Data Wrangling & Munging: Organising, Managing and Enabling Data for Analysis©

Prof. John MacInnes & Dr Diarmuid McDonnell

University of Edinburgh

July 2018

© Diarmuid McDonnell

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

AQMEN

THE UNIVERSITY of EDINBURGH

Better Knowledge
Better Society
Supporting researchers to use administrative data for economic and social research
Enabling studies that benefit the public
Helping inform and improve policy-making
W. adrn.ac.uk
E. adrc-s@ed.ac.uk
AQMeN
www.aqmen.ac.uk

Administrative Data Research Centre Scotland
An ESRC Data Investment
Better Knowledge
Better Society
Supporting researchers to use administrative data for economic and social research
Enabling studies that benefit the public
Helping inform and improve policy-making
W. adrn.ac.uk
E. adrc-s@ed.ac.uk
AQMeN
Applied Quantitative Methods Network

Part I

The Workflow

A Thought Experiment

Be honest….

1. Have you ever lost a file?

2. Have you ever wondered if you have deleted a file?

3. Have you and a colleague ever been working on different versions of a file?

A Thought Experiment

Be honest….

Have you ever struggled to identify which data file is the correct one?

custinfo_2014.xlsx

custinfo_recent_2014.xlsx

cinfo_report2014.xlsx

DEATH VALLEY MARATHON
383
MARATHON
791
WHAT IS QUANTITATIVE LONGITUDINAL DATA ANALYSIS?

The Workflow

- Planning, organising and documenting work

- This includes…

Cleaning data

Analysing data

Presenting results

Backing up and archiving material

The Workflow

Workflow should be planned and carefully orchestrated

Workflow MUST not be *adhoc*

(e.g. piece-meal, developed as a reaction to mistakes etc.)

The Workflow

Better supporting YOU and what YOU DO

Not changing you into something YOU ARE NOT

Shopping without a list?
Cooking with a list?

APPROVED B-17F and G CHECKLIST

REVISED 3-1-44

PILOT'S DUTIES IN RED

COPILOT'S DUTIES IN BLACK

BEFORE STARTING

1. Pilot's Preflight—COMPLETE
2. Form 1A—CHECKED
3. Controls and Seats—CHECKED
4. Fuel Transfer Valves & Switch—OFF
5. Intercoolers—Cold
6. Gyros—UNCAGED
7. Fuel Shut-off Switches—OPEN
8. Gear Switch—NEUTRAL
9. Cowl Flaps—Open Right—OPEN LEFT—Locked
10. Turbos—OFF
11. Idle cut-off—CHECKED
12. Throttles—CLOSED
13. High RPM—CHECKED
14. Autopilot—OFF
15. De-icers and Anti-icers, Wing and Prop—OFF
16. Cabin Heat—OFF
17. Generators—OFF

STARTING ENGINES

1. Fire Guard and Call Clear—LEFT Right
2. Master Switch—ON
3. Battery switches and inverters—ON & CHECKED
4. Parking Brakes—Hydraulic Check—On-CHECKED
5. Booster Pumps—Pressure—ON & CHECKED
6. Carburetor Filters—Open
7. Fuel Quantity—Gallons per tank
8. Start Engines: both magnetos on after one revolution
9. Flight Indicator & Vacuum Pressures CHECKED
10. Radio—On
11. Check Instruments—CHECKED
12. Crew Report
13. Radio Call & Altimeter—SET

ENGINE RUN-UP

1. Brakes—Locked
2. Trim Tabs—SET
3. Exercise Turbos and Props
4. Check Generators—CHECKED & OFF
5. Run up Engines

BEFORE TAKEOFF

1. Tailwheel—Locked
2. Gyro—Set
3. Generators—ON

AFTER TAKEOFF

1. Wheel—PILOT'S SIGNAL
2. Power Reduction
3. Cowl Flaps
4. Wheel Check—OK right—OK LEFT

BEFORE LANDING

1. Radio Call, Altimeter—SET
2. Crew Positions—OK
3. Autopilot—OFF
4. Booster Pumps—On
5. Mixture Controls—AUTO-RICH
6. Intercooler—Set
7. Carburetor Filters—Open
8. Wing De-icers—Off
9. Landing Gear
   a. Visual—Down Right—DOWN LEFT Tailwheel Down, Antenna in, Ball Turret Checked
   b. Light—OK
   c. Switch Off—Neutral
10. Hydraulic Pressure—OK Valve closed
11. RPM 2100—Set
12. Turbos—Set
13. Flaps ⅓—⅓ Down

