Introduction to Sampling Theory

Lecture 21
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 Method of Estimation:

The ratio method of estimation uses the auxiliary information which is correlated with the study variable to improve the precision which results in the improved estimators when the regression of Y on X is linear and passes through origin.

When the regression of Y on X is linear, it is not necessary that the line should always pass through origin.

Under such conditions, it is more appropriate to use the regression type estimators to estimate the population mean.

Regression Method of Estimation:

In ratio method, the conventional estimator sample mean $\bar{y}$ was improved by multiplying it by a factor $\frac{\bar{X}}{\bar{x}}$ where $\bar{x}$ is an unbiased estimator of population mean $\bar{X}$ which is chosen as population mean of auxiliary variable.

Now we consider another idea based on difference.

Regression Method of Estimation:

Consider an estimator $(\bar{x} - \bar{X})$ for which

$$E(\bar{x} - \bar{X}) = 0.$$

Consider an improved estimator of $\bar{Y}$ as

$$\hat{\bar{Y}}^{*} = \bar{y} + \mu(\bar{x} - \bar{X})$$

which is an unbiased estimator of $\bar{Y}$ and μ is any constant.

Now find μ such that the $Var(\hat{\bar{Y}}^{*})$ is minimum.

Regression Method of Estimation:

$$Var(\hat{\overline{Y}}*) = Var(\overline{y}) + \mu^2 Var(\overline{x}) + 2\mu Cov(\overline{x}, \overline{y})$$

$$\frac{\partial Var(\overline{Y}^*)}{\partial \mu} = 0$$

$$\Rightarrow \mu = -\frac{Cov(\overline{x}, \overline{y})}{Var(\overline{x})}$$

$$= -\frac{\frac{N-n}{Nn} S_{XY}}{\frac{N-n}{Nn} S_X^2}$$

$$= -\frac{S_{XY}}{S_X^2}$$

where $S_{XY} = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y}), \quad S_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X}).$

Regression Method of Estimation:

Consider a linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$ **where *y* is the dependent variable, *x* is the independent variable and** ε **is the random error component which takes care of the difference arising due to lack of exact relationship between *x* and *y*.**

So we can write the model for each observation as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1, 2, .., n.$$

Minimize $\sum_{i=1}^{n} \varepsilon_i^2$ **for given *n* paired data set** $(x_i, y_i),\ i = 1, 2, .., n.$

,

.

Regression Method of Estimation:

Minimizes the sum of squares

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

with respect to β_0 **and** β_1.

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = -2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)$$

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = -2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i$$

The solutions of β_0 **and** β_1 **are obtained by setting**

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = 0$$

,

.

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = 0.$$

Regression Method of Estimation:

We obtain the solutions b_0 **of** β_0 **and** b_1 **of** β_1

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{s_{xy}}{s_{xx}}$$

where

$$s_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}), \quad s_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2, \quad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \; \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i .$$

$b_1 = \dfrac{s_{xy}}{s_{xx}}$ **corresponds to** $\beta = \dfrac{Cov(x,y)}{Var(x)} = \dfrac{S_{xy}}{S_x^2}.$

Regression Method of Estimation:

Thus the optimum value of μ is same as the regression coefficient of y on x with a negative sign, *i.e.*,

$$\mu = -\beta.$$

So the estimator $\hat{\bar{Y}}^*$ with optimum value of μ is

$$\hat{\bar{Y}}_{reg} = \bar{y} + b_1(\bar{X} - \bar{x})$$

which is the estimator of $\bar{Y}$ and the procedure of estimation is called as the regression method of estimation.

Regression estimates with pre-assigned β:

If value of β is known as β_0 (say) then the regression estimator is

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

Bias of $\hat{\bar{Y}}_{reg}$

Now, assuming that the random sample $(x_i, y_i),\ i = 1, 2, .., n$ is drawn by SRSWOR,

$$\begin{aligned} E(\hat{\bar{Y}}_{reg}) &= E(\bar{y}) + \beta_0\left[\bar{X} - E(\bar{x})\right] \\ &= \bar{Y} + \beta_0\left[\bar{X} - \bar{X}\right] \\ &= \bar{Y}. \end{aligned}$$

Thus $\hat{\bar{Y}}_{reg}$ is an unbiased estimator of $\bar{Y}$ when β is known.

