Stata Markdown and Reproducible Research

Rob Hicks

Note: course syllabus is here:

https://rlhick.people.wm.edu/stories/syllabus_econ407.html

Introduction

From Wikipedia, reproducible research is defined as:

> The term reproducible research refers to the idea that the ultimate product of academic research is the paper **along with** the full computational environment used to produce the results in the paper such as the code, data, etc. that can be used to reproduce the results and create new work based on the research.

The reproducible research movement (especially for the statistical sciences) takes this a step further by advocating for dynamic documents. The idea is that a researcher should provide a file (the dynamic document) that can execute the statistical analysis, generate figures, and contains accompanying text narrative. This file can be executed to produce the **academic paper**. The researcher shares this file with other researchers rather than the only the paper. It is my view that within 20 years nearly every scientific journal in applied statistics will require this approach.

This document shows how to use MarkStat and markdown syntax for reproducible research and dynamic documents in stata. The idea behind MarkStat is that you share your research by sharing your do file. This do file performs the full suite of statistical analysis and can produce the pdf (with extra configuration), MS Word, or html documents describing your analysis. You will use this workflow for producing pdf or word documents for class assignments.

For every problem set, you will turn in

- The stata `stmd` (similar to a do file) file containing all commands and written text that produces your problem set responses.
- A hardcopy of the pdf or word version produced after running your do file [the hardcopy]

The only exception to this rule is for questions involving proofs or other equation heavy assigments where handwritten responses can be attached to the hardcopy problem set response.

Installation Instructions

In `Stata`, issue these commands:

1. `ssc install markstat`
2. `ssc install whereis`
3. Install pandoc from `http://pandoc.org/installing`
4. Tell markstat where to find pandoc. Probably the command you need to run in stata is:

- Windows: `whereis pandoc "C:\Users\username\AppData\Local\Pandoc\pandoc.exe"`
- Mac: `whereis pandoc /usr/local/bin/pandoc`
- Linux/Unix: `whereis pandoc /usr/bin/pandoc`

Windows users should substitute your username for "username" in the `whereis` commands above

Some Features of MarkStat

Markdoc allows for most features of Markdown, which is a liteweight and readable **text-based** language that allows files to be easily converted to nice looking pdf, html, or even word documents. Some features you will likely want to use:

- Equations and Math Notation using latex math
- Headers
- Emphasizing text (bold and italics)
- Numeric and bulletted lists
- Turning stata output on and off
- Pagebreaks can be inserted using `\newpage` on a separate line

A simple example analysis using Markdoc

Below we'll be modeling the following regression equation for cars back in the day:

$$price_i = \beta_0 + \beta_1 mpg_i + \beta_2 foreign_i + \epsilon_i$$

Load Data and Summarize

```
. cd ~/Dropbox/Current/Teaching/courses/ECON407/do_files/reproducible_research/
> markstat
/home/robhicks/Dropbox/Current/Teaching/courses/ECON407/do_files/reproducible_r
> esearch/markstat

. webuse auto
(1978 Automobile Data)

. reg price mpg
```

Source	SS	df	MS	Number of obs	=	74
				F(1, 72)	=	20.26
Model	139449474	1	139449474	Prob > F	=	0.0000
Residual	495615923	72	6883554.48	R-squared	=	0.2196
				Adj R-squared	=	0.2087
Total	635065396	73	8699525.97	Root MSE	=	2623.7

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mpg	-238.8943	53.07669	-4.50	0.000	-344.7008	-133.0879
_cons	11253.06	1170.813	9.61	0.000	8919.088	13587.03

```
.     sum
```

Variable	Obs	Mean	Std. Dev.	Min	Max
make	0				
price	74	6165.257	2949.496	3291	15906
mpg	74	21.2973	5.785503	12	41
rep78	69	3.405797	.9899323	1	5
headroom	74	2.993243	.8459948	1.5	5
trunk	74	13.75676	4.277404	5	23
weight	74	3019.459	777.1936	1760	4840
length	74	187.9324	22.26634	142	233
turn	74	39.64865	4.399354	31	51
displacement	74	197.2973	91.83722	79	425
gear_ratio	74	3.014865	.4562871	2.19	3.89
foreign	74	.2972973	.4601885	0	1

```
. hist price
(bin=8, start=3291, width=1576.875)

. graph export price.png, replace
(file price.png written in PNG format)
```

