

cmset — Declare data to be choice model data

[Description](#)
[Options](#)[Quick start](#)
[Remarks and examples](#)[Menu](#)
[Stored results](#)[Syntax](#)
[Also see](#)

Description

`cmset` manages the choice model settings of a dataset. You use `cmset` to declare the data in memory to be choice model data. With cross-sectional data, you designate which variables identify cases and alternatives. With panel data, you designate which variables identify panels, time periods, and alternatives. You must `cmset` your data before you can use the other `cm` commands.

`cmset` without arguments displays how the data are currently set. `cmset` also sorts the data based on the variables that identify cases, alternatives, and panels.

Quick start

Declare dataset to be choice model data with case identifier `caseid` and alternatives (choice-set) identifier `choiceset`

```
cmset caseid choiceset
```

Declare dataset to be choice model data with unspecified alternatives

```
cmset caseid, noalternatives
```

Declare dataset to be panel choice model data with panel identifier `pvar`, time identifier `tvar`, and alternatives identifier `choiceset`

```
cmset pvar tvar choiceset
```

Declare dataset to be panel choice model data with unspecified alternatives

```
cmset pvar tvar, noalternatives
```

Indicate that observations in the panel choice model data are made monthly; `tvar2` is not formatted

```
cmset pvar tvar2 choiceset, monthly
```

As above, and apply `%tm` format to `tvar2`

```
cmset pvar tvar2 choiceset, format(%tm)
```

View `cm` settings

```
cmset
```

Menu

Statistics > Choice models > Setup and utilities > Declare data to be choice model data

Syntax

Declare data to be cross-sectional choice model data

```
cmset caseidvar altvar [, force]
cmset caseidvar, noalternatives
```

Declare data to be panel choice model data

```
cmset panelvar timevar altvar [, tsoptions force]
cmset panelvar timevar, noalternatives
```

Display how data are currently cmset

```
cmset
```

Clear cm settings

```
cmset, clear
```

caseidvar identifies the cases in the cross-sectional data syntax.

altvar identifies the alternatives (choice sets).

panelvar identifies the panels, and *timevar* identifies the times within panels.

<i>tsoptions</i>	Description
<i>unitoptions</i>	specify units of <i>timevar</i>
<i>deltaoption</i>	specify period between observations in <i>timevar</i> units

<i>unitoptions</i>	Description
(<i>default</i>)	<i>timevar</i> 's units to be obtained from <i>timevar</i> 's display format
<i>clocktime</i>	<i>timevar</i> is %tc: 0 = 1jan1960 00:00:00.000, 1 = 1jan1960 00:00:00.001, ...
<i>daily</i>	<i>timevar</i> is %td: 0 = 1jan1960, 1 = 2jan1960, ...
<i>weekly</i>	<i>timevar</i> is %tw: 0 = 1960w1, 1 = 1960w2, ...
<i>monthly</i>	<i>timevar</i> is %tm: 0 = 1960m1, 1 = 1960m2, ...
<i>quarterly</i>	<i>timevar</i> is %tq: 0 = 1960q1, 1 = 1960q2, ...
<i>halfyearly</i>	<i>timevar</i> is %th: 0 = 1960h1, 1 = 1960h2, ...
<i>yearly</i>	<i>timevar</i> is %ty: 1960 = 1960, 1961 = 1961, ...
<i>generic</i>	<i>timevar</i> is %tg: 0 = ?, 1 = ?, ...
<i>format(%fmt)</i>	specify <i>timevar</i> 's format and then apply default rule

In all cases, negative *timevar* values are allowed.

<i>deltaoption</i>	Example
<i>delta(#)</i>	delta(1) or delta(2)
<i>delta((exp))</i>	delta((7*24))
<i>delta(# units)</i>	delta(7 days) or delta(15 minutes) or delta(7 days 15 minutes)
<i>delta((exp) units)</i>	delta((2+3) weeks)

Allowed units for %tc and %tC *timevars* are

seconds	second	secs	sec
minutes	minute	mins	min
hours	hour		
days	day		
weeks	week		

and for all other %t *timevars* are

days	day
weeks	week

Options

`noalternatives` specifies that alternatives are not explicitly identified. That is, there is no alternatives variable. The default is that you must specify an alternatives variable.

