Y} = t_{xy} - \frac{t_x t_y}{N}$$

③

To estimate β_0 and β_1, replace $t_x, t_Y, t_{x^2}, t_{xY}, \bar{Y}, \bar{X}$ with their estimators.

The Horvitz-Thompson estimator of t_Y is

$$\hat{t}_Y = \sum_{i=1}^{n} \frac{y_i}{\pi_i} = \sum_{i=1}^{n} w_i y_i \text{ , where } w_i = \frac{1}{\pi_i}$$

Similarly, $\hat{t}_x = \sum_{i=1}^{n} w_i x_i$, $\hat{t}_{x^2} = \sum_{i=1}^{n} w_i x_i^2$, $\hat{t}_{xy} = \sum_{i=1}^{n} w_i x_i y_i$

But it may be that we don't know N, and we also need to estimate it.

④

$$E[w_i] = E\left[\frac{1}{\pi_i}\right] = \sum_{j=1}^{N} \frac{1}{\pi_j} \cdot p_j \text{ , where}$$

p_j is the probability of selecting Y_j on a particular draw.

$$\pi_j = \text{Prob}(Y_j \text{ is in the sample})$$
$$= \sum_{i=1}^{n} \text{Prob}(Y_j \text{ was drawn on the } i^{th} \text{ draw})$$
$$= n p_j \qquad \Rightarrow p_j = \frac{\pi_j}{n}$$

So $E[w_i] = \sum_{j=1}^{N} \frac{1}{\pi_j} \frac{\pi_j}{n} = \frac{N}{n}$

5

Then $E\left[\sum_{i=1}^{n} w_i\right] = \sum_{i=1}^{n} \frac{N}{n} = n\frac{N}{n} = N$

That is, the sum of the sample weights is an unbiased estimator of the population size.

[Note: In SRS, each $\pi_i = \frac{n}{N}$ so $w_i = \frac{N}{n}$ and $\sum_{i=1}^{n} w_i = N$ exactly.]

6

Now, $$\hat{\beta}_1 = \frac{\hat{t}_{xy} - \dfrac{\hat{t}_x \hat{t}_y}{\hat{N}}}{\hat{t}_{x^2} - \dfrac{\hat{t}_x^2}{\hat{N}}}$$

$$= \frac{\sum_{i=1}^{n} w_i x_i y_i - \dfrac{\left(\sum_{i=1}^{n} w_i x_i\right)\left(\sum_{i=1}^{n} w_i y_i\right)}{\sum_{i=1}^{n} w_i}}{\sum_{i=1}^{n} w_i x_i^2 - \dfrac{\left(\sum_{i=1}^{n} w_i x_i\right)^2}{\sum_{i=1}^{n} w_i}}$$

(7)

Also, $$\hat{\beta}_0 = \frac{\hat{t}_y}{\hat{N}} - \hat{\beta}_1 \frac{\hat{t}_x}{\hat{N}}$$

$$= \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} - \hat{\beta}_1 \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

These estimators are exactly the same as those in weighted least squares!

(8)

In ordinary least squares, the matrix version of the model is $Y = X\vec{\beta} + \vec{\varepsilon}$, where $\vec{\varepsilon} \sim N(\vec{0}, \sigma^2 I)$, and the solution is

$$\hat{\beta} = (X'X)^{-1}X'Y$$

In weighted least squares, the model is the same, but we assume $\vec{\varepsilon} \sim N(\vec{0}, \sigma^2 V)$, where V is a diagonal matrix of constants.

The solution is $\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}Y$

the elements of V^{-1} are the <u>weights</u>.

In our case, $V = \begin{bmatrix} \pi_1 & & 0 \\ & \ddots & \\ 0 & & \pi_n \end{bmatrix}$ and $V^{-1} = \begin{bmatrix} w_1 & & 0 \\ & \ddots & \\ 0 & & w_n \end{bmatrix}$

Thus, what we are doing is equivalent to assuming that $\varepsilon_i \sim N(0, \sigma^2 \pi_i)$

WLS is designed to give smaller weights to y-values that have larger variances.

This is analogous to the H-T estimator giving smaller weights to y values that had higher probabilities