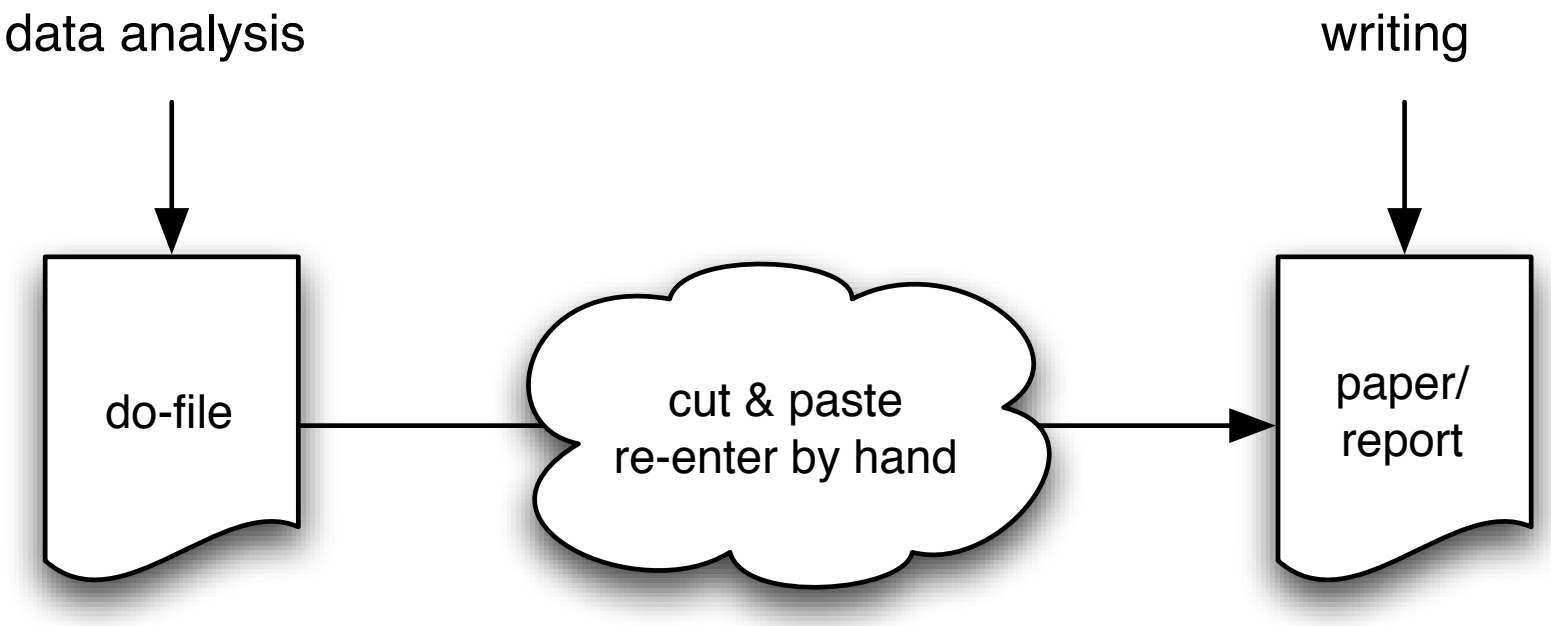
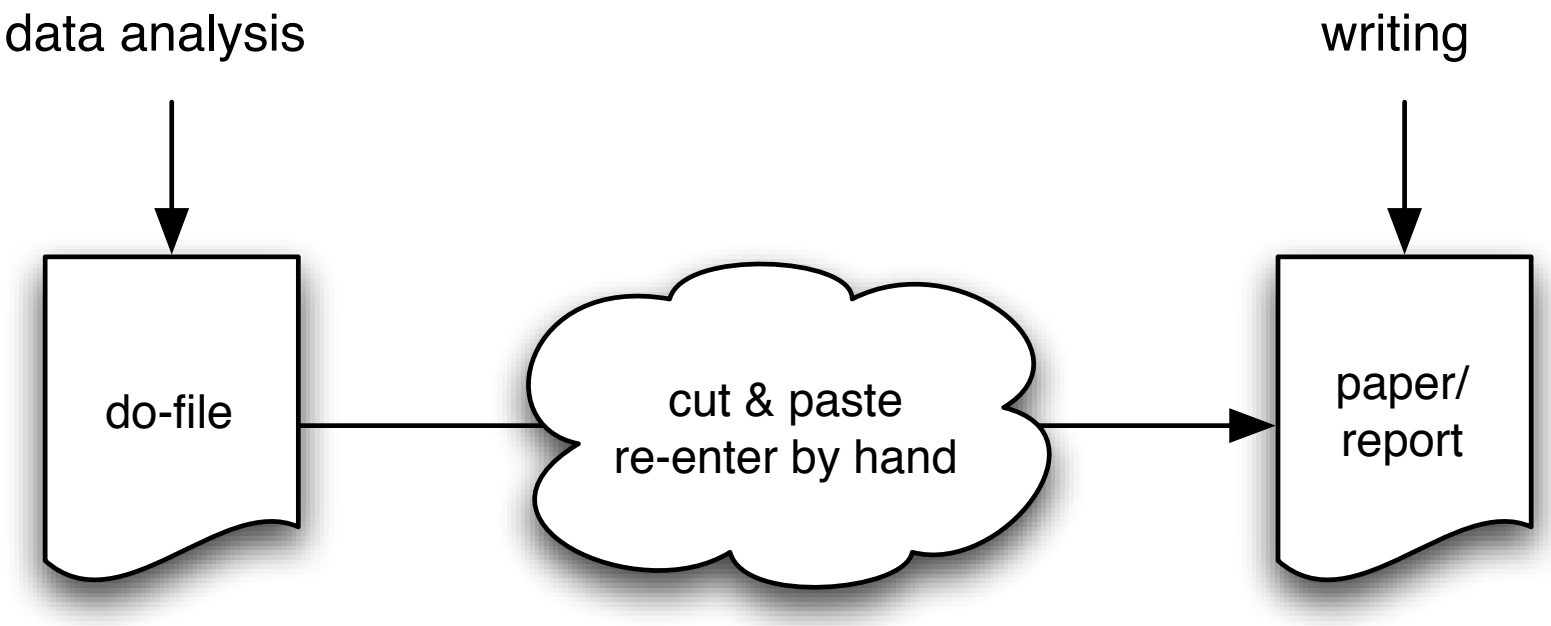
Reproducible Research Using Stata