`force` suppresses error messages caused by the alternatives variable *altvar*. This option is rarely used. The alternatives variable must be free of errors before `cm` commands can run, so this option changes only the point at which error messages will be issued. One use of the `force` option is to specify it with `cmset` and then run `cmsample` to identify the observations with bad values for the alternatives variable. `force` does not suppress all error messages. Error messages in the case ID variable and error messages in the time variable for panel data are not suppressed.

unitoptions `clocktime`, `daily`, `weekly`, `monthly`, `quarterly`, `halfyearly`, `yearly`, `generic`, and `format(%fmt)` specify the units in which *timevar* is recorded when *timevar* is specified.

timevar will often simply be a variable of counts such as 1, 2, ..., or years such as 2001, 2002, In other cases, *timevar* will be a formatted %t variable; see [D] [Datetime](#). In any of these cases, you do not need to specify a *unitoption*.

Only when *timevar* is an unformatted time variable would you use these options. When you `cmset` panel choice model data, it becomes `xtset` as well. These options are simply passed to `xtset`. See [XT] [xtset](#) for option details.

`delta()` specifies the period of *timevar* and is commonly used when *timevar* is %tc or %tC. `delta()` is rarely used with other %t formats or with unformatted time variables. If `delta()` is not specified, `delta(1)` is assumed. See [XT] [xtset](#) for option details.

`clear`—used in `cmset`, `clear`—makes Stata forget that the data were ever `cmset`. This option is rarely used. Note that if you `cmset` your data as panel choice model data with an alternatives variable, they also become `xtset`. Typing `cmset`, `clear` does not clear the `xt` settings. To do this, you must type `xtset`, `clear` as well.

Remarks and examples

[stata.com](http://www.stata.com)

`cmset` declares the dataset in memory to be choice model data. You need to do this before you can use the other `cm` commands.

`cmset` sets cross-sectional choice data and panel choice data. The usual syntax for cross-sectional data is to give `cmset` two variables:

```
cmset caseidvar altvar
```

The case ID variable *caseidvar* must be numeric, and its values must be integers. The variable *altvar* containing the alternatives can be either numeric or string.

The usual syntax for panel data is to give `cmset` three variables:

```
cmset panelvar timevar altvar
```

The variable *panelvar* identifies panels, which are typically IDs for individuals or decision makers. The variable *timevar* identifies times within panels, points at which choices were made. Both *panelvar* and *timevar* must be numeric, and both must contain integers only.

For some choice models, alternatives are not explicitly identified. Alternatives are known only by their characteristics as given by alternative-specific variables. In this case, the syntax for cross-sectional data is

```
cmset caseidvar, noalternatives
```

and the syntax for panel data is

```
cmset panelvar timevar, noalternatives
```

For a brief introduction to other choice models, see [\[CM\] Intro 4](#).

► Example 1: Cross-sectional choice data

Here is an example of cross-sectional choice data:

```
. use https://www.stata-press.com/data/r16/carchoice
(Car choice data)
. list consumerid car purchase if consumerid <= 4, sepby(consumerid) abbr(10)
```

	consumerid	car	purchase
1.	1	American	1
2.	1	Japanese	0
3.	1	European	0
4.	1	Korean	0
5.	2	American	1
6.	2	Japanese	0
7.	2	European	0
8.	2	Korean	0
9.	3	American	0
10.	3	Japanese	1
11.	3	European	0
12.	4	American	1
13.	4	Japanese	0
14.	4	European	0

The variable `consumerid` is the case ID variable, and the variable `car` defines the alternatives. These fictitious data represent persons who purchased a car with their choices categorized by the nationality of the manufacturer, American, Japanese, European, or Korean.