FINAL APPROACH

14. Flaps—PILOT'S SIGNAL
15. RPM 2200—PILOT'S SIGNAL

In the late 1930s, military aviators in the American Army and Navy began using aviation checklists. Checklist became part of a new paradigm for how to fly, which consisted of

- Elaborate standardized procedures for many activities
- Checklists to ensure all critical steps had been done
- Quantitative tables and formulas that specified the best settings, under different conditions, for speed, engine RPM, gasoline/air mixture, engine cooling, and many other parameters.

This new paradigm (Standard Procedure Flying) had a major influence on reducing aviation accidents and increasing military effectiveness during World War II, particularly because of the rapidly increasing complexity of military aircraft, and the huge number of new pilots.

Despite the benefits of Standard Procedure Flying for both safety and efficiency, by the end of WWII only a few air forces had fully embraced it

Roger Bohn http://www.vs29.org/Links/NATOPS/SOP-bohn-2013-1.pdf

The Workflow

The Four Pillars of Wisdom

A Planned Workflow Has Benefits

Four Pillars of Wisdom

- Accuracy
  - minimising information loss and errors in analyses and output
- Efficiency
  - automation, maximising features in software
- Transparency
  - showing what you did, why, when, how
- Reproducibility
  - same results every time whoever or wherever
  - editing, rewriting reports or re-submission of papers

J. Scott Long has posted a really good pdf version of a talk on the workflow
http://www.ihrp.uic.edu/files/Workflow%20Slides%20JSLong%20110410.pdf

The best habit that you can get into

is to get into good habits!

The Workflow

Drukker's dictum: Never type anything that you can obtain from a saved result

Gayle's dictum:

You can't be too fit or have too many publications

However…

- 500+ scientific publications in peer reviewed journals (15,000+ citations and an H-index of 66)

- Has run more than 70 marathon and ultra-marathon races, including seven 90km Comrades Marathons and fifteen 56km Two Oceans Marathons

http://www.essm.uct.ac.za/ESSM/Tim_Noakes

Long’s Law

It is always easier to document today than it is tomorrow!

Corollary 1:

Nobody likes to write documentation

Corollary 2:

Nobody every regrets having written documentation

Long's Law

Has anyone in the history of data analysis ever said

"these files are too well documented"

scarc.library.oregonstate.edu/coll/pauling/rnb/index.html

As with many scientists, Linus Pauling utilized bound notebooks to keep track of the details of his research as it unfolded. A testament to the remarkable length and diversity of Dr. Pauling's career, the Pauling Papers holdings include forty-six research notebooks spanning the years of 1922 to 1994 and covering any number of the scientific fields in which Dr. Pauling involved himself. In this regard, the notebooks contain many of Pauling's laboratory calculations and experimental data, as well as scientific conclusions, ideas for further research and numerous autobiographical musings.

Research Notebook 01
1922
Research Notebook 02
1922-1923, 1932, 1934, 1936, 1973, 1985
Research Notebook 03
1923-1925
Research Notebook 04
1923-1924, 1928-1930
Research Notebook 05

Research Notebook 13
1935-1936, 1938-1939
Research Notebook 14
1936-1939, 1949, 1952
Research Notebook 15
1935, 1937, 1968
Research Notebook 16
1935-1956
Research Notebook 17
1939-1941, 1971, 1988

Research Notebook 24
1953, 1956, 1962, 1963, 1967, 1968, 1969, 1970, 1973
Research Notebook 25
1958, 1964-1966
Research Notebook 26
1955, 1964-1969, 1974-1976, 1980-1982, 1987, 1990-1991
Research Notebook 27
1952-1954, 1960-1961, 1964, 1971-

Research Notebook 35b
1938-1939, 1946, 1955, 1968, 1986-1988
Research Notebook 36
1980-1981, 1986-1987
Research Notebook 37
1971, 1983
Research Notebook 38
1980-1981, 1983, 1985, 1989
Research Notebook 39

Book 24 - Page 151 - Linu
scarc.library.oregonstate.edu/coll/pauling/rnb/24/24-151.html

21 June 1973
Portola Valley, Cal.
Linus Pauling.

Golden Wedding Anniversary

Three days ago Ava Helen and I celebrated our 50th wedding anniversary. We had been married in Salem, Oregon, on 17 June 1923.