L. Philip Schumm Ronald A. Thisted

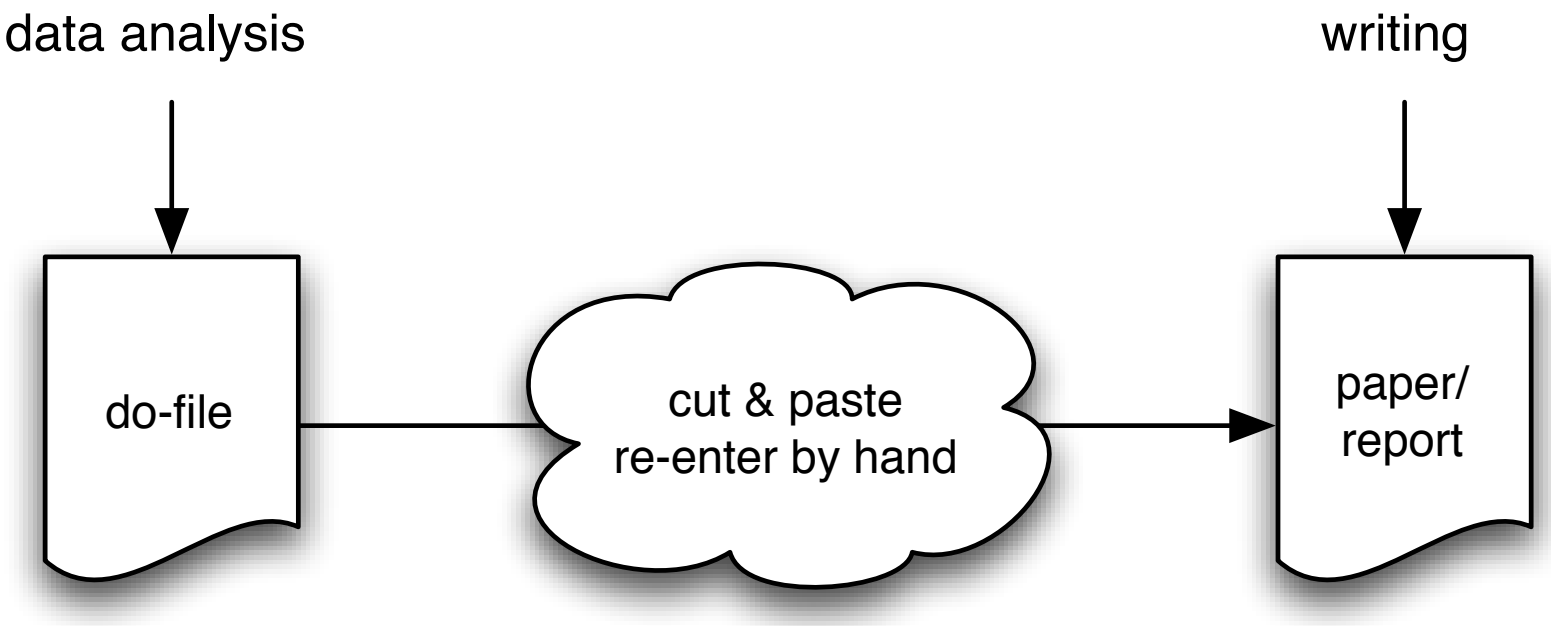
Department of Health Studies
University of Chicago

