Introduction to Sampling Theory

Lecture 22
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 when β is Computed from Sample:

Suppose a random sample of size *n* on paired observations,

$(x_i, y_i),\ i = 1, 2, .., n$ **is drawn by SRSWOR.**

When β is unknown, it is estimated as

$$\hat{\beta} = \frac{S_{xy}}{S_x^2}$$

and then the regression estimator of $\bar{Y}$ is given by

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

Bias and Mean Squared Error of the Regression Estimates:

It is difficult to find the exact expressions of $E(\hat{\bar{Y}}_{reg})$ **and** $Var(\hat{\bar{Y}}_{reg})$.

So we approximate them using the same methodology as in the case of ratio method of estimation.

Let

$$\varepsilon_0 = \frac{\bar{y} - \bar{Y}}{\bar{Y}} \Rightarrow \bar{y} = \bar{Y}(1 + \varepsilon_0)$$

$$\varepsilon_1 = \frac{\bar{x} - \bar{X}}{\bar{x}} \Rightarrow \bar{x} = \bar{X}(1 + \varepsilon_1)$$

$$\varepsilon_2 = \frac{s_{xy} - S_{XY}}{S_{XY}} \Rightarrow s_{xy} = S_{XY}(1 + \varepsilon_2)$$

$$\varepsilon_3 = \frac{s_x^2 - S_X^2}{S_X^2} \Rightarrow s_x^2 = S_X^2(1 + \varepsilon_3).$$

Bias and Mean Squared Error of the Regression Estimates:

Then

$$E(\varepsilon_0) = 0$$

$$E(\varepsilon_1) = 0$$

$$E(\varepsilon_2) = 0$$

$$E(\varepsilon_3) = 0$$

$$E(\varepsilon_0^2) = \frac{f}{n} C_Y^2$$

$$E(\varepsilon_1^2) = \frac{f}{n} C_X^2$$

$$E(\varepsilon_0 \varepsilon_1) = \frac{f}{n} \rho C_X C_Y$$

Bias and Mean Squared Error of the Regression Estimates:

Consider

$$\hat{\bar{Y}}_{reg} = \bar{y} + \frac{s_{xy}}{s_x^2}(\bar{X} - \bar{x})$$

$$= \bar{Y}(1+\varepsilon_0) + \frac{S_{XY}(1+\varepsilon_2)}{S_X^2(1+\varepsilon_3)}(-\varepsilon_1 \bar{X}).$$

The estimation error of $\hat{\bar{Y}}_{reg}$ **is**

$$(\hat{\bar{Y}}_{reg} - \bar{Y}) = \bar{Y}\varepsilon_0 - \beta \bar{X}\varepsilon_1(1+\varepsilon_2)(1+\varepsilon_3)^{-1}$$

where $\beta = \frac{S_{XY}}{S_X^2}$ **is the population regression coefficient.**

Assuming $|\varepsilon_3| < 1,$

$$(\hat{\bar{Y}}_{reg} - \bar{Y}) = \bar{Y}\varepsilon_0 - \beta \bar{X}(\varepsilon_1 + \varepsilon_1\varepsilon_2)(1 - \varepsilon_3 + \varepsilon_3^2 - ....).$$

Bias and Mean Squared Error of the Regression Estimates:

Assuming $|\varepsilon_3| < 1,$

$$(\hat{\bar{Y}}_{reg} - \bar{Y}) = \bar{Y}\varepsilon_0 - \beta\bar{X}(\varepsilon_1 + \varepsilon_1\varepsilon_2)(1 - \varepsilon_3 + \varepsilon_3^2 - ....)$$

retaining the terms up to second power of $\varepsilon's$ **and ignoring other terms, we have**

$$(\hat{\bar{Y}}_{reg} - \bar{Y}) \simeq \bar{Y}\varepsilon_0 - \beta\bar{X}(\varepsilon_1 + \varepsilon_1\varepsilon_2)(1 - \varepsilon_3 + \varepsilon_3^2)$$

$$\simeq \bar{Y}\varepsilon_0 - \beta\bar{X}(\varepsilon_1 - \varepsilon_1\varepsilon_3 + \varepsilon_1\varepsilon_2).$$

Bias and Mean Squared Error of the Regression Estimates:

Now the bias of $\hat{\bar{Y}}_{reg}$ up to the second order of approximation is given as

$$E(\hat{\bar{Y}}_{reg} - \bar{Y}) \simeq E\left[\bar{Y}\varepsilon_0 - \beta\bar{X}\varepsilon_1(\varepsilon_1 + \varepsilon_1\varepsilon_2)(1 - \varepsilon_3 + \varepsilon_3^2)\right]$$

$$= -\frac{\beta\bar{X}f}{n}\left[\frac{\mu_{21}}{\bar{X}S_{XY}} - \frac{\mu_{30}}{\bar{X}S_X^2}\right]$$

where $f = \dfrac{N-n}{N}$ **and** $(r,s)^{th}$ **cross product moment is given by**

$$\mu_{rs} = E\left[(x-\bar{X})^r(y-\bar{Y})^s\right]$$

So that $\mu_{21} = E\left[(x-\bar{X})^2(y-\bar{Y})\right], \quad \mu_{30} = E\left[(x-\bar{X})^3\right].$

