Combining datasets

You have two datasets that you wish to combine. Below, we will draw a dataset as a box where, in the box, the variables go across and the observations go down.

See [D] `append` if you want to combine datasets vertically:

\[
\begin{array}{c}
\text{A} \\
\end{array} +
\begin{array}{c}
\text{B} \\
\end{array} =
\begin{array}{c}
\text{A} \\
\text{B}
\end{array}
\]

`append` adds observations to the existing variables. That is an oversimplification because `append` does not require that the datasets have the same variables. `append` is appropriate, for instance, when you have data on hospital patients and then receive data on more patients.

See [D] `merge` if you want to combine datasets horizontally:

\[
\begin{array}{c}
\text{A} \\
\end{array} +
\begin{array}{c}
\text{B} \\
\end{array} =
\begin{array}{c}
\text{A} \\
\text{B}
\end{array}
\]

`merge` adds variables to the existing observations. That is an oversimplification because `merge` does not require that the datasets have the same observations. `merge` is appropriate, for instance, when you have data on survey respondents and then receive data on part 2 of the questionnaire.

There is another way to combine datasets horizontally, or more precisely, hierarchically, by loading them into separate frames and linking them. See [D] `frlink` for a discussion of when you might want to use `merge` versus `frlink`.

See [D] `joinby` when you want to combine datasets horizontally but form all pairwise combinations within group:

\[
\begin{array}{c}
\text{A} \\
\end{array} \times
\begin{array}{c}
\text{B} \\
\end{array} =
\begin{array}{c}
\text{A} \\
\text{B}
\end{array}
\]
joinby is similar to merge but forms all combinations of the observations where it makes sense. joinby would be appropriate, for instance, where A contained data on parents and B contained data on their children. joinby *familyid* would form a dataset of each parent joined with each of his or her children.

Also see [D] cross for a less frequently used command that forms every pairwise combination of two datasets.

See Mitchell (2020, chap. 7) for more information on combining datasets in Stata.

References


