Title

stata.com

Intro 2 — Data layout

Description Remarks and examples Also see

Description

This introduction describes the data layout required by all `cm` commands and describes how to `cmset` your data.

Remarks and examples

stata.com

Remarks are presented under the following headings:

Data layout for choice models

`cm` commands require data in a different form than the usual Stata data format. Typically in Stata, a single Stata observation corresponds to a single statistical observation—that is why Stata calls rows in a Stata dataset "observations". But `cm` commands need multiple Stata observations to hold the data for a single statistical observation. So as not to confuse statistical observations with Stata observations, we call a single statistical observation a "case" and use this terminology throughout the CM manual.

Here is an example of choice data. We show data for the first three individuals.

```
. use https://www.stata-press.com/data/r16/carchoice
(Car choice data)
. list consumerid car purchase gender income dealers if consumerid <= 3,
> sepby(consumerid) abbrev(10)
```

	consumerid	car	purchase	gender	income	dealers
1.	1	American	1	Male	46.7	9
2.	1	Japanese	0	Male	46.7	11
3.	1	European	0	Male	46.7	5
4.	1	Korean	0	Male	46.7	1
5.	2	American	1	Male	26.1	10
6.	2	Japanese	0	Male	26.1	7
7.	2	European	0	Male	26.1	2
8.	2	Korean	0	Male	26.1	1
9.	3	American	0	Male	32.7	8
10.	3	Japanese	1	Male	32.7	6
11.	3	European	0	Male	32.7	2

These fictitious data represent persons who purchased a car with their choices categorized by the nationality of the manufacturer, American, Japanese, European, or Korean. The first variable is `consumerid`, a variable identifying individual consumers; it is called the case ID variable.

The second variable shown is `car`, which holds the possible choices available to the consumer. The possible choices are called "alternatives", and this variable is referred to as the alternatives variable. We see that the first two consumers had all four nationalities of cars as alternatives. The third had only American, Japanese, and European as alternatives because there were no Korean dealerships in his or her community.

The third variable `purchase` is a 0/1 variable indicating which car the person purchased. For discrete choice models (`cmclogit`, `cmmprobit`, `cmmixlogit`, and `cmxtmixlogit`), this variable is the dependent variable in the estimation.

The variables `gender` and `income` are case-specific variables; they are constant within case. The variable `dealers` contains the number of dealerships of each nationality that are located in the consumer's community. It varies both by alternative and by individual. It is an alternative-specific variable. The case-specific variables and the alternative-specific variables will be used as independent variables in the estimation. It is important to distinguish between them because they are handled differently by the estimation commands and grouped separately when you run the command. See [CM] **Intro 5** for examples.

If you are familiar with Stata, you know that this data arrangement is called "long data". There are multiple Stata observations for each distinct value of the case ID variable. Wide data would be just one Stata observation for each case. All `cm` commands require data be in the long form.

The long-data format has implications for how missing values are handled by the `cm` commands. By default, any missing value within any of the observations for a case causes the entire case to be dropped from the analysis. The option `altwise` (meaning alternativewise), which all `cm` commands allow, causes only the observations with missing values to be dropped. See [CM] **cmsample** for a longer discussion about missing values. See example 3 in [CM] **cmclogit** for an estimation example.

cmset: Cross-sectional data

If a command begins with `cm`, you must `cmset` your data before you can run the command.

For cross-sectional data with identified alternatives, we pass the case ID variable and alternatives variable as arguments:

```
. cmset consumerid car
note: alternatives are unbalanced across choice sets; choice sets of different
      sizes found
            caseid variable:  consumerid
      alternatives variable:  car
```

The command echoed back the variable names that we set and, in this instance, also displayed a message, "alternatives are unbalanced across choice sets; choice sets of different sizes found." A "choice set" is the set of available alternatives for a case. This message is merely saying that the number of alternatives per case differs across the cases.

To see a tabulation of the choice sets, we type `cmchoiceset`:

```
. cmchoiceset
Tabulation of choice-set possibilities
```

Choice set	Freq.	Percent	Cum.
1 2 3	380	42.94	42.94
1 2 3 4	505	57.06	100.00
Total	885	100.00	

```
Total is number of cases.
. label list nation
nation:
           1 American
           2 Japanese
           3 European
           4 Korean
```

The output shows there are two choice sets, $\{1, 2, 3\}$ and $\{1, 2, 3, 4\}$. We also listed the value label `nation`, which is the value label for the alternatives variable `car`, to see the correspondence between the numerical values and the nationalities. The two choice sets are all four nationalities and all nationalities except Korean. See [CM] **cmchoiceset** for more ways to use this command.

For some CM estimators, such as `cmmixlogit`, having explicitly identified alternatives is optional. For the model fit by the `cmrologit` estimator, the alternatives are not identified, so there is no alternatives variable. When there is no alternatives variable, we `cmset` our data using the option `noalternatives` and pass the case ID variable as an argument:

```
. cmset consumerid, noalternatives
note: alternatives are unbalanced across choice sets; choice sets of different
      sizes found
            caseid variable:  consumerid
   no alternatives variable
```

cmset: Panel data

Here is an example of panel choice data:

```
. use https://www.stata-press.com/data/r16/transport, clear
(Transportation choice data)
. list id t alt if id == 1, sepby(t)
```

	id	t	alt
1.	1	1	Car
2.	1	1	Public
3.	1	1	Bicycle
4.	1	1	Walk
5.	1	2	Car
6.	1	2	Public
7.	1	2	Bicycle
8.	1	2	Walk
9.	1	3	Car
10.	1	3	Public
11.	1	3	Bicycle
12.	1	3	Walk

The first variable, id, is an ID for individuals, and the second variable, t, is the time. The set of data for an individual makes up a "panel", so the individual ID is the panel ID.

For panel choice data, cmset takes three variables—when there is an alternatives variable, as there is in this example. The first variable identifies the panels, the second gives the time, and the third is the alternatives variable.

```
. cmset id t alt
panel data: panels id and time t
note: case identifier _caseid generated from id t
note: panel by alternatives identifier _panelaltid generated from id alt

                     caseid variable:  _caseid
                alternatives variable:  alt
      panel by alternatives variable:  _panelaltid (strongly balanced)
                       time variable:  t, 1 to 3
                               delta:  1 unit

note: data have been xtset
```

cmset has created two new variables: _caseid and _panelaltid. See example 2 in [CM] **cmset** for details about these variables. You do not need to concern yourself with them, however. Just leave them in your dataset, and the cm commands will use them automatically to make things work.

Also see

[CM] **Intro 3** — Descriptive statistics

[CM] **cmchoiceset** — Tabulate choice sets

[CM] **cmsample** — Display reasons for sample exclusion

[CM] **cmset** — Declare data to be choice model data