by — Repeat Stata command on subsets of the data

<table>
<thead>
<tr>
<th>Description</th>
<th>Quick start</th>
<th>Syntax</th>
<th>Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>Remarks and examples</td>
<td>References</td>
<td>Also see</td>
<td></td>
</tr>
</tbody>
</table>

Description

Most Stata commands allow the by prefix, which repeats the command for each group of observations for which the values of the variables in varlist are the same. by without the sort option requires that the data be sorted by varlist; see [D] sort.

Stata commands that work with the by prefix indicate this immediately following their syntax diagram by reporting, for example, “by is allowed; see [D] by” or “bootstrap, by, etc., are allowed; see [U] 11.1.10 Prefix commands”.

by and bysort are really the same command; bysort is just by with the sort option.

The varlist1 (varlist2) syntax is of special use to programmers. It verifies that the data are sorted by varlist1 varlist2 and then performs a by as if only varlist1 were specified. For instance,

    by pid (time): generate growth = (bp - bp[_n-1])/bp

performs the generate by values of pid but first verifies that the data are sorted by pid and time within pid.

Quick start

Generate newv as an observation number within each level of catvar

    by catvar: generate newv = _n

As above, but sort data by catvar first

    by catvar, sort: generate newv = _n

Same as above

    bysort catvar: generate newv = _n

As above, but sort by v within values of catvar

    bysort catvar (v): generate newv = _n

Generate newv as an observation number for each observation in levels of catvar and v

    bysort catvar v: generate newv = _n

Note: Any command that accepts the by prefix may be substituted for generate above.
Syntax

by varlist: stata_cmd

bysort varlist: stata_cmd

The above diagrams show by and bysort as they are typically used. The full syntax of the commands is

by varlist1 [ (varlist2) ] [ , sort rc0 ]: stata_cmd

bysort varlist1 [ (varlist2) ] [ , rc0 ]: stata_cmd

Options

sort specifies that if the data are not already sorted by varlist, by should sort them.

rc0 specifies that even if the stata_cmd produces an error in one of the by-groups, then by is still to run the stata_cmd on the remaining by-groups. The default action is to stop when an error occurs. rc0 is especially useful when stata_cmd is an estimation command and some by-groups have insufficient observations.

Remarks and examples

Example 1

. use https://www.stata-press.com/data/r16/autornd
   (1978 Automobile Data)
   . keep in 1/20
      (54 observations deleted)
   . by mpg: egen mean_w = mean(weight)
      not sorted
      r(5);
   . sort mpg
   . by mpg: egen mean_w = mean(weight)
by — Repeat Stata command on subsets of the data

. list

<table>
<thead>
<tr>
<th>make</th>
<th>weight</th>
<th>mpg</th>
<th>mean_w</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Buick Electra</td>
<td>4000</td>
<td>15</td>
<td>3916.667</td>
</tr>
<tr>
<td>2. Cad. Deville</td>
<td>4500</td>
<td>15</td>
<td>3916.667</td>
</tr>
<tr>
<td>3. Chev. Impala</td>
<td>3500</td>
<td>15</td>
<td>3916.667</td>
</tr>
<tr>
<td>4. AMC Pacer</td>
<td>3500</td>
<td>15</td>
<td>3916.667</td>
</tr>
<tr>
<td>5. Buick Riviera</td>
<td>4000</td>
<td>15</td>
<td>3916.667</td>
</tr>
<tr>
<td>6. Cad. Eldorado</td>
<td>4000</td>
<td>15</td>
<td>3916.667</td>
</tr>
<tr>
<td>7. AMC Concord</td>
<td>3000</td>
<td>20</td>
<td>3350</td>
</tr>
<tr>
<td>8. Chev. Malibu</td>
<td>3000</td>
<td>20</td>
<td>3350</td>
</tr>
<tr>
<td>9. Buick Skylark</td>
<td>3500</td>
<td>20</td>
<td>3350</td>
</tr>
<tr>
<td>10. Buick LeSabre</td>
<td>3500</td>
<td>20</td>
<td>3350</td>
</tr>
<tr>
<td>11. Buick Regal</td>
<td>3500</td>
<td>20</td>
<td>3350</td>
</tr>
<tr>
<td>12. Chev. Monte Carlo</td>
<td>3000</td>
<td>20</td>
<td>3350</td>
</tr>
<tr>
<td>13. Chev. Nova</td>
<td>3500</td>
<td>20</td>
<td>3350</td>
</tr>
<tr>
<td>14. Cad. Seville</td>
<td>4500</td>
<td>20</td>
<td>3350</td>
</tr>
<tr>
<td>15. AMC Spirit</td>
<td>2500</td>
<td>20</td>
<td>3350</td>
</tr>
<tr>
<td>16. Buick Century</td>
<td>3500</td>
<td>20</td>
<td>3350</td>
</tr>
<tr>
<td>17. Chev. Monza</td>
<td>3000</td>
<td>25</td>
<td>2500</td>
</tr>
<tr>
<td>18. Buick Opel</td>
<td>2000</td>
<td>25</td>
<td>2500</td>
</tr>
</tbody>
</table>

by requires that the data be sorted. In the above example, we could have typed by mpg, sort: egen mean_w = mean(weight) or bysort mpg: egen mean_w = mean(weight) rather than the separate sort; all would yield the same results.

For more examples, see [U] 11.1.2 by varlist; [U] 11.5 by varlist: construct, and [U] 13.7 Explicit subscripting. For extended introductions with detailed examples, see Cox (2002) and Mitchell (2010, chap. 7).

Technical note

by repeats the stata_cmd for each group defined by varlist. If stata_cmd stores results, only the results from the last group on which stata_cmd executes will be stored.

References


Also see

[D] sort — Sort data
[D] statsby — Collect statistics for a command across a by list
[P] byable — Make programs byable
[P] foreach — Loop over items
[P] forvalues — Loop over consecutive values
[P] while — Looping
[U] 11.1.2 by varlist:
[U] 11.1.10 Prefix commands
[U] 11.4 varname and varlists
[U] 11.5 by varlist: construct