Thus

$$Bias(\hat{\bar{Y}}_{reg}) = -\frac{\beta f}{n}\left[\frac{\mu_{21}}{S_{XY}} - \frac{\mu_{30}}{S_X^2}\right].$$

Bias and Mean Squared Error of the Regression Estimates:

Also

$$E(\hat{\bar{Y}}_{reg}) = E(\bar{y}) + E[\hat{\beta}(\bar{X} - \bar{x})]$$

$$= \bar{Y} + \bar{X}E(\hat{\beta}) - E(\hat{\beta}\bar{x})$$

$$= \bar{Y} + E(\bar{x})E(\hat{\beta}) - E(\hat{\beta}\bar{x})$$

$$= \bar{Y} - Cov(\hat{\beta}, \bar{x})$$

$$Bias(\hat{\bar{Y}}_{reg}) = E(\hat{\bar{Y}}_{reg}) - \bar{Y} = -Cov(\hat{\beta}, \bar{x}).$$

Bias and Mean Squared Error of the Regression Estimates:

To obtain the *MSE* of $\hat{\bar{Y}}_{reg}$, consider

$$E(\hat{\bar{Y}}_{reg} - \bar{Y})^2 \approx E\left[\varepsilon_0 \bar{Y} - \beta \bar{X}(\varepsilon_1 - \varepsilon_1\varepsilon_3 + \varepsilon_1\varepsilon_2)\right]^2.$$

Retaining the terms of $\varepsilon's$ up to the second power and ignoring others, we have

$$E(\hat{\bar{Y}}_{reg} - \bar{Y})^2 \approx E\left[\varepsilon_0^2 \bar{Y}^2 + \beta^2 \bar{X}^2 \varepsilon_1^2 - 2\beta \bar{X}\bar{Y}\varepsilon_0\varepsilon_1\right]$$

$$= \bar{Y}^2 E(\varepsilon_0^2) + \beta^2 \bar{X}^2 E(\varepsilon_1^2) - 2\beta \bar{X}\bar{Y} E(\varepsilon_0\varepsilon_1)$$

$$= \frac{f}{n}\left[\bar{Y}^2 \frac{S_Y^2}{\bar{Y}^2} + \beta^2 \bar{X}^2 \frac{S_X^2}{\bar{X}^2} - 2\beta \bar{X}\bar{Y}\rho \frac{S_X S_Y}{\bar{X}\bar{Y}}\right]$$

Bias and Mean Squared Error of the Regression Estimates:

Thus

$$MSE(\hat{\bar{Y}}_{reg}) = E(\hat{\bar{Y}}_{reg} - \bar{Y})^2$$

$$= \frac{f}{n}(S_Y^2 + \beta^2 S_X^2 - 2\beta\rho S_X S_Y).$$

Since $\beta = \frac{S_{XY}}{S_X^2} = \rho\frac{S_Y}{S_X}$,

so substituting it in *MSE* $(\hat{\bar{Y}}_{reg})$, **we get**

$$MSE(\hat{\bar{Y}}_{reg}) = \frac{f}{n}S_Y^2(1-\rho^2).$$

Bias and Mean Squared Error of the Regression Estimates:

So up to the second order of approximation, the regression estimator is better than the conventional sample mean estimator under SRSWOR.

This is because the regression estimator uses some extra information also. Moreover, such extra information requires some extra cost also.

This shows a false superiority in some sense.

So the regression estimators and SRS estimates can be combined if cost aspect is also taken into consideration.

Comparison of $\hat{\bar{Y}}_{reg}$ with Ratio Estimate and SRS Sample Mean Estimate:

$$MSE(\hat{\bar{Y}}_{reg}) = \frac{f}{n} S_Y^2 (1-\rho^2)$$

$$MSE(\hat{\bar{Y}}_{R}) = \frac{f}{n} (S_Y^2 + R^2 S_X^2 - 2\rho R S_X S_Y)$$

$$Var_{SRS}(\bar{y}) = \frac{f}{n} S_Y^2.$$

(i) **As** $MSE\ (\hat{\bar{Y}}_{reg}) = Var_{SRS}(\bar{y})(1-\rho^2)$

and because $\rho^2 < 1,$

so $\hat{\bar{Y}}_{reg}$ **is always superior to sample mean estimate.**

Comparison of $\hat{\bar{Y}}_{reg}$ with Ratio Estimate and SRS Sample Mean Estimate:

(ii) $\hat{\bar{Y}}_{reg}$ **is better than** $\hat{\bar{Y}}_R$ **if** $MSE(\hat{\bar{Y}}_{reg}) \leq MSE(\hat{\bar{Y}}_R)$

or if $\frac{f}{n}S_Y^2(1-\rho^2) \leq \frac{f}{n}(S_Y^2 + R^2S_X^2 - 2\rho RS_XS_Y)$

or if $(RS_X - \rho S_Y)^2 \geq 0$

which always holds true.

So regression estimate is always superior to the ratio estimate up to the second order of approximation.