Regression estimates with pre-assigned β : Variance of $\hat{\bar{Y}}_{reg}$

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

$$= E\left[\bar{y} + \beta_0(\bar{X} - \bar{x}) - \bar{Y}\right]^2$$

$$= E\left[(\bar{y} - \bar{Y}) - \beta_0(\bar{x} - \bar{X})\right]^2$$

$$= E\left[(\bar{y} - \bar{Y})^2 + \beta_0^2(\bar{x} - \bar{X}) - 2\beta_0 E(\bar{x} - \bar{X})(\bar{y} - \bar{Y})\right]$$

$$= Var(\bar{y}) + \beta_0^2 Var(\bar{x}) - 2\beta_0 Cov(\bar{x}, \bar{y})$$

$$= \frac{f}{n}\left[S_Y^2 + \beta_0^2 S_X^2 - 2\beta_0 S_{XY}\right]$$

$$= \frac{f}{n}\left[S_Y^2 + \beta_0^2 S_X^2 - 2\beta_0 \rho S_X S_Y\right]$$

where

$$f = \frac{N-n}{N}; \quad S_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2; \quad S_Y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$$

ρ : **Correlation coefficient between** X **and** Y.

Regression estimates with pre-assigned β:

Comparing $Var(\hat{\bar{Y}}_{reg})$ with $Var(\bar{y})$, we note that

$$Var(\hat{\bar{Y}}_{reg}) < Var(\bar{y})$$

If $$\beta_0^2 S_X^2 - 2\beta_0 S_{XY} < 0$$

or $$\beta_0 S_X^2 \left(\beta_0 - \frac{2S_{XY}}{S_X^2} \right) < 0$$

which is possible when

either $$\beta_0 < 0 \text{ and } \left(\beta_0 - \frac{2S_{XY}}{S_X^2} \right) > 0 \Rightarrow \frac{2S_{XY}}{S_X^2} < \beta_0 < 0$$

or $$\beta_0 > 0 \text{ and } \left(\beta_0 - \frac{2S_{XY}}{S_X^2}) \right) < 0 \Rightarrow 0 < \beta_0 < \frac{2S_{XY}}{S_X^2}.$$

Optimal value of β:

Choose β **such that** $Var(\hat{\bar{Y}}_{reg})$ **is minimum .**

So $$\frac{\partial Var(\hat{\bar{Y}}_{reg})}{\partial \beta} = \frac{\partial}{\partial \beta}\left[S_Y^2 + \beta^2 S_X^2 - 2\beta\rho S_X S_Y\right] = 0$$

$$\Rightarrow \beta = \rho\frac{S_Y}{S_X} = \frac{S_{XY}}{S_X^2}.$$

The minimum value of variance of $\hat{\bar{Y}}_{reg}$ **with optimum value of** $\beta_{opt} = \frac{\rho S_Y}{S_X}$ **is**

$$Var_{min}(\hat{\bar{Y}}_{reg}) = \frac{f}{n}\left[S_Y^2 + \rho^2\frac{S_Y^2}{S_X^2}S_X^2 - 2\rho\frac{S_Y}{S_X}\rho S_X S_Y\right]$$

$$= \frac{f}{n}S_Y^2(1-\rho^2).$$

Optimal value of β:

We see from

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

Since $-1 \leq \rho \leq 1,$ so

$$Var(\hat{\bar{Y}}_{reg}) \leq Var_{SRS}(\bar{y})$$

which always holds true. So the regression estimator is always better than the sample mean under SRSWOR.

Departure from β:

If β_0 is the preassigned value of regression coefficient, then

$$Var_{\min}(\hat{\bar{Y}}_{reg}) = \frac{f}{n}\left[S_Y^2 + \beta_0^2 S_X^2 - 2\beta_0 \rho S_X S_Y\right]$$

$$= \frac{f}{n}\left[S_Y^2 + \beta_0^2 S_X^2 - 2\rho\beta_0 S_X S_Y - \rho^2 S_Y^2 + \rho^2 S_Y^2\right]$$

$$= \frac{f}{n}\left[(1-\rho^2)S_Y^2 + \beta_0^2 S_X^2 - 2\beta_0 S_X^2 \beta_{opt} + \beta_{opt}^2 S_X^2\right]$$

$$= \frac{f}{n}\left[(1-\rho^2)S_Y^2 + (\beta_0 - \beta_{opt})^2 S_X^2\right]$$

where $\beta_{opt} = \dfrac{\rho S_Y}{S_X}.$

Estimate of Variance:

An unbiased sample estimate of $Var(\hat{\bar{Y}}_{reg})$ **is**

$$\widehat{Var}(\hat{\bar{Y}}_{reg}) = \frac{f}{n(n-1)} \sum_{i=1}^{n} \left[(y_i - \bar{y}) - \beta_0 (x_i - \bar{x}) \right]^2$$

$$= \frac{f}{n} (s_y^2 + \beta_0^2 s_x^2 - 2\beta_0 s_{xy})$$

Note that the variance of $\hat{\bar{Y}}_{reg}$ **increases as the difference between** β_0 **and** β_{opt} **increases.**