July 11, 2005


data analysis
writing
do-file
cut & paste
re-enter by hand
paper/
report

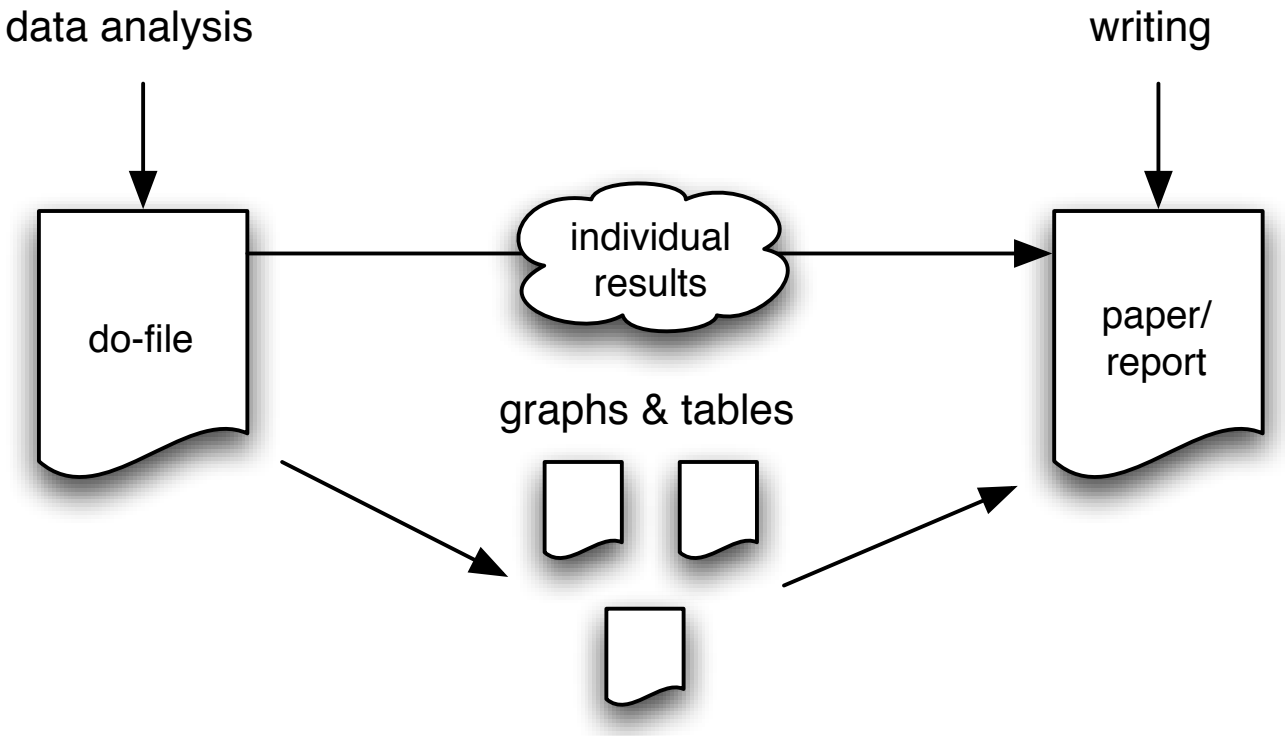


- Inefficient and time-consuming

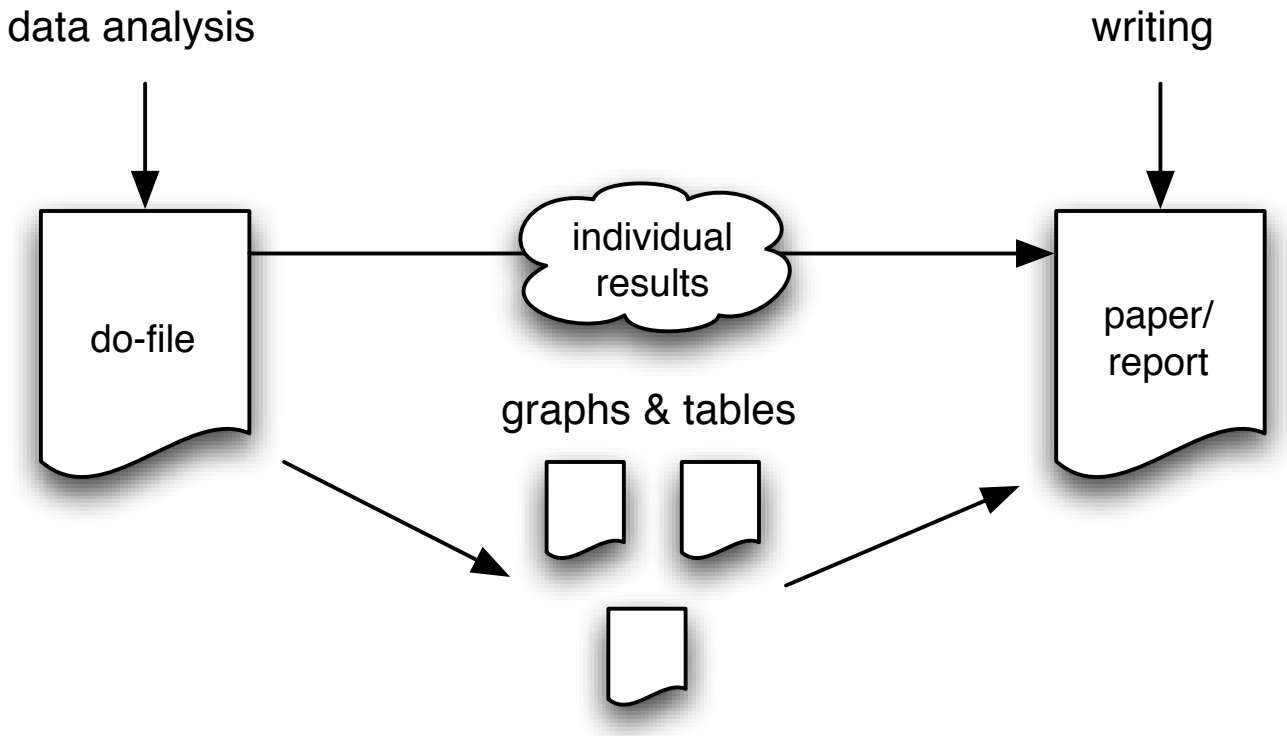


- Inefficient and time-consuming
- Can lead to non-reproducible results

A big improvement: Intermediary files

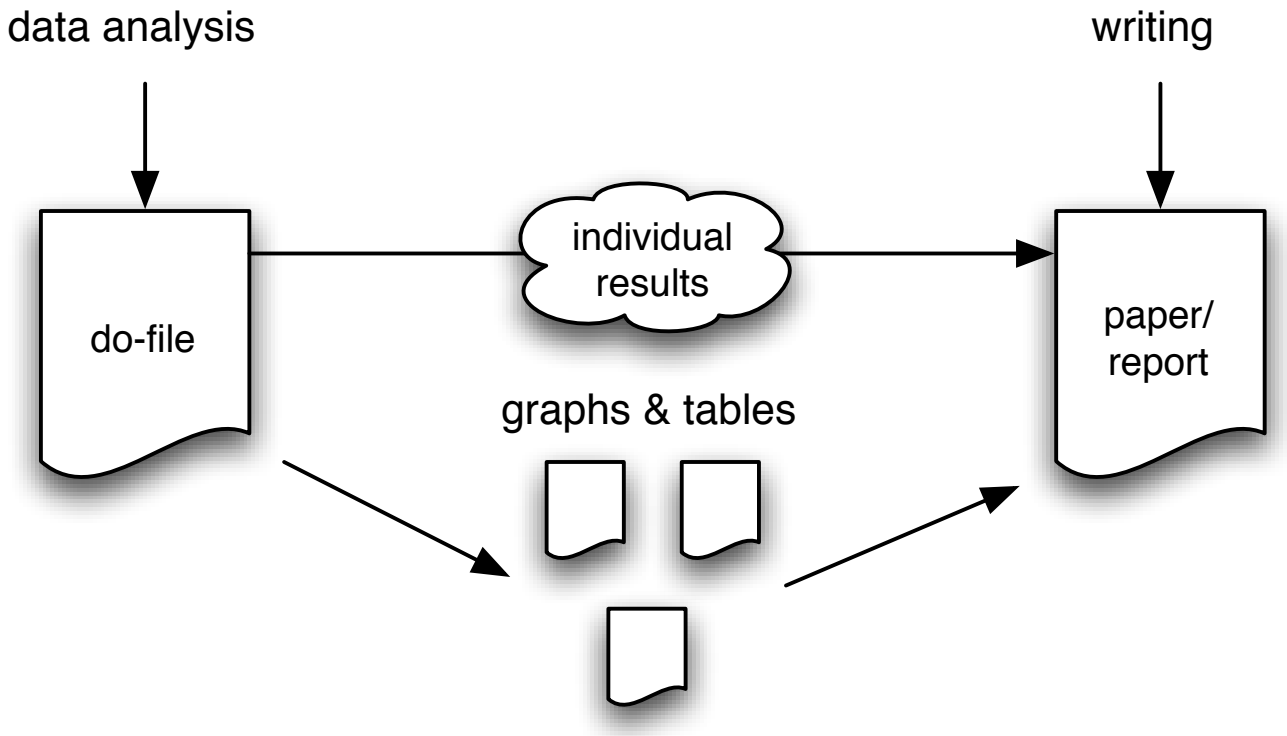


