Introduction to Sampling Theory

Lecture 23

Regression Method of Estimation

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Slides can be downloaded from

http://home.iitk.ac.in/~shalab/sp

Regression Estimates in Stratified Sampling:

Under the set up of stratified sampling, let the population of *N* sampling units be divided into *k* strata.

The strata sizes are $N_1, N_2, \ldots, N_k$ such that $\sum_{i=1}^{k} N_i = N.$

A sample of size n_i on $(x_{ij}, y_{ij}),\ j = 1, 2, .., n_i,$ **is drawn from i^{th} strata (i = 1,2,..,k) by SRSWOR where x_{ij} and y_{ij} denote the j^{th} unit from i^{th} strata on auxiliary and study variables, respectively.**

Regression Estimates in Stratified Sampling:

Regression Estimates in Stratified Sampling:

In order to estimate the population mean, there are two approaches:

1. **Separate regression estimator**
2. **Combined regression estimator**

1. Separate Regression Estimator:

- **Estimate regression estimator**

$$\hat{\bar{Y}}_{reg} = \bar{y} + \beta_0(\bar{X} - \bar{x})$$

from each stratum separately, i.e., the regression estimate in the i^{th} stratum is $\hat{\bar{Y}}_{reg(i)} = \bar{y}_i + \beta_i(\bar{X}_i - \bar{x}_i).$

- **Find the stratified mean as the weighted mean of** $\hat{\bar{Y}}_{reg(i)}$ $i = 1, 2, .., k$ **as**

$$\hat{\bar{Y}}_{sreg} = \sum_{i=1}^{k} \frac{N_i \hat{\bar{Y}}_{reg(i)}}{N}$$

$$= \sum_{i=1}^{k} [w_i \{\bar{y}_i + \beta_i(\bar{X}_i - \bar{x}_i)\}]$$

where $\beta_i = \frac{S_{ixy}}{S_{ix}^2}, \; w_i = \frac{N_i}{N}.$

1. Separate Regression Estimator:

In this approach, the regression estimator is separately obtained in each of the stratum and then combined using the philosophy of stratified sample.

So $\hat{\bar{Y}}_{sreg}$ **is termed as separate regression estimator.**

2. Combined Regression Estimator:

Another strategy is to estimate $\overline{x}$ and $\overline{y}$ in the $\hat{\overline{Y}}_{reg}$ as respective stratified mean.

Replacing $\overline{X}$ by $\overline{x}_{st} = \sum_{i=1}^{k} w_i \overline{x}_i$ and $\overline{y}$ by $\overline{y}_{st} = \sum_{i=1}^{k} w_i \overline{y}_i$, we have

$$\hat{\overline{Y}}_{creg} = \overline{y}_{st} + \beta(\overline{X} - \overline{x}_{st}).$$

In this case, all the sample information is combined first and then implemented in regression estimator, so $\hat{\overline{Y}}_{reg}$ is termed as combined regression estimator.

Properties of Separate and Combined Regression Estimators:

In order to derive the mean and variance of $\hat{\bar{Y}}_{sreg}$ and $\hat{\bar{Y}}_{creg}$, there are two cases

- when β is pre-assigned as β_0.
- when β is estimated from the sample.

We consider here the case that β is pre-assigned as β_0.

Other case when β is estimated as $\hat{\beta} = \frac{s_{xy}}{s_x^2}$ can be dealt with the same approach based on defining various ε's and using the approximation theory as in the case of $\hat{\bar{Y}}_{reg}$.

1. Separate Regression Estimator:

Assume β is known, say β_0. Then

$$\hat{\bar{Y}}_{sreg} = \sum_{i=1}^{k} w_i [\bar{y}_i + \beta_{0i}(\bar{X}_i - \bar{x}_i)]$$

$$E(\hat{\bar{Y}}_{sreg}) = \sum_{i=1}^{k} w_i \left[E(\bar{y}_i) + \beta_{0i} \left(\bar{X}_i - E(\bar{x}_i) \right) \right]$$

$$= \sum_{i=1}^{k} w_i [\bar{Y}_i + (\bar{X}_i - \bar{X}_i)]$$

$$= \bar{Y}.$$

1. Separate Regression Estimator:

$$Var(\hat{\bar{Y}}_{sreg}) = E\left[\hat{\bar{Y}}_{sreg} - E(\hat{\bar{Y}}_{sreg})\right]^2$$

