Regression Estimation

Survey Sampling
Statistics 4234/5234
Fall 2018

October 11, 2018

The regression estimator

If the population $\{(x_i, y_i) : i = 1, 2, \ldots, N\}$ satisfies

$$y_i \approx B_0 + B_1 x_i$$

for each i, and the population mean for the auxiliary variables, $\bar{x}_U$, is known, we can use this information to estimate $\bar{y}_U$ with increased precision.

Assume our data consist of a simple random sample $\mathcal{S}$ of size n, let $\bar{x}$ and $\bar{y}$ denote the sample means, and define

$$\widehat{B}_1 = \frac{\sum_{i \in \mathcal{S}} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i \in \mathcal{S}} (x_i - \bar{x})^2}$$

and

$$\widehat{B}_0 = \bar{y} - \widehat{B}_1 \bar{x}$$

Then the *regression estimator* of $\bar{y}_U$ is defined by

$$\hat{\bar{y}}_{\text{reg}} = \hat{B}_0 + \hat{B}_1 \bar{x}_U = \bar{y} + \hat{B}_1 \left(\bar{x}_U - \bar{x}\right)$$

The regression estimator is *biased*.

$$\mathsf{E}\left(\hat{\bar{y}}_{\text{reg}} - \bar{y}_U\right) = \mathsf{E}\left[\hat{B}_1(\bar{x}_U - \bar{x})\right] = -\mathsf{Cov}\left(\hat{B}_1, \bar{x}\right)$$

“As with ratio estimation, for large SRSs the MSE for regression estimation is approximately equal to the variance; the bias is often negligible in large samples.”

Consider

$$\mathsf{MSE}\left(\widehat{\bar{y}}_{\mathsf{reg}}\right) = \mathsf{E}\left\{\left[\bar{y} + \widehat{B}_1(\bar{x}_U - \bar{x}) - \bar{y}_U\right]^2\right\}$$

Define the population quantities

$$\bar{x}_U \quad \text{and} \quad \bar{y}_U \quad \text{and} \quad S_x \quad \text{and} \quad S_y$$

as usual; also

$$R = \frac{\sum_{i=1}^{N}(x_i - \bar{x}_U)(y_i - \bar{y}_U)}{(N-1)S_x S_y}$$

and

$$B_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x}_U)(y_i - \bar{y}_U)}{\sum_{i=1}^{N}(x_i - \bar{x}_U)^2} = \frac{R\,S_y}{S_x}$$

Letting

$$d_i = y_i - [\bar{y}_u + B_1(x_i - \bar{x}_U)]$$

for each $i = 1, \dots, N$ we have

$$\text{MSE}\left(\hat{\bar{y}}_{\text{reg}}\right) \approx V(\bar{d}) = \frac{S_d^2}{n}\left(1 - \frac{n}{N}\right) \tag{1}$$

where

$$S_d^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \bar{y}_U - B_1\left[x_i - \bar{x}_U\right]\right)^2 = S_y^2(1 - R^2)$$

Thus we have

$$\text{MSE}\left(\hat{\bar{y}}_{\text{reg}}\right) \approx \frac{1}{n}S_y^2(1 - R^2)\left(1 - \frac{n}{N}\right) \tag{2}$$

Recall that

$$\text{MSE}\,(\bar{y}) = \frac{S_y^2}{n}\left(1 - \frac{n}{N}\right)$$

Thus the regression estimator pretty much *always* has a lower MSE than the ordinary sample mean.

The closer R is to ± 1 the better.

Standard error

As usual,

$$\mathsf{SE}\left(\widehat{\bar{y}}_{\mathsf{reg}}\right) = \sqrt{\widehat{V}\left(\widehat{\bar{y}}_{\mathsf{reg}}\right)}$$

Using (1) we have

$$\widehat{V}\left(\widehat{\bar{y}}_{\mathsf{reg}}\right) = \frac{s_e^2}{n}\left(1 - \frac{n}{N}\right)$$

where

$$s_e^2 = \frac{1}{n-1}\sum_{i\in\mathcal{S}}\left(y_i - \hat{B}_0 - \hat{B}_1 x_i\right)^2$$

and using (2) we get

$$\hat{V}\left(\widehat{\bar{y}}_{\text{reg}}\right) = \frac{1}{n} s_y^2 (1 - r^2)\left(1 - \frac{n}{N}\right)$$

Are these the same thing?

