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 $(x_{ij}, y_{ij})$, $j = 1, 2, \ldots, n_i$, is drawn from $i^{th}$ strata $(i = 1, 2, \ldots, k)$ 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
  \[ \hat{Y}_{\text{reg}} = \bar{y} + \beta_0 (\bar{X} - \bar{x}) \]
  from each stratum separately, i.e., the regression estimate in the \( i^{th} \) stratum is \( \hat{Y}_{\text{reg}(i)} = \bar{y}_i + \beta_i (\bar{X}_i - \bar{x}_i) \).
- Find the stratified mean as the weighted mean of \( \hat{Y}_{\text{reg}(i)} \) \( i = 1, 2, \ldots, k \) as
  \[
  \hat{Y}_{\text{sreg}} = \sum_{i=1}^{k} \frac{N_i \hat{Y}_{\text{reg}(i)}}{N} = \sum_{i=1}^{k} \left[ w_i \{ \bar{y}_i + \beta_i (\bar{X}_i - \bar{x}_i) \} \right]
  \]
  where \( \beta_i = \frac{S_{i xy}}{S_{ix}^2}, w_i = \frac{N_i}{N} \).
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 \( \hat{Y}_{sreg} \) is termed as separate regression estimator.
2. Combined Regression Estimator:

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

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

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

In this case, all the sample information is combined first and then implemented in regression estimator, so $\hat{Y}_{reg}$ is termed as combined regression estimator.
Properties of Separate and Combined Regression Estimators:

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

- when $\beta$ is pre-assigned as $\beta_0$.
- when $\beta$ is estimated from the sample.

We consider here the case that $\beta$ is pre-assigned as $\beta_0$.

Other case when $\beta$ is estimated as $\hat{\beta} = \frac{S_{xy}}{S_x^2}$ can be dealt with the same approach based on defining various $\varepsilon^t$s and using the approximation theory as in the case of $\hat{Y}_{reg}$.
1. Separate Regression Estimator:

Assume $\beta$ is known, say $\beta_0$. Then

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

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

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

$$= \bar{Y}.$$
1. Separate Regression Estimator:

\[
\text{Var}(\hat{Y}_{s\text{reg}}) = E \left[ \hat{Y}_{s\text{reg}} - E(\hat{Y}_{s\text{reg}}) \right]^2
\]

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

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

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

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

\[
= \sum_{i=1}^{k} \frac{w_i^2 f_i}{n_i} \left( S_{iY}^2 + \beta_{0i}^2 S_{iX}^2 - 2 \beta_{0i} S_{iXY} \right)
\]
1. Separate Regression Estimator:

\[ \text{Var} \left( \hat{Y}_{s\text{reg}} \right) \text{ is minimum when } \beta_{0i} = \frac{S_{iXY}}{S_{iX}^2} \text{ and so substituting } \beta_{0i}, \text{ we have} \]

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

where \( f_i = \frac{N_i - n_i}{N_i} \).

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

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

and

\[
\widehat{Var}_{\text{min}}(\hat{Y}_{sreg}) = \sum_{i=1}^{k} \left[ \frac{w_i^2 f_i}{n_i} \left( s_{iy}^2 - \beta_{0i}^2 s_{ix}^2 \right) \right].
\]
2. Combined Regression Estimator:

Assume \( \beta \) is known, say \( \beta_0 \). Then

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

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

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

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

\[
= \bar{Y}.
\]

Thus \( \hat{Y}_{creg} \) is an unbiased estimator of \( \bar{Y} \).
2. Combined Regression Estimator:

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

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

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

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

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

\( \text{Var}(\hat{Y}_{\text{creg}}) \) is minimum when

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

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

and the minimum variance is given by

\[
\text{Var}_{\text{min}}(\hat{Y}_{\text{creg}}) = \sum_{i=1}^{k} \frac{w_i^2 f_i}{n_i} (S_{iY}^2 - \beta_0^2 S_{iX}^2).
\]
2. Combined Regression Estimator:

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

\[ E \left( s_{ix}^2 \right) = S_{ix}^2, \quad E \left( s_{iy}^2 \right) = S_{iy}^2 \text{ and } E \left( s_{iXY} \right) = S_{iXY}, \]

we get the estimate of variance as

\[
\hat{Var} \left( \hat{Y}_{c\text{reg}} \right) = \sum_{i=1}^{k} \left[ \frac{w_i^2 f_i}{n_i} \left( s_{iy}^2 + \beta_{0i}^2 s_{ix}^2 - 2 \beta_{0i} s_{iXY} \right) \right]
\]

and

\[
\hat{Var}_{\text{min}} \left( \hat{Y}_{c\text{reg}} \right) = \sum_{i=1}^{k} \left[ \frac{w_i^2 f_i}{n_i} \left( s_{iy}^2 - \beta_{0i}^2 s_{ix}^2 \right) \right].
\]
Comparison of Separate and Combined Regression Estimator:
The variance of $\hat{Y}_{s\text{reg}}$ is minimum when $\beta_{0i} = \beta_0$ for all $i$.

The variance of $\hat{Y}_{c\text{reg}}$ is minimum when $\beta_0 = \frac{\text{Cov}(\bar{x}_{st}, \bar{y}_{st})}{\text{Var}(\bar{x}_{st})} = \beta^*_0$. 
Comparison of Separate and Combined Regression Estimator:

The minimum variance is

\[ \text{Var}(\hat{Y}_{\text{c reg}})_{\text{min}} = \text{Var}(\bar{y}_{st})(1 - \rho^*_2) \]

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

\[ \text{Var}(\hat{Y}_{\text{c reg}}) - \text{Var}(\hat{Y}_{\text{s reg}}) = \sum_{i=1}^{k} (\beta_{0i}^2 - \beta_0^2) \frac{w_i f_i}{n_i} S_{lX}^2 \]

\[ \text{Var}(\hat{Y}_{\text{c reg}})_{\text{min}} - \text{Var}(\hat{Y}_{\text{s reg}})_{\beta_{0i} = \beta_0} = \sum_{i=1}^{k} \frac{f_i}{n_i} (\beta_{0i} - \beta_0)^2 w_i^2 S_{lX}^2 \geq 0 \]
Comparison of Separate and Combined Regression Estimator:

We observe that

\[
Var(\hat{Y}_{c\text{reg}})_{\min} - Var(\hat{Y}_{s\text{reg}})_{\beta_0=\beta_0} = \sum_{i=1}^{k} \frac{f_i}{n_i} (\beta_{0i} - \beta_0)^2 w_i S_{iX}^2 \geq 0
\]

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.