$$= E\left[\sum_{i=1}^{k} w_i \bar{y}_i + \sum_{i=1}^{k} w_i \beta_{0i}(\bar{X}_i - \bar{x}_i) - \bar{Y}\right]^2$$

$$= E\left[\sum_{i=1}^{k} w_i(\bar{y}_i - \bar{Y}) - \sum_{i=1}^{k} w_i \beta_{0i}(\bar{x}_i - \bar{X}_i)\right]^2$$

$$= \sum_{i=1}^{k} w_i^2 E(\bar{y}_i - \bar{Y}_i)^2 + \sum_{i=1}^{k} w_i^2 \beta_{0i}^2 E(\bar{x}_i - \bar{X}_i)^2 - 2\sum_{i=1}^{k} w_i^2 \beta_{0i} E(\bar{x}_i - \bar{X}_i)(\bar{y}_i - \bar{Y}_i)$$

$$= \sum_{i=1}^{k} w_i^2 Var(\bar{y}_i) + \sum_{i=1}^{k} w_i^2 \beta_{0i}^2 Var(\bar{x}_i) - 2\sum_{i=1}^{k} w_i^2 \beta_{0i} Cov(\bar{x}_i, \bar{y}_i)$$

$$= \sum_{i=1}^{k} \frac{w_i^2 f_i}{n_i}(S_{iY}^2 + \beta_{0i}^2 S_{iX}^2 - 2\beta_{0i} S_{iXY})]$$

1. Separate Regression Estimator:

$Var(\hat{\bar{Y}}_{sreg})$ **is minimum when** $\beta_{0i} = \dfrac{S_{iXY}}{S_{iX}^2}$ **and so substituting** β_{0i}, **we have**

$$V_{\min}(\hat{\bar{Y}}_{sreg}) = \sum_{i=1}^{k} \left[\frac{w_i^2 f_i}{n_i} (S_{iY}^2 - \beta_{0i}^2 S_{iX}^2) \right]$$

where $f_i = \dfrac{N_i - n_i}{N_i}$.

Since SRSWOR is followed in drawing the samples from each stratum, so

$$E(s_{ix}^2) = S_{iX}^2$$

$$E(s_{iy}^2) = S_{iY}^2$$

$$E(s_{ixy}) = S_{iXY}.$$

1. Separate Regression Estimator:

Thus an unbiased estimator of variance can be obtained by replacing S_{iX}^2 and S_{iY}^2 by their respective unbiased estimators s_{ix}^2 and s_{iy}^2, respectively as

$$\widehat{Var}(\hat{\bar{Y}}_{s\,reg}) = \sum_{i=1}^{k}\left[\frac{w_i^2 f_i}{n_i}(s_{iy}^2 + \beta_{0i}^2 s_{ix}^2 - 2\beta_{0i} s_{ixy})\right]$$

and

$$\widehat{Var}_{\min}(\hat{\bar{Y}}_{s\,reg}) = \sum_{i=1}^{k}\left[\frac{w_i^2 f_i}{n_i}(s_{iy}^2 - \beta_{0i}^2 s_{ix}^2)\right].$$

2. Combined Regression Estimator:

Assume β is known, say β_0. Then

$$\hat{\bar{Y}}_{creg} = \sum_{i=1}^{k} w_i \bar{y}_i + \beta_0 (\bar{X} - \sum_{i=1}^{k} w_i \bar{x}_i)$$

$$E\left(\hat{\bar{Y}}_{creg}\right) = \sum_{i=1}^{k} w_i E(\bar{y}_i) + \beta_0 [\bar{X} - \sum_{i=1}^{k} w_i E(\bar{x}_i)]$$

$$= \sum_{i=1}^{k} w_i \bar{Y}_i + \beta_0 [\bar{X} - \sum_{i=1}^{k} w_i \bar{X}_i]$$

$$= \bar{Y} + \beta_0 (\bar{X} - \bar{X})$$

$$= \bar{Y}.$$

Thus $\hat{\bar{Y}}_{creg}$ is an unbiased estimator of $\bar{Y}$.