Our celebration was at the ranch. Our four children were there, also Linda's husband (Barclay) and their four boys; also Art and Laurelee Robinson; also Art Cherkin; also Ewan Cameron from Scotland, his wife, and two children; also L. Gorgo Miller, Ava Helen's sister; also my two sisters, Pauline Pauling-Ney and Lucille Jenkins

The Workflow

- Improving the workflow with a modest amount of effort

- The less experience you have the better
  - start from the very beginning

The Workflow

ALL SERIOUS WORK MUST BE REPRODUCIBLE!

There MUST be an audit trail

The Workflow

Why is it all so difficult?

Social science data tends to come in messy formats

Administrative data often is even more complex in nature than social survey data

The Workflow

Why is it all so difficult?

Minor decisions have major consequences…

Which cases?
Which variables?
How to code (e.g. education)?
How to recode?
Where do I truncate?

Can I trace these decisions in my audit trail?

The Workflow

Why is it all so difficult?

Poor discipline and insufficient documentation

The Workflow

File Naming Protocols

File Name =

name_date_depositor's initials_version_type

The Workflow

File Naming Protocols

File Name = name_date_depositor's initials_version_type

Therefore **bhpsaindresp_20140506_vg_v1.dta**

Would be a

a.. The British Household Panel Survey File “aindresp”
b.. Deposited on 6th May 2014
c.. Deposited by vg (Vernon Gayle)
d.. Version v1
e.. File type (e.g. a Stata .dta file)

	A	C	D	E	F	G	H	I
1	File Register							
2								
3								
4		File Name (name_subname_date[year/month/day]_depositor's initials_version_type)	File Type					Brief Description of the file and it purpose
5	Directory Name	(e.g. bhps_aindresp_140129_vg_v1.dta)	(e.g. Stata data file)	Name of Author	Initials of Author	Date of Creation	Date of last revision	(e.g. Stata .do file MSc dissertation; Draft Chapter 1 PhD)
6								
7								
8								
9								
10								
11								
12								
13								
14								
15								
16								
17								

Other seemingly small issues such as ‘Directory Structures’ and ‘Variable Naming Conventions’ are similarly worth thinking about!

Hidden Realities

- Collaboration
- Team working
- Version Control

Google docs

OneDrive

Dropbox

EVERNOTE

GitLab

Part II

Stata

Excel

Spreadsheet can not easily be scripted –

So there will be no clear audit trail of activity

If you use a ‘gui’ then one day you will inevitably end up in a sticky mess

We advise against undertaking data analyses using software packages in interactive modes, for example through a graphical user interface (gui) such as drop-down menus.

Which tool?

Which software (or data analysis package) should I learn?

"This is like asking which five countries should I know about, given that I don't have an interest in either geology or politics – the answers is it depends"

Larry Wall (the creator of the programming language Pearl)

SPSS

IBM SPSS Statistics 22
File Edit View Data
IBM SPSS Statistics
IBM
New Files:
New Dataset
New Database Query...
Recent Files:
...ample_20170921_vg_v1.sav
...\stata8lycs9sw1.dta
..._spss_20160221_vg_v1.sav
..._spss_20160221_vg_v1.sav
Open another file...
What's New:
Cognos BI users: Further enhance your analysis using SPSS Statistics
Now you can import Cognos BI data into SPSS Statistics to gain further insights into your analysis.
Modules and Programmability:
Learn more about the modules and programmbility extensions
Show: Installed
IBM SPSS Statistics
IBM SPSS Regression
IBM SPSS Advanced Statistics
Tutorials:
Learn how to use SPSS Statistics to get the results you need
Introduction
Reading Data
Don't show this dialog in the future
OK
Cancel
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode:ON

SPSS

- Statistical Package for the Social Sciences
- Was widely taught in university departments (e.g. psychology, sociology, health etc.)
- Commercially available
- www.ibm.com/DataStatistics/SPSS

SAS

- Now only widely used in government departments
- Re-launched as a data science package
- Commercially available
- Fiddly to get started with
- www.sas.com

R

- Growing in popularity (e.g. data science, statistics, science etc.)