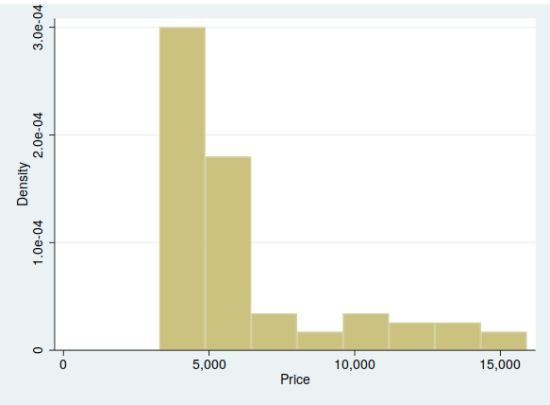

Figure 1: Histogram of Price

Regression Model

Here are the regression results:

```
. reg price mpg foreign
```

Source	SS	df	MS	Number of obs	=	74
				F(2, 71)	=	14.07
Model	180261702	2	90130850.8	Prob > F	=	0.0000
Residual	454803695	71	6405685.84	R-squared	=	0.2838
				Adj R-squared	=	0.2637
Total	635065396	73	8699525.97	Root MSE	=	2530.9

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mpg	-294.1955	55.69172	-5.28	0.000	-405.2417	-183.1494
foreign	1767.292	700.158	2.52	0.014	371.2169	3163.368
_cons	11905.42	1158.634	10.28	0.000	9595.164	14215.67

Discussion

Looks like back in the day, foreign cars sell for more!

Markdoc and Mata

Mata is the matrix algebra environment in stata. We can embed markdown (including equations) inside mata too:

Define $\mathbf{A}_{2\times 2}$ as

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

```
.     mata
------------------------------------------------- mata (type end to exit) -----

: A = (1,2\3,4)
: A
       1   2
    +-------+
  1 |   1   2 |
  2 |   3   4 |
    +-------+

: end
--------------------------------------------------------------------------------
```

Compiling your document

You will be creating a file with an `stmd` extension that contains your code and writeup. You can create and edit this file in any text editor including the stata do file editor.

Suppose your problem set document called `script.stmd` contained this text:

```
% Problem Set 1
% Johnny Appleseed
% Sept 1, 2018
Let us read the fuel efficiency data that ships with Stata

    sysuse auto, clear

To study how fuel efficiency depends on weight it is useful to transform
the dependent variable from "miles per gallon" to "gallons per 100
miles"

    gen gphm = 100/mpg

We then obtain a fairly linear relationship

    twoway scatter gphm weight || lfit gphm weight ///
    ytitle(Gallons per 100 Miles) legend(off)
    graph export auto.png, width(500) replace

![Fuel Efficiency by Weight](auto.png)

The regression equation estimated by OLS is
$$
    gphm = \beta_0 + \beta_1 weight + \epsilon
$$
Estimating in stata, yields:
```

```
    regress gphm weight

Thus, a car that weighs 1,000 pounds more than another requires on
average an extra 1.4 gallons to travel 100 miles.
```

You can then generate a word, pdf, or html document containing all code and results with these commands in stata (assuming your current working directory contains `script.stmd`):

- `markstat using script, mathjax`: produces an html file
- `markstat using script, mathjax docx`: produces a word document
- `markstat using script, mathjax pdf`: produces a pdf document (requires working latex environment)

Problem set responses produced by `markstat` in small fonts will be immediately returned to the student and considered not turned in until font sizes are fixed. Shoot for 11pt fonts.

Document Details

This document is written entirely in `stata` using `markstat`. To see the source code, see http://rlhick.people.wm.edu/bin/reproducible_research.stmd (clickable).