Stata and the newcomer

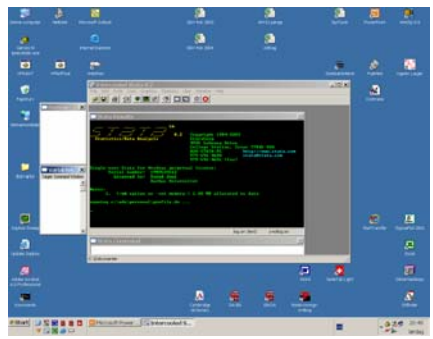
Svend Juul, 5 April 2004

29 March 2004: Meeting at a Danish hospital

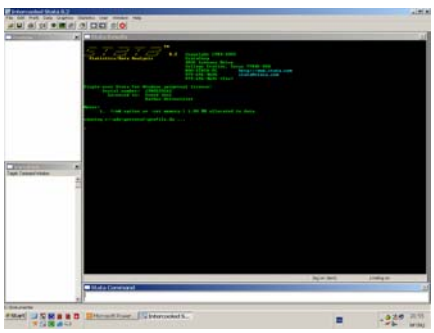
- We want to improve the research facilities for our staff. We need a good, flexible, affordable statistical package, including graphing facilities.
- ”Stata is the answer to that need.”
- ”I’m not so sure. People tell me that it is difficult to get started. It may be OK for a full-time researcher, but our junior doctors can’t spend weeks to find out how to use it.”
- ”I heard that SPSS is a lot friendlier.”
- ”But it’s expensive. Wouldn’t Excel do?”

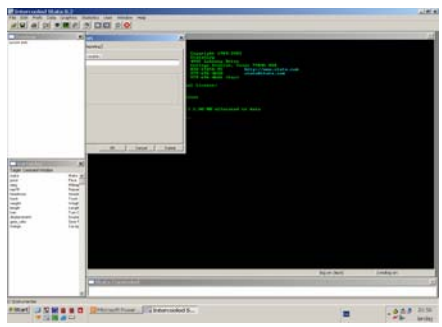
- First impression.
- Names are important
- Manual structure

I just opened my new Stata 8. It looks silly.



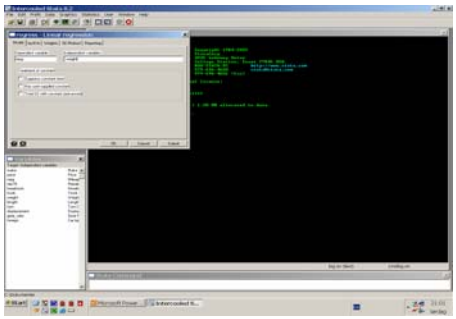
I do as [GSW] tells me.

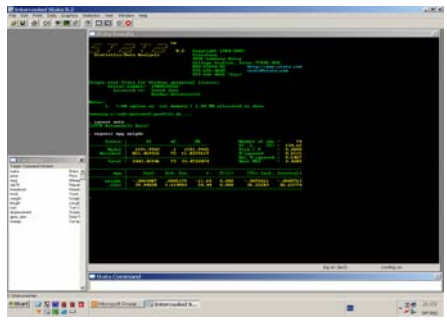




Oops! The Review window hides the dialog.

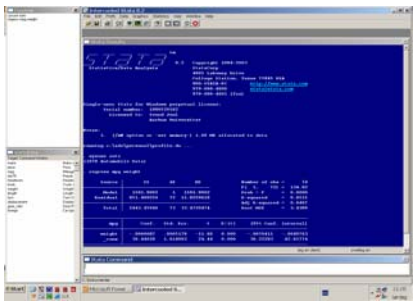
I must close the Review window.





Where is the Review window now?

As we all know (except the new user):
This is how it should be:



This was the user's first impression:

- It looks ugly (especially the Viewer window font).
- It behaves confusingly.
- The documentation [GSW] is misleading.

- *"I give up using Stata. It's unprofessional."*
- *"I give up using Stata. I'm not bright enough."*

The Results window is clumsy

- Previous output disappears.
- I can't edit (remove junk) before printing.
- I get confused by the many colours; they generate more heat than light.
- I can only move around using the mouse.
- We know the solution: use a log file
  - but most often I forgot to open one beforehand.