To declare the data to be `cm` data, we type

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

We have to `cmset` our data only once if we save our data after we `cmset` it. Let's illustrate this. Typing `cmset` without arguments shows the current settings.

```
. save carchoice_cmset
file carchoice_cmset.dta saved

. use carchoice_cmset
(Car choice data)

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

For these data, the choice sets are unbalanced, and `cmset` gave us a message telling us this. If we want to see the distinct choice-set possibilities, we can 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.



► Example 2: Data errors with `cmset`

If there were errors in the alternatives variable, `cmset` would give an error message. Here is an example with a dataset where we added errors:

```
. use https://www.stata-press.com/data/r16/carchoice_errors, clear
(Car choice data with errors)

. cmset consumerid car
at least one choice set has more than one instance of the same alternative
r(459);
```

When `cmset` detects errors in the alternatives variable, you may want to type `cmset` again with the option `force`, and then use `cmsample`:

```
. cmset consumerid car, force
note: at least one choice set has more than one instance of the same
      alternative
      caseid variable:  consumerid
      alternatives variable:  car

. cmsample, generate(flag)
```

Reason for exclusion	Freq.	Percent	Cum.
observations included	3,153	99.78	99.78
repeated alternatives within case*	7	0.22	100.00
Total	3,160	100.00	

* indicates an error

```
. list consumerid car flag if flag != 0, sepby(consumerid) abbr(10)
```

	consumerid	car	flag
397.	111	American	repeated alternatives within case*
398.	111	Japanese	repeated alternatives within case*
399.	111	Japanese	repeated alternatives within case*
1035.	290	American	repeated alternatives within case*
1036.	290	Japanese	repeated alternatives within case*
1037.	290	Japanese	repeated alternatives within case*
1038.	290	Korean	repeated alternatives within case*

Some `cm` estimators such as `cmrologit` do not require an alternatives variable. In this case, you use the `noalternatives` option and just specify the case ID variable:

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



▶ Example 3: Panel choice data

When you have panel choice data, you will have a panel ID variable and a time variable. Typically, you will also have a variable specifying the alternatives.

Here is an example in which `id` is the panel ID variable, `t` is the time variable, and variable `alt` contains the alternatives. The first panel of these data looks like

```
. 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

To `cmset` the data, we type

```
. 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
```

Look at the notes displayed by `cmset`. It has created two new variables: `_caseid` and `_panelaltid`. Let's list their values for the first two panels.

```
. sort id t alt
. list id t alt _caseid _panelaltid if inlist(id, 1, 2), sepby(t) abbr(11)
```

	id	t	alt	_caseid	_panelaltid
1.	1	1	Car	1	1
2.	1	1	Public	1	2
3.	1	1	Bicycle	1	3
4.	1	1	Walk	1	4
5.	1	2	Car	2	1
6.	1	2	Public	2	2
7.	1	2	Bicycle	2	3
8.	1	2	Walk	2	4
9.	1	3	Car	3	1
10.	1	3	Public	3	2
11.	1	3	Bicycle	3	3
12.	1	3	Walk	3	4
13.	2	1	Car	4	5
14.	2	1	Public	4	6
15.	2	1	Bicycle	4	7
16.	2	1	Walk	4	8
17.	2	2	Car	5	5
18.	2	2	Public	5	6
19.	2	2	Bicycle	5	7
20.	2	2	Walk	5	8
21.	2	3	Car	6	5
22.	2	3	Public	6	6
23.	2	3	Bicycle	6	7
24.	2	3	Walk	6	8

`_caseid` is a variable that identifies cases. For choice model data, remember that a case is a single statistical observation but consists of multiple Stata observations. Each distinct value of panel ID \times time represents a single statistical observation, that is, a case. The values of `_caseid` correspond to the distinct values of panel ID \times time, in this example, the values of `id \times t`.

`_panelaltid` is a variable that uniquely identifies the distinct values of panel ID \times alternative. Why do you need this variable? It is created so you can use [Stata's time-series operators](#). Imagine that you want to include lags of alternative-specific variables in your model. The lags must be specific to the alternative, and Stata's time-series lag operator needs to know how to do this.