2. Combined Regression Estimator:

$$Var(\hat{\bar{Y}}_{creg}) = E[\bar{Y}_{creg} - E(\bar{Y}_{creg})]^2$$

$$= E\left[\sum_{i=1}^{k} w_i \bar{y}_i + \beta_0 (\bar{X} - \sum_{i=1}^{k} w_i \bar{x}_i) - \bar{Y}\right]^2$$

$$= E\left[\sum_{i=1}^{k} w_i (\bar{y}_i - \bar{Y}) - \beta_0 \sum_{i=1}^{k} w_i (\bar{x}_i - \bar{X}_i)\right]^2$$

$$= \sum_{i=1}^{k} w_i^2 Var(\bar{y}_i) + \beta_0^2 \sum_{i=1}^{k} w_i^2 Var(\bar{x}_i) - 2\sum_{i=1}^{k} w_i^2 \beta_0 Cov(\bar{x}_i, \bar{y}_i)$$

$$= \sum_{i=1}^{k} \frac{w_i^2 f_i}{n_i}\left[S_{iY}^2 + \beta_0^2 S_{iX}^2 - 2\beta_0 S_{iXY}\right].$$

2. Combined Regression Estimator:

$Var(\hat{\bar{Y}}_{creg})$ **is minimum when**

$$\beta_0 = \frac{Cov(\bar{x}_{st}, \bar{y}_{st})}{Var(\bar{x}_{st})}$$

$$= \frac{\sum_{i=1}^{k} \frac{w_i^2 f_i}{n_i} S_{iXY}}{\sum_{i=1}^{k} \frac{w_i^2 f_i}{n_i} S_{iX}^2}$$

and the minimum variance is given by

$$Var_{\min}(\hat{\bar{Y}}_{creg}) = \sum_{i=1}^{k} \frac{w_i^2 f_i}{n_i} (S_{iY}^2 - \beta_0^2 S_{iX}^2).$$

2. Combined Regression Estimator:

Since SRSWOR is followed to draw the sample from strata, so using

$$E\left(s_{ix}^2\right)=S_{iX}^2,\ E\left(s_{iy}^2\right)=S_{iY}^2 \text{ and } E\left(s_{ixy}\right)=S_{iXY},$$

we get the estimate of variance as

$$\widehat{Var}(\hat{\bar{Y}}_{creg})=\sum_{i=1}^{k}\left[\frac{w_i^2 f_i}{n_i}(s_{iy}^2+\beta_0^2 s_{ix}^2-2\beta_{0i}s_{ixy})\right]$$

and

$$\widehat{Var}_{\min}(\hat{\bar{Y}}_{creg})=\sum_{i=1}^{k}\left[\frac{w_i^2 f_i}{n_i}(s_{iy}^2-\beta_{0i}^2 s_{ix}^2)\right].$$

Comparison of Separate and Combined Regression Estimator:

The variance of $\hat{\bar{Y}}_{sreg}$ is minimum when $\beta_{0i} = \beta_0$ for all i.

The variance of $\hat{\bar{Y}}_{creg}$ is minimum when $\beta_0 = \dfrac{Cov(\bar{x}_{st}, \bar{y}_{st})}{Var(\bar{x}_{st})} = \beta_0^*.$

Comparison of Separate and Combined Regression Estimator:

The minimum variance is

$$Var(\hat{\bar{Y}}_{creg})_{\min} = Var(\bar{y}_{st})(1-\rho_*^2)$$

where $\rho_* = \dfrac{Cov(\bar{x}_{st}, \bar{y}_{st})}{\sqrt{Var(\bar{x}_{st})Var(\bar{y}_{st})}}$.

$$Var(\hat{\bar{Y}}_{creg}) - Var(\hat{\bar{Y}}_{sreg}) = \sum_{i=1}^{k}(\beta_{0i}^2 - \beta_0^2)\frac{w_i^2 f_i}{n_i} S_{iX}^2$$

$$Var(\hat{\bar{Y}}_{creg})_{\min} - Var(\hat{\bar{Y}}_{sreg})_{\beta_{0i}=\beta_0} = \sum_{i=1}^{k}\frac{f_i}{n_i}(\beta_{0i} - \beta_0)^2 w_i^2 S_{iX}^2 \geq 0$$

Comparison of Separate and Combined Regression Estimator:

We observe that

$$Var(\hat{\bar{Y}}_{creg})_{\min} - Var(\hat{\bar{Y}}_{sreg})_{\beta_{0i}=\beta_0} = \sum_{i=1}^{k} \frac{f_i}{n_i} (\beta_{0i} - \beta_0)^2 w_i^2 S_{iX}^2 \geq 0$$

which is always true.

So if the regression line of *y* on *x* is approximately linear and the regression coefficients do not vary much among the strata, then separate regression estimate is more efficient than combined regression estimator.