The Viewer window is clumsy

- No, actually it is perfect for looking up help.
- But it is not good for displaying, editing and printing output.
- The SMCL translation to fonts is strange.

	Column		
Row	1	2	Total
1	**17**	**11**	**28**
2	**34**	**14**	**48**
Total	**51**	**25**	**76**

Recommendation: Re-design

- Improve the immediate appearance after installation.
- Automatic log file – and an editable output window.
- (Forget SMCL-formatted output; just plain text).

- First impression
- Names are important
- Manual structure

Names are important

A

- **`summarize`** has the option **`meanonly`**.
- It displays nothing, but it saves results.
- No, it saves several results.

B

- I guess it displays the mean only.
- I guess it saves the mean only, then.
- Why, then, isn't the name **`nodisplay`**?

`table` – `tabulate`

`table` and **`tabulate`** with very similar names actually do very different things.

- **`tabulate`** does three very different things (**`tab1`**, **`tab2`** and tabsum).
- [R] **`tabulate`** lists 26 options. Only 4 are common to **`tab1`** and **`tab2`**.

tab1	tab2
	all
	cchi2
	cell
	chi2
	clrchi2
	column
	exact
	expected
	gamma
generate()	
	lrchi2
matcell()	matcell()
	matcol()

tab1	tab2
	matrow()
missing	missing
	nofreq
	nokey
nolabel	nolabel
plot	
	replace
	row
sort	
subpop()	subpop()
	taub
	V
	wrap

Suggestions

- Let **`tab1`**, **`tab2`** and **`tabsum`** be three separate commands, with separate manual entries.
- Now, the tabsum syntax is:
  `tabulate` *indpevar* [*indepvar*], **`summarize(`***depvar***`)`**
- Suggestion, like **`oneway`** and **`anova`**:
  `tabsum` *depvar indepvar* [*indepvar*] , *options*
- Other tabulation commands need rethinking too.

generate and **egen**

egen supplies some functions that ought to be in **generate**
- but I can't predict which functions belong where

```
generate y = max(x1, x2, x3)
egen y = rmax(x1 x2 x3)
```
They do the same, but with different syntax.

```
generate y = sum(x1)
egen y = sum(x1)
```
They do different things.
I don't recall which does what.

Suggestion:

- Incorporate as many as possible of **egen**'s functions in **generate**.
- New names for sum functions etc.:

rsum(*args*)	Sum of arguments within observation (**egen**'s **rsum**)
csum(*var*)	Cumulated sum (**generate**'s **sum**)
tsum(*var*)	Total sum, all observations (**egen**'s **sum**)

This is a trap:

anycommand **if age>60**
includes those with missing **age**
because **missing** is > 60 (!)

This is designed to generate errors.

- First impression
- Names are important
- Manual structure

Finding commands

- Commands in unpredictable places:
  - **dotplot** and **histogram** in [R]
  - **graph dot** and **graph bar** in [G]
- Is the alphabetic structure viable?
  - [U] is excellent
  - [G] is a disaster

Give up the alphabetic manual structure.
E.g. make a chapter on the family of fitting and smoothing functions:

[R] **fracpoly**	[G] **twoway lfitci**
[R] **lowess**	[G] **twoway lowess**
[R] **mkspline**	[G] **twoway mband**
[R] **smooth**	[G] **twoway mspline**
[G] **twoway fpfit**	[G] **twoway qfit**
[G] **twoway fpfitci**	[G] **twoway qfitci**
[G] **twoway lfit**	

- The gap between the Getting Started manual and the “real” manuals is huge.
- Make a Newcomer’s Guide (300 pages).
  More than [GS], less than [U] and [R].
  It should include commands for elementary data management and analysis.

Conclusion

- Many users learn to love Stata – some of us even get addicted.
- Many potential users feel discouraged
- Give high priority to:
  - The first impression
  - Friendliness towards new users
  - Manual structure
  - Names, names, names
- It can be done without harm to the current users.