A big improvement: Intermediary files



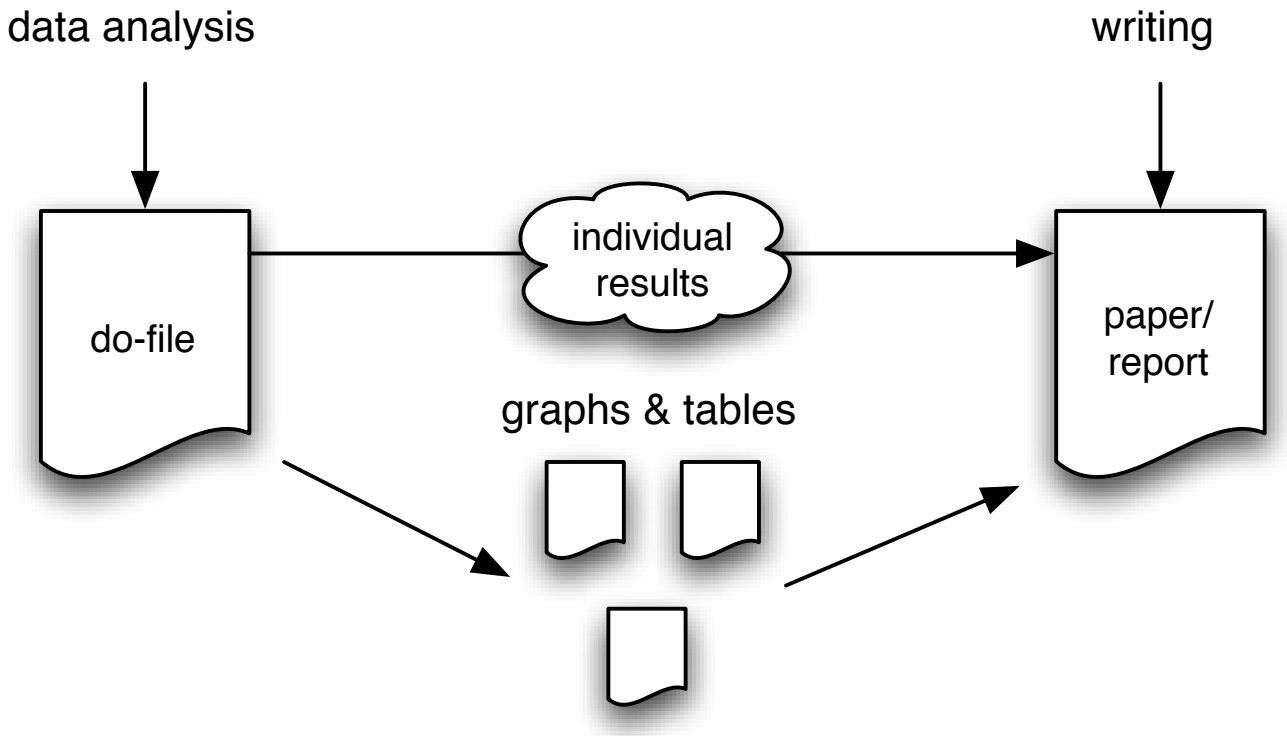
- Not all results automatically transferred

A big improvement: Intermediary files



- Not all results automatically transferred
- Can be difficult to manage

A big improvement: Intermediary files



- Not all results automatically transferred
- Can be difficult to manage
- Data analysis and writing still asynchronous

