Vectorized Algorithm - Stute (1997) Test

Diego Ciccia*

July 11, 2024

Let $\underset{(N\times 1)}{Y}$ be a vector of sample realizations of outcome variables Y. The correspondent sample realizations of a treatment variable D, as well as a constant and other covariates, are all contained in matrix $\underset{(N\times k)}{X}$. We wish to implement Stute (1997) linearity test with B wild bootstrap replications. To this end, we have preallocated matrix $\underset{(N\times B)}{F}$ with N realizations of parametric transformations of B i.i.d. random draws from $U(0,1)$.

Matrix Setup. The `stute_stat()` function in the C++ script takes as input a matrix $\underset{(N\times p)}{V}$, where $p = 1$ for the computation of the test statistic, $p = B$ for the bootstrap routine. Each column of V corresponds to a vector of residuals from a linear regression of Y on X. For the point estimate, let $\underset{(N\times 1)}{E} = \underset{(N\times 1)}{V} = \underset{(N\times 1)}{Y} - \underset{(N\times k)}{X}\underset{(k\times 1)}{b}$ be the vector of residuals from the original dataset. The computation of residuals from all the bootstrap resampling of Y follows three steps:

1. Resample Y using F:
$$\underset{(N\times B)}{Y^{st}} = \left[\underset{(N\times 1)}{Y}\right]_B - \left[\underset{(N\times 1)}{E}\right]_B + \underset{(N\times B)}{F} \odot \left[\underset{(N\times 1)}{E}\right]_B$$
where the $[\underset{(m\times 1)}{.}]_n$ operator appends n copies of the argument vector to form a $m \times n$ matrix and $\odot$ is the element-wise product.

2. Compute the coefficients from a linear regression of each column of Y^{st} on X *all at once*:
$$\underset{(k\times B)}{b^{st}} = (\underset{(k\times N)}{X'} \cdot \underset{(N\times k)}{X})^{-1}(\underset{(k\times N)}{X'} \cdot \underset{(N\times B)}{Y^{st}})$$

3. Use b^{st} to compute the residual matrix:
$$\underset{(N\times B)}{V} = \underset{(N\times B)}{Y^{st}} - \underset{(N\times k)}{X} \cdot \underset{(k\times B)}{b^{st}}$$

*cicciadiego99@gmail.com.

Vectorized Computation. The test statistic from Stute (1997) in a dataset sorted by D can be rewritten as follows:

$$S = \frac{1}{N^2} \sum_{i=1}^{N} \left(\sum_{j=1}^{i} E_j \right)^2 .$$

In other words, the test statistic is equal to the sum of the squares of the cumulative sums. The double-summation can be converted into a vectorized algorithm. Let $\underset{(N\times N)}{L} = [i \geq j]_{i,j\in\{1,\ldots,N\}^2}$ be a lower diagonal matrix, with entries on and below the main diagonal equal to 1, elsewhere 0. Let $\underset{(1\times N)}{I}$ be a row unity vector. Then,

$$S = \frac{1}{N^2} \underset{(1\times N)}{I} \cdot (\underset{(N\times N)}{L} \cdot \underset{(N\times 1)}{E})^{\circ 2}$$

where $\circ$ is the element-wise power operator. Similarly, the B boostrap statistics are stacked in the following vector

$$\underset{(1\times B)}{S^{st}} = \frac{1}{N^2} \underset{(1\times N)}{I} \cdot (\underset{(N\times N)}{L} \cdot \underset{(N\times B)}{V})^{\circ 2}$$

and the corresponding p-value is computed as follows:

$$\frac{1}{B} \sum_{i=1}^{B} 1\{S_i^{st} > S\}.$$

Notice that LV generates the vector of cumulative sums of the columns of V. After raising every element to its square, left multiplication by I yields the column total. To see this, consider the following example:

$$V = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix} \quad L = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \quad I = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \quad LV = \begin{bmatrix} 1 & 4 \\ 3 & 9 \\ 6 & 15 \end{bmatrix} \quad ILV = \begin{bmatrix} 10 & 28 \end{bmatrix}$$

References

Winfried Stute. Nonparametric model checks for regression. *The Annals of Statistics*, pages 613–641, 1997.