- Popular with statisticians

- Free (open source)

- Difficult to learn
- Development and support is not commercial
- Help resources are under-developed

- https://cran.r-project.org

Other Approaches

- Python (www.python.org)
  - Open source
  - Fewer libraries than R
  - Tricky to learn and almost no help resources

- Julia (https://julialang.org/)
  - Open source
  - Fewer libraries than R
  - Tricky to learn and no help resources

Stata

- Popular in economics globally

- Popular with US social science (e.g. sociology)

- Commercial software

- Easy to learn
- Excellent development
- Outstanding help files and teaching resources

- www.stata.com

Stata

- Does all the routine stuff (SPSS, SAS & R)
- Fits many more models than standard software (esp. for selection, longitudinal)
- Specialist complex survey analysis functions (Svy) (useful for UKHLS, MCS etc)
- Human readable format
- Very strong documentation
- Worldwide user community (lists etc)
- New features almost daily

‘*…[Stata] has gotten better and better over time, so that by now it can happily be used as a general-purpose package. Stata is powerful and fast which makes it viable to carry out analysis on a PC…*

Overall, Stata is a very good choice for our kind of work’

(Treiman 2009 p.66)

Data Analysis Using Stata
Third Edition
ULRICH KOHLER
FRAUKE KREUTER
The Workflow of Data Analysis Using Stata
J. SCOTT LONG
Mehmet Mehmetoglu
Tor Georg Jakobsen
APPLIED STATISTICS USING STATA
A Guide for the Social Sciences
David Pevalin
Karen Robson
the stata survival manual

http://www.timberlake.co.uk/software/stata/business/single/prices.html

Please select:

- Qty: 1 - Business single user / volume user license - £955
- Qty: 1 - Business single user / volume user (annual license) - £475
- Qty: 2 - Business single user / volume user license - £1,595
- Qty: 3 - Business single user / volume user license - £2,010
- Qty: 4 - Business single user / volume user license - £2,420
- Qty: 5 - Business single user / volume user license - £2,830
- Qty: 6 - Business single user / volume user license - £3,170
- Qty: 7 - Business single user / volume user license - £3,505
- Qty: 8 - Business single user / volume user license - £3,840
- Qty: 9 - Business single user / volume user license - £4,175
- Qty: 10 - Business single user / volume user license - £4,510
- Qty: 11 - Business single user / volume user license - £4,775
- Qty: 12 - Business single user / volume user license - £5,040

Using Stata

Stata/SE 15.0

File Edit Data Graphics Statistics User Window Help

```
  ___  ____  ____  ____  ____ (R)
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   15.0   Copyright 1985-2017 StataCorp LLC
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
     Special Edition                  College Station, Texas 77845 USA
                                      800-STATA-PC        http://www.stata.com
                                      979-696-4600        stata@stata.com
                                      979-696-4601 (fax)

Unlimited-user Stata network license expires 14 Sep 2018:
       Serial number:  401509006065
         Licensed to:  Vernon Gayle
                       University of Edinburgh

Notes:
      1.  Unicode is supported; see help unicode_advice.
      2.  Maximum number of variables is set to 5000; see help set_maxvar.
      3.  New update available; type -update all-

.
```

Variables

Filter variables here

Name | Label

There are no items to show.

Command

Data Editor (familiar spread sheet)

Data Editor (Browse) - [auto]