What is reproducible research?

- Emerging literature (e.g., Buckheit and Donoho, 1995; Gentleman and Lang, 2003)

What is reproducible research?

- Emerging literature (e.g., Buckheit and Donoho, 1995; Gentleman and Lang, 2003)
- Dynamic document composed of *code chunks* and *text chunks*

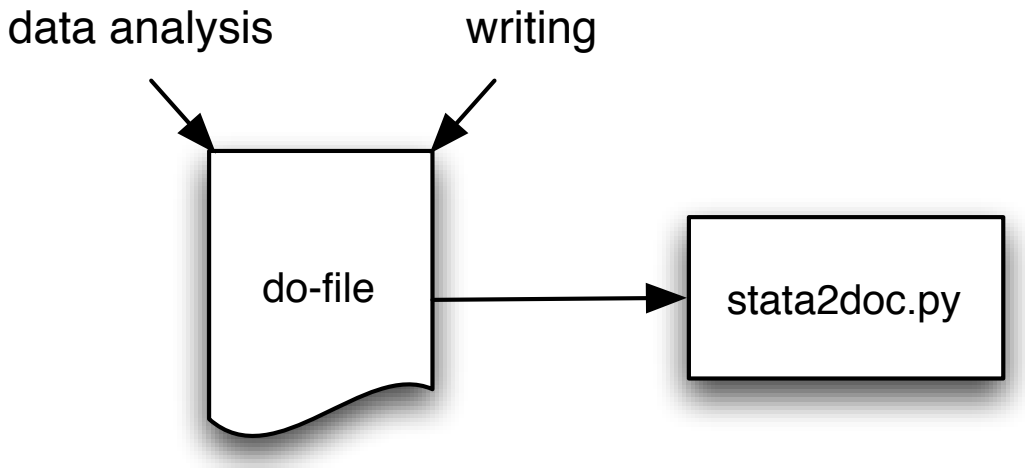
What is reproducible research?

