STA304

Neil Montgomery
2016-03-21

ratio and regression estimators

ratio recap

If one has a sample $\{(y_1, x_1), \ldots, (y_n, x_n)\}$ the following may be of interest:

1. To estimate the population ratio $R = \tau_y \big/ \tau_x$ using $\hat{R} = r = \overline{y} \big/ \overline{x}$.

2. To enable estimation of τ_y when N is unknown.

3. To enable improved estimation of τ_y or μ_y when y and x are correlated.

The improvement in 3. occurs when the relationship between y and x is a straight line through the origin and $\hat{\rho} > 0.5$.

regression estimation

There is a more general form of (but not precisely a generalization of) the 3rd use ratio estimation ("improved estimation"), good for when the relationship is linear but not through the origin.

The setup is similar.

One has a population $\{(y_1, x_1), \ldots, (y_N, x_N)\}$ and simple random sample $\{(y_1, x_1), \ldots, (y_n, x_n)\}$. The true mean μ_x of the x variable is known.

We seek an improvement over $\bar{y}$, using the information contained in the x variable in the sample and from our knowledge of μ_x.

quick review of regression basics - I

The "least squares" regression line $y = \alpha + \beta x + \varepsilon$ fit through the points $\{(y_1, x_1), \ldots, (y_n, x_n)\}$ is given by slope and intercept estimates respectively:

$$\hat{\beta} = b = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\alpha} = a = \overline{y} - b\overline{x}$$

We can define the fitted values as:

$$\hat{y}_i = a + bx_i$$

and the "residuals" as:

$$\varepsilon_i = y_i - \hat{y}_i$$

quick review of regression basics - II

The sum of squared residuals divided by $n - 2$ is called the Mean Squared Error:

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$$

and is an estimate of the amount of variation around the regression line.

A plot of residuals on the horizontal axis versus fitted values on the vertical axis is an effective way to verify that there is a linear relationship between y and x.

the regression estimator for μ_y - I

A regression line could be used to estimate the mean y value at *any* x value, but we are mainly interested in the overall mean μ_y, which corresponds to the regression line evaluated at μ_x.

Note that regression lines always pass through the point $(\overline{x}, \overline{y})$, which is a reasonable estimator for the point (μ_y, μ_x).

So starting from:

$$\hat{y}_i = a + bx_i$$

substituting the formula for a we get:

$$\hat{y}_i = \overline{y} + b(x_i - \overline{x})$$

the regression estimator for μ_y - II

Finally substituting μ_x for x_i we end up with the regression estimator:

$$\hat{\mu}_{yL} = \bar{y} + b(\mu_x - \bar{x})$$

The estimated variance for $\hat{\mu}_{yL}$ is similar in spirit to that of the ratio estimator for μ_y, with the variance of the residuals playing the key role:

$$\hat{V}(\hat{\mu}_{yL}) = \left(1 - \frac{n}{N}\right) \frac{MSE}{n}$$

For estimating τ_y multiply both by N.