Yes. Just as

$$S_d^2 = S_y^2(1 - R^2)$$

in the population,

$$s_e^2 = s_y^2(1 - r^2)$$

in the sample.

Example

(Example 4.9 on page 139)

Estimate the number of dead trees.

The region is divided into 100 plots, analysts examine photographs of each, and count the number of dead trees. This is an imperfect method.

Also, take an SRS of 25 of the plots, and conduct a field count. This method is accurate.

Let $x_i =$ photo count for plot i for $i = 1, \ldots, 100$, the x_i are *all* known.

Let $y_i =$ field count, these values are only known for $i \in \mathcal{S}$.

Goal: Estimate $t_y = \sum_{i=1}^{100} y_i$.

We know that $\bar{x}_U = 11.3$.

The sample data give $\bar{x} = 10.60$ and $\bar{y} = 11.56$ and $s_y^2 = 9.09$. Also $\widehat{B}_1 = 0.6133$.

The regression estimator of average number of dead trees per plot is

$$\hat{\bar{y}}_{\text{reg}} = \bar{y} + \hat{B}_1(\bar{x}_U - \bar{x}) = 11.56 + 0.6133(11.3 - 10.6) = 11.99$$

The estimator is adjusted *up* because the sample plots have fewer photo counted dead trees than the unsampled plots ($\bar{x} < \bar{x}_U$); this suggests that the sample mean *under*estimates $\bar{y}_U$.

Standard error?

Well

$$\widehat{V}\left(\widehat{\bar{y}}_{\text{reg}}\right) = \frac{s_e^2}{n}\left(1 - \frac{n}{N}\right) = \frac{5.55}{25}\left(1 - \frac{25}{100}\right)$$

and thus

$$\text{SE}\left(\widehat{\bar{y}}_{\text{reg}}\right) = \sqrt{\widehat{V}\left(\widehat{\bar{y}}_{\text{reg}}\right)} = 0.408$$

Thus $\widehat{t}_{y,\text{reg}} = 1199$; we estimate a total of 1199 dead trees in the area, with a standard error of $\text{SE}\,(\widehat{t}_{y,\text{reg}}) = 41$ trees.

Do

$$\widehat{t}_{y,\text{reg}} \pm 2\ \text{SE}\,(\widehat{t}_{y,\text{reg}})$$

and we are about 95% confident that there are between 1117 and 1281 dead trees in this region.

Using the sample mean $\widehat{t}_y = N\bar{y} = 100(11.56) = 11.56$ we have

$$\widehat{V}\left(\bar{y}\right) = \frac{s_y^2}{n}\left(1 - \frac{n}{N}\right) = \frac{9.09}{25}\left(1 - \frac{25}{75}\right) = 0.2727$$

and thus

$$\mathsf{SE}\left(\bar{y}\right) = \sqrt{.2727} = .5252$$

and thus $\mathsf{SE}(t_y) = 52$.

Regression estimation gets our standard error from 52 trees down to 41 trees.

Ratio estimation would not have been nearly as effective in this problem.

$$\widehat{B} = \frac{\bar{y}}{\bar{x}} = \frac{11.56}{10.60} = 1.0956$$

and thus

$$\widehat{\bar{y}}_r = \widehat{B}\bar{x}_U = 1.096 \cdot 11.3 = 12.32$$

and we estimate a total of $t_y = 1232$ dead trees.

Now

$$\mathsf{SE}(\widehat{\bar{y}}_r) = \frac{\bar{x}_U}{\bar{x}} \frac{s_e}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$$

where s_e is the sample standard deviation of

$$e_i = y_i - \widehat{B} x_i$$

We obtain

$$\mathsf{SE}\left(\widehat{\bar{y}}_r\right) = 0.512$$

and thus

$$\mathsf{SE}\left(\hat{t}_{yr}\right) = 51$$

which is no real improvement over the ordinary sample mean.