- Emerging literature (e.g., Buckheit and Donoho, 1995; Gentleman and Lang, 2003)
- Dynamic document composed of *code chunks* and *text chunks*
- Literate programming (Knuth, 1992)
  - tangling
  - weaving

What is reproducible research?

- Emerging literature (e.g., Buckheit and Donoho, 1995; Gentleman and Lang, 2003)
- Dynamic document composed of *code chunks* and *text chunks*
- Literate programming (Knuth, 1992)
  - tangling
  - weaving
- R package called Sweave (Leisch, 2002)

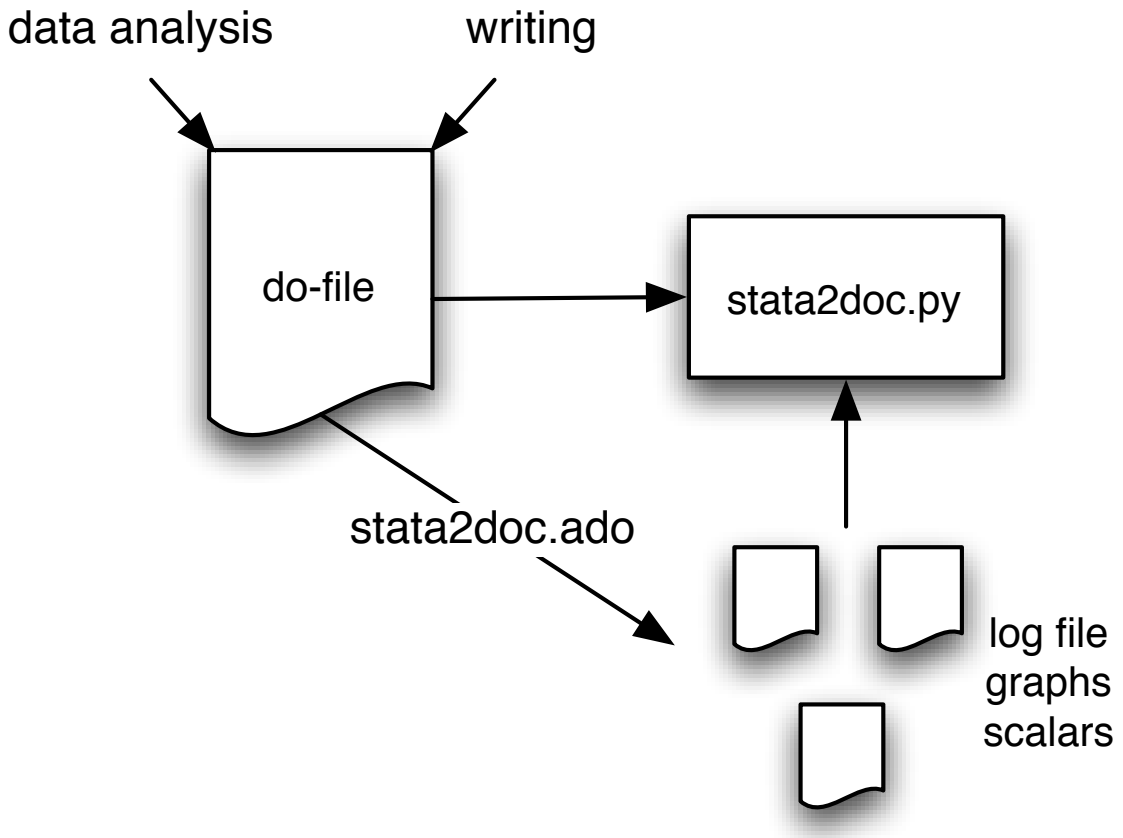
A "dynamic" do-file



Comments, commands, and docstrings

```
// Here is an example dynamic do-file.

* here is the docstring for the first command
sysuse auto

* weightsq equals weight squared
gen weightsq=weight^2

reg mpg weight weightsq foreign

/* As you can see, commands don't have to have
   docstrings. */
```

