The Stata Journal (2007)
7, Number 4, pp. 590–592

Stata tip 55: Better axis labeling for time points and time intervals

Nicholas J. Cox
Department of Geography
Durham University
Durham City, UK
n.j.cox@durham.ac.uk

Plots of time-series data show time on one axis, usually the horizontal or x axis. Unless the number of time points is small, axis labels are usually given only for selected times. Users quickly find that Stata's default time axis labels are often not suitable for use in public. In fact, the most suitable labels may not correspond to *any* of the data points. This will arise when it is better to label longer time intervals, rather than any individual times in the dataset.

For example,

```
. webuse turksales
```

reads in 40 quarterly observations for 1990q1 to 1999q4 with a response variable of turkey sales. The default time axis labels with both `line sales t` and `tsline sales` are 1990q1, 1992q3, 1995q1, 1997q3, and 2000q1. These are not good choices for any purpose, even exploration of the data in private.

Label choice is partly a matter of taste, but you might well agree with Stata that labeling every time point would be busy and the result difficult to read. With 40 quarterly values, possible choices include one point per year (10 labels) and one point every other year (5 labels). One possibility is to label every fourth quarter, as that is usually the quarter with highest turkey sales. `summarize` reveals that the times range from 120 to 159 quarters (0 means the first quarter of 1960), so we can type

```
. line sales t, xlabel(123(4)159)
```

Note how we use a *numlist*, `123(4)159`, to avoid spelling out every value. The step length is 4 for four quarters. See [U] **11.1.8 numlist** or `help numlist` for more details of *numlists*. This graph too would need more work before publication, as the labels are still crowded. The text of the labels (e.g., 1990q4) may or may not be judged suitable, depending partly on the readership for the graph.

However, there is another choice: label time intervals (years) and mark the boundaries between those time intervals by ticks. Consider 1990. The four quarters in Stata's units are 120, 121, 122, and 123. Thus we could put text showing the year at a midpoint of 121.5 and ticks showing year boundaries at 119.5 and 123.5. For all years, we should use the *numlist* idea again with the following command to produce figure 1.

```
. line sales t, xtick(119.5(4)159.5, tlength(*1.5))
> xlabel(121.5(4)157.5, noticks format(%tqCY)) xtitle("")
```

gr0030

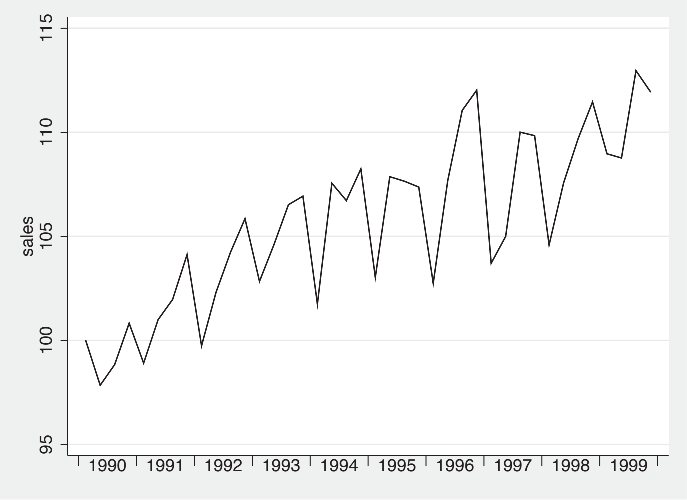

Figure 1: Turkey sales in each quarter. Time axis labels show years (with ticks suppressed) and time axis ticks show year ends.

The most important details here are suppressing the ticks for the axis labels and specifying a format for them. Cosmetic additions include lengthening the ticks compared with the default and suppressing the axis title, which would otherwise be the variable name `t` (or a variable label if it existed). It is usually clear from the labels what is being shown. Other possibilities include changing the text size for the axis label, changing the angle at which the axis label is shown, and suppressing the century by using a format like `%tqY`. Those may not be especially attractive, but nevertheless might be forced upon you by practicalities.

The main idea is clearly more general. The axis labels and the axis ticks need not correspond to each other, and it might be good to have fewer labels than ticks for longer series. Monthly and half-yearly data naturally yield to the same method, but use 12 or 2 and not 4 as the step length. Weekly and daily data are more awkward but still manageable.

If you were producing many similar graphs, you might want to automate this process to some degree. The mental arithmetic might easily be more challenging than in the turkey example. Let us imagine daily data for several years. Thus we could put ticks every January 1 and year labels every July 1. That will be adequate precision in practice. Find the first and last years in your data, if necessary by a command like `gen year = year(date)` followed by `summarize`. Suppose again that the years are 1990–1999. We can put the needed dates in local macros with a loop:

```
. forvalues y = 1990/1999 {
        local jan `jan´ `=mdy(1,1,`y´)´
        local jul `jul´ `=mdy(7,1,`y´)´
}
```

Each time around the loop the daily dates for January 1 and July 1 in each year are calculated on the fly with a call to the `mdy()` function and added to a macro. For more details, see [P] **forvalues** and [P] **macro**, the corresponding help files, or Cox (2002). Once done, the graph command is something like

```
. line whatever date, xlabel(`jul´, format(%tdCY) noticks)
> xtick(`jan´, tlength(*1.5))
```

A key requirement is that the local macros used in the graph command must be visible, by virtue of being in the same interactive session, do-file, or program. That is in essence what `local` means.

Calendar years, meaning here Western calendar years, are clearly not the only possibilities. You could use other boundaries and midpoints for years or other periods defined by other criteria (e.g., academic, financial, fiscal, hydrological, political, religious).

Reference

Cox, N. J. 2002. Speaking Stata: How to face lists with fortitude. *Stata Journal* 2: 202–222.

The Stata Journal (2007)
7, Number 4, p. 593

Software Updates

gr0012_1: Density probability plots. N. J. Cox. *Stata Journal* 5: 259–273.

> The program has been updated so that users of Stata 9 and later can use an `addplot()` option.

st0133_1: Fitting mixed logit models by using maximun simulated likelihood. A. R. Hole. *Stata Journal* 7: 388–401.

> New features include options for specifying weights (including sampling weights) and for obtaining robust and cluster−robust standard errors. The estimation speed has also been improved by using analytical instead of numerical derivatives when maximizing the simulated log-likelihood function. This change has the side effect of producing somewhat different estimation results compared with the previous version for some datasets and model specifications. The new `numerical` option may be used to replicate estimation results produced with the old version, but it should only be used for that purpose, as it causes the command to run slowly.