When you `cmset` panel data with specified alternatives, your data are automatically `xtset`. You can type `xtset` to see the settings:

```
. xtset
      panel variable:  _panelaltid (strongly balanced)
      time variable:  t, 1 to 3
                   delta: 1 unit
```

`_panelaltid` becomes the “panel” identifier viewing the data as `xt` data. See [CM] [Intro 7](#) and [CM] [cmxtmixlogit](#) for more on using time-series operators with panel CM data.

`cmxtmixlogit` allows you to fit a model with unspecified alternatives. To do this, you use the option `noalternatives`:

```
. use https://www.stata-press.com/data/r16/transport, clear
(Transportation choice data)
. cmset id t, noalternatives
panel data: panels id and time t
note: case identifier _caseid generated from id t
      caseid variable:  _caseid
      no alternatives variable
      time variable:  t, 1 to 3
                   delta: 1 unit
```

`_caseid` is again created, and its values are the same as in the previous `cmset` results.

There is no `_panelaltid` variable because there are no specified alternatives. The data are not `xtset` because there is no way to match up the alternatives.

```
. xtset
panel variable not set; use xtset varname ...
r(459);
```

Because the data are not `xtset`, you cannot use time-series operators for panel CM models with unspecified alternatives.

◀

Stored results

`cmset` stores the following in `r()`:

Scalars

<code>r(n_cases)</code>	number of cases
<code>r(n_alt_min)</code>	minimum number of alternatives per case
<code>r(n_alt_avg)</code>	average number of alternatives per case
<code>r(n_alt_max)</code>	maximum number of alternatives per case
<code>r(altvar_min)</code>	minimum of alternatives variable (if set when numeric)
<code>r(altvar_max)</code>	maximum of alternatives variable (if set when numeric)

Macros

<code>r(caseid)</code>	name of case ID variable
<code>r(altvar)</code>	name of alternatives variable (if set)

For panel data, `cmset` also stores the following in `r()`:

Scalars

<code>r(imin)</code>	minimum panel ID
<code>r(imax)</code>	maximum panel ID
<code>r(tmin)</code>	minimum time
<code>r(tmax)</code>	maximum time
<code>r(tdelta)</code>	delta
<code>r(gaps)</code>	1 if there are gaps, 0 otherwise

Macros

<code>r(origpanelvar)</code>	name of original panel variable passed to <code>cmset</code>
<code>r(panelvar)</code>	name of panel variable
<code>r(timevar)</code>	name of time variable
<code>r(tdeltas)</code>	formatted delta
<code>r(tmins)</code>	formatted minimum time
<code>r(tmaxs)</code>	formatted maximum time
<code>r(tsfmt)</code>	<i>%fmt</i> of time variable
<code>r(unit)</code>	units of time variable: <code>Clock</code> , <code>clock</code> , <code>daily</code> , <code>weekly</code> , <code>monthly</code> , <code>quarterly</code> , <code>halfyearly</code> , <code>yearly</code> , or <code>generic</code>
<code>r(unit1)</code>	units of time variable: <code>C</code> , <code>c</code> , <code>d</code> , <code>w</code> , <code>m</code> , <code>q</code> , <code>h</code> , <code>y</code> , or <code>"</code>
<code>r(balanced)</code>	<code>unbalanced</code> , <code>weakly balanced</code> , or <code>strongly balanced</code> ; a set of panels are <code>strongly balanced</code> if they all have the same time values, otherwise <code>weakly balanced</code> if same number of time values, otherwise <code>unbalanced</code>

Also see

- [CM] [cmchoiceset](#) — Tabulate choice sets
- [CM] [cmsample](#) — Display reasons for sample exclusion
- [CM] [cmsummarize](#) — Summarize variables by chosen alternatives
- [CM] [cmtab](#) — Tabulate chosen alternatives
- [XT] [xtset](#) — Declare data to be panel data