Two stata commands: stata2doc and s2d



Syntax

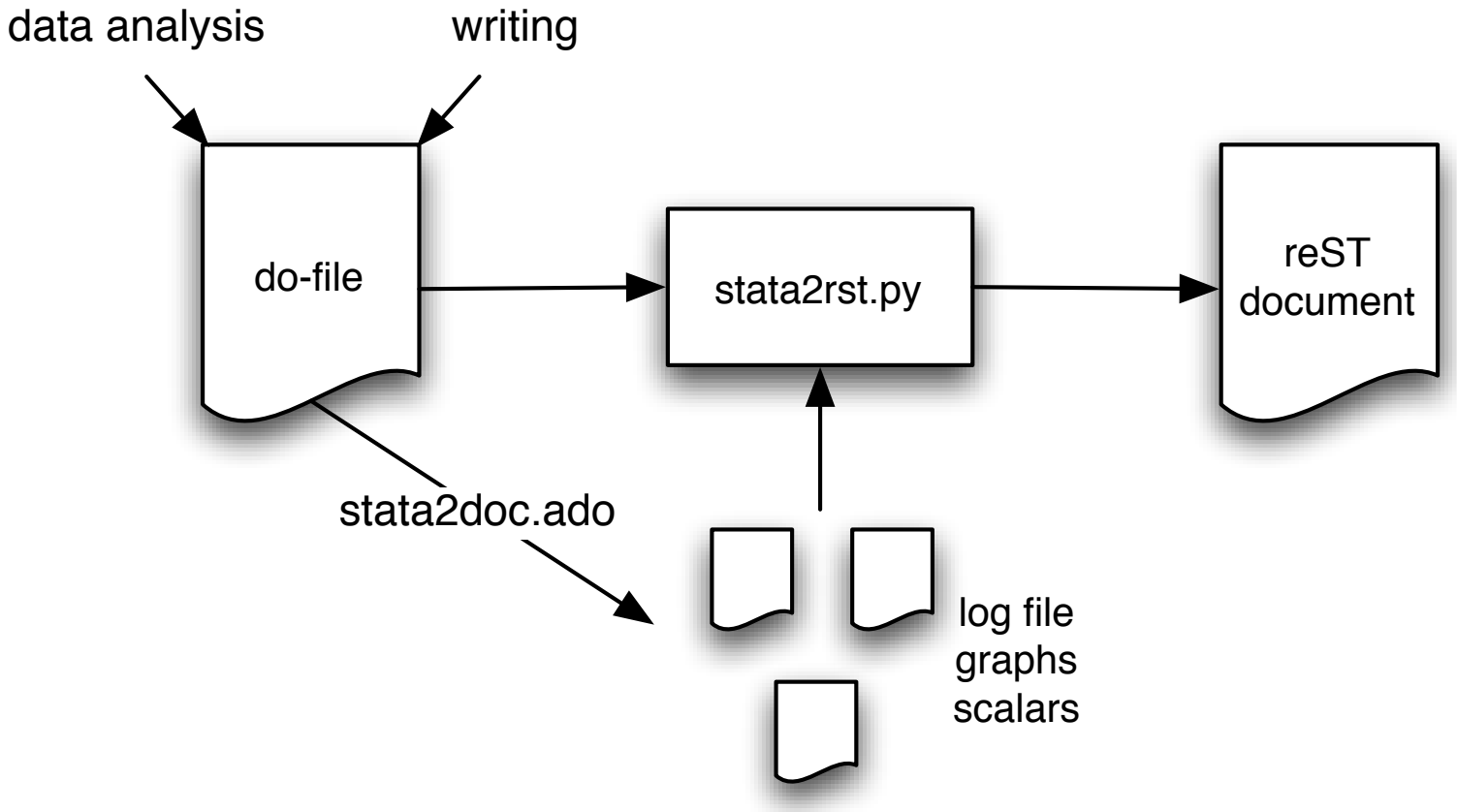
stata2doc using *do-file*, [dirname(*dirname*) linesize(#) as(*type*)
replace *override_options*]

s2d [*exp_list*, nodisplay table noisily warn name(*name*)] :

Examples of -s2d- usage

```
. s2d w2coef=_b[weightsq] rsq=e(r2): reg mpg weight weightsq foreign
<output omitted>

. scalar li
   s2d_rsq =   .69129599
s2d_w2coef =  1.591e-06

. s2d two = (1 + 1), noi

   s2d_two =           2
```

Putting it all together



What is reStructuredText?

- A plaintext markup syntax and parser system

What is reStructuredText?

- A plaintext markup syntax and parser system
- Intuitive, readable, and easy-to-use

What is reStructuredText?

- A plaintext markup syntax and parser system
- Intuitive, readable, and easy-to-use
- Powerful and extensible

What is reStructuredText?