File Edit View Data Tools

make[1] AMC Concord

	make	price	mpg	rep78	headroom	trunk	weight	length	turn	displacement	gear_ratio	foreign
1	AMC Concord	4,099	22	3	2.5	11	2,930	186	40	121	3.58	Domestic
2	AMC Pacer	4,749	17	3	3.0	11	3,350	173	40	258	2.53	Domestic
3	AMC Spirit	3,799	22	.	3.0	12	2,640	168	35	121	3.08	Domestic
4	Buick Century	4,816	20	3	4.5	16	3,250	196	40	196	2.93	Domestic
5	Buick Electra	7,827	15	4	4.0	20	4,080	222	43	350	2.41	Domestic
6	Buick LeSabre	5,788	18	3	4.0	21	3,670	218	43	231	2.73	Domestic
7	Buick Opel	4,453	26	.	3.0	10	2,230	170	34	304	2.87	Domestic
8	Buick Regal	5,189	20	3	2.0	16	3,280	200	42	196	2.93	Domestic
9	Buick Riviera	10,372	16	3	3.5	17	3,880	207	43	231	2.93	Domestic
10	Buick Skylark	4,082	19	3	3.5	13	3,400	200	42	231	3.08	Domestic
11	Cad. Deville	11,385	14	3	4.0	20	4,330	221	44	425	2.28	Domestic
12	Cad. Eldorado	14,500	14	2	3.5	16	3,900	204	43	350	2.19	Domestic
13	Cad. Seville	15,906	21	3	3.0	13	4,290	204	45	350	2.24	Domestic
14	Chev. Chevette	3,299	29	3	2.5	9	2,110	163	34	231	2.93	Domestic
15	Chev. Impala	5,705	16	4	4.0	20	3,690	212	43	250	2.56	Domestic
16	Chev. Malibu	4,504	22	3	3.5	17	3,180	193	31	200	2.73	Domestic
17	Chev. Monte Carlo	5,104	22	2	2.0	16	3,220	200	41	200	2.73	Domestic
18	Chev. Monza	3,667	24	2	2.0	7	2,750	179	40	151	2.73	Domestic
19	Chev. Nova	3,955	19	3	3.5	13	3,430	197	43	250	2.56	Domestic
20	Dodge Colt	3,984	30	5	2.0	8	2,120	163	35	98	3.54	Domestic
21	Dodge Diplomat	4,010	18	2	4.0	17	3,600	206	46	318	2.47	Domestic
22	Dodge Magnum	5,886	16	2	4.0	17	3,600	206	46	318	2.47	Domestic
23	Dodge St. Regis	6,342	17	2	4.5	21	3,740	220	46	225	2.94	Domestic
24	Ford Fiesta	4,389	28	4	1.5	9	1,800	147	33	98	3.15	Domestic

Syntax File .do

Do-file Editor - example_20180611_vg_v1

File Edit View Project Tools

example_20180611_vg_v1

```
** This file has Stata commands**

mean price

** This file contains notes **

/**

*******************************************************************************

Latest Update:

11th June 2018 Poundworld enters administration after rescue talks fail.


Previous Updates:

21st January 2018 The Kinks bassist Jim Rodford has dies at the age of 76.

20th January 2018 Alan Carr gets married to long-term boyfriend in LA.

*******************************************************************************


**/
```

Stata Main Window (output)

Do-file Editor - example_20180611_vg_v1

File Edit View Project Tools

example_20180611_vg_v1

```
** This file has Stata commands**

mean price

** This file contains notes **

/**

*****************************************************************************

Latest Update:

11th June 2018 Poundworld enters administration after rescue talks fail.


Previous Updates:

21st January 2018 The Kinks bassist Jim Rodford has dies at the age of 76.

20th January 2018 Alan Carr gets married to long-term boyfriend in LA.

*****************************************************************************


**/
```

Data Editor (Browse) - [auto]

File Edit View Data Tools

make[1] AMC Concord

	make	price	mpg	rep78	headroom	trunk	weight	length	turn	displacement	gear_ratio	foreign
1	AMC Concord	4,099	22	3	2.5	11	2,930	186	40	121	3.58	Domestic
2	AMC Pacer	4,749	17	3	3.0	11	3,350	173	40	258	2.53	Domestic
3	AMC Spirit	3,799	22	.	3.0	12	2,640	168	35	121	3.08	Domestic
4	Buick Century	4,816	20	3	4.5	16	3,250	196	40	196	2.93	Domestic
5	Buick Electra	7,827	15	4	4.0	20	4,080	222	43	350	2.41	Domestic
6	Buick LeSabre	5,788	18	3	4.0	21	3,670	218	43	231	2.73	Domestic
7	Buick Opel	4,453	26	.	3.0	10	2,230	170	34	304	2.87	Domestic
8	Buick Regal	5,189	20	3	2.0	16	3,280	200	42	196	2.93	Domestic
9	Buick Riviera	10,372	16	3	3.5	17	3,880	207	43	231	2.93	Domestic
10	Buick Skylark	4,082	19	3	3.5	13	3,400	200	42	231	3.08	Domestic
11	Cad. Deville	11,385	14	3	4.0	20	4,330	221	44	425	2.28	Domestic
12	Cad. Eldorado	14,500	14	2	3.5	16	3,900	204	43	350	2.19	Domestic
13	Cad. Seville	15,906	21	3	3.0	13	4,290	204	45	350	2.24	Domestic
14	Chev. Chevette	3,299	29	3	2.5	9	2,110	163	34	231	2.93	Domestic
15	Chev. Impala	5,705	16	4	4.0	20	3,690	212	43	250	2.56	Domestic
16	Chev. Malibu	4,504	22	3	3.5	17	3,180	193	31	200	2.73	Domestic
17	Chev. Monte Carlo	5,104	22	2	2.0	16	3,220	200	41	200	2.73	Domestic
18	Chev. Monza	3,667	24	2	2.0	7	2,750	179	40	151	2.73	Domestic
19	Chev. Nova	3,955	19	3	[illegible]	13	3,430	197	43	250	2.56	Domestic
20	Dodge Colt	3,984	30	5	[illegible]	8	2,120	163	35	98	3.54	Domestic
21	Dodge Diplomat	4,010	18	2	[illegible]	17	3,600	206	46	318	2.47	Domestic
22	Dodge Magnum	5,886	16	2	[illegible]	17	3,600	206	46	318	2.47	Domestic
23	Dodge St. Regis	6,342	17	2	[illegible]	21	3,740	220	46	225	2.94	Domestic
24	Ford Fiesta	4,389	28	4	[illegible]	9	1,800	147	33	98	3.15	Domestic