- A plaintext markup syntax and parser system
- Intuitive, readable, and easy-to-use
- Powerful and extensible
- via Docutils may be translated into a variety of formats (e.g., HTML, LaTeX, PDF, Open Office)

(see http://docutils.sourceforge.net for more information)

Simple command: do-file

```
/*
----------------
A Simple Example
----------------

This is a *very* simple example in which I shall demonstrate the following:

1) a simple command
2) graphs
3) substitution
4) tables

The Venerable Auto Data
-----------------------

Let's start by reading them in:
*/

sysuse auto
```

Simple command: reStructuredText

```
---------------
A Simple Example
---------------

This is a *very* simple example in which I shall demonstrate the following:

1) a simple command
2) graphs
3) substitution
4) tables

The Venerable Auto Data
----------------------

Let's start by reading them in:

::

 . sysuse auto
 (1978 Automobile Data)
```

Simple command: PDF via LaTeX

A Simple Example

This is a *very* simple example in which I shall demonstrate the following:

1) a simple command
2) graphs
3) substitution
4) tables

The Venerable Auto Data

Let's start by reading them in:

```
. sysuse auto
(1978 Automobile Data)
```

Graph: do-file

```
* Now lets look at a boxplot comparing mpg between
* domestic and foreign.

* Boxplot comparing domestic and foreign.
gr box mpg, over(foreign) name(fig1)
```

Graph: reStructuredText

Now lets look at a boxplot comparing mpg between domestic and foreign.

```
.. gr box mpg, over(foreign) name(fig1)
.. figure:: fig1.pdf
   :scale: 33

   Boxplot comparing domestic and foreign.
```

Graph: PDF via LaTeX

Now lets look at a boxplot comparing mpg between domestic and foreign.

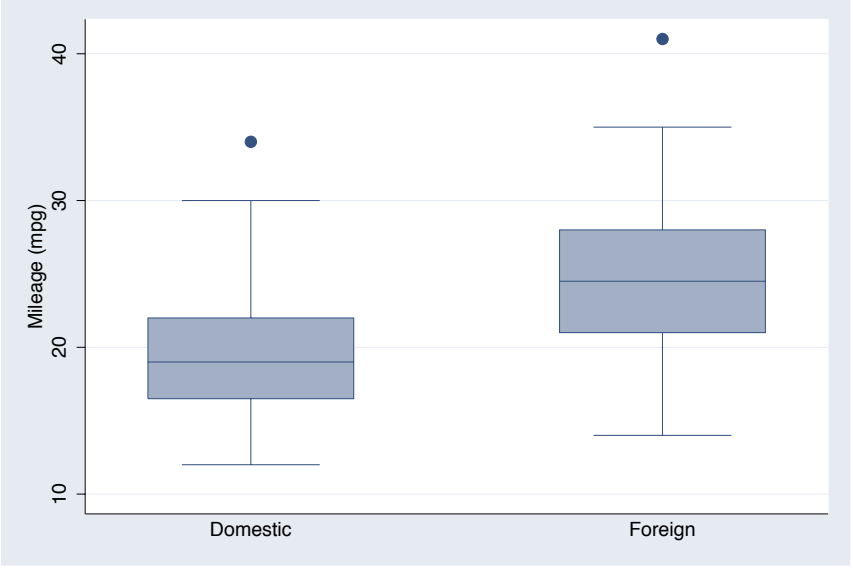


Figure 1: Boxplot comparing domestic and foreign.

Substitution: do-file

```
* Using a t-test to compare mpg between foreign and domestic
* cars yields a p-value of |s2d_ttp|.

s2d ttp=(string(r(p),"%05.4f")): ttest mpg, by(foreign)
```

Substitution: reStructuredText

Using a t-test to compare mpg between foreign and domestic cars yields a p-value of |s2d_ttp|.

```
.. s2d ttp=(string(r(p),"%05.4f")): ttest mpg, by(foreign)
.. |s2d_ttp| replace:: 0.0005
```

Substitution: PDF via LaTeX

Using a t-test to compare mpg between foreign and domestic cars yields a p-value of 0.0005.

Table: do-file

```
* Finally, we'll try regressing ``mpg'' on ``weight'', ``weightsq'',
* and ``foreign''.

* Regression of mpg on several covariates.
s2d, t: reg mpg weight weightsq foreign
```

Table: reStructuredText

```
Finally, we'll try regressing ``mpg`` on ``weight``, ``weightsq``,
and ``foreign``.

.. s2d, t: reg mpg weight weightsq foreign
.. stata-table:: Regression of mpg on several covariates.
```

mpg	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
weight	-.0165729	.0039692	-4.18	0.000	-.0244892	-.0086567
weightsq	1.59e-06	6.25e-07	2.55	0.013	3.45e-07	2.84e-06
foreign	-2.2035	1.059246	-2.08	0.041	-4.3161	-.0909002
_cons	56.53884	6.197383	9.12	0.000	44.17855	68.89913

Table: PDF via LaTeX

Finally, we'll try regressing `mpg` on `weight`, `weightsq`, and `foreign`.

Table 1: Regression of mpg on several covariates.

mpg	**Coef.**	**Std. Err.**	**t**	**P>|t|**	**[95% Conf.**	**Interval]**
weight	-.0165729	.0039692	-4.18	0.000	-.0244892	-.0086567
weightsq	1.59e-06	6.25e-07	2.55	0.013	3.45e-07	2.84e-06
foreign	-2.2035	1.059246	-2.08	0.041	-4.3161	-.0909002
_cons	56.53884	6.197383	9.12	0.000	44.17855	68.89913