Stata/SE 15.0 - C:\Program Files (x86)\Stata15\ado\base\a\auto.dta

File Edit Data Graphics Statistics User Window Help

```
. mean price

Mean estimation                   Number of obs    =         74

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
       price |   6165.257   342.8719      5481.914       6848.6
--------------------------------------------------------------

. * output appears here in the main Stata (black) window
```

Log File (plain text)

```
. mean price

Mean estimation                   Number of obs    =         74

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
       price |   6165.257   342.8719      5481.914      6848.6
--------------------------------------------------------------

.
. ** Output appears is copied to the log file
.
```

Do-file Editor - example_20180611_vg_v1

File Edit View Project Tools

example_20180611_vg_v1

```
** This file has Stata commands**

mean price

** This file contains notes **

/**

*****************************************************************************

Latest Update:

11th June 2018 Poundworld enters administration after rescue talks fail.


Previous Updates:

21st January 2018 The Kinks bassist Jim Rodford has dies at the age of 76.

20th January 2018 Alan Carr gets married to long-term boyfriend in LA.

*****************************************************************************


**/
```

Data Editor (Browse) - [auto]

File Edit View Data Tools

make[1] AMC Concord

	make	price	mpg	rep78	headroom	trunk	weight	length	turn	displacement	gear_ratio	foreign
1	AMC Concord	4,099	22	3	2.5	11	2,930	186	40	121	3.58	Domestic
2	AMC Pacer	4,749	17	3	3.0	11	3,350	173	40	258	2.53	Domestic
3	AMC Spirit	3,799	22	.	3.0	12	2,640	168	35	121	3.08	Domestic
4	Buick Century	4,816	20	3	4.5	16	3,250	196	40	196	2.93	Domestic
5	Buick Electra	7,827	15	4	4.0	20	4,080	222	43	350	2.41	Domestic
6	Buick LeSabre	5,788	18	3	4.0	21	3,670	218	43	231	2.73	Domestic
7	Buick Opel	4,453	26	.	3.0	10	2,230	170	34	304	2.87	Domestic
8	Buick Regal	5,189	20	3	2.0	16	3,280	200	42	196	2.93	Domestic
9	Buick Riviera	10,372	16	3	3.5	17	3,880	207	43	231	2.93	Domestic
10	Buick Skylark	4,082	19	3	3.5	13	3,400	200	42	231	3.08	Domestic
11	Cad. Deville	11,385	14	3	4.0	20	4,330	221	44	425	2.28	Domestic
12	Cad. Eldorado	14,500	14	2	3.5	16	3,900	204	43	350	2.19	Domestic
13	Cad. Seville	15,906	21	3	3.0	13	4,290	204	45	350	2.24	Domestic
14	Chev. Chevette	3,299	29	3	2.5	9	2,110	163	34	231	2.93	Domestic
15	Chev. Impala	5,705	16	4	4.0	20	3,690	212	43	250	2.56	Domestic
16	Chev. Malibu	4,504	22	3	3.5	17	3,180	193	31	200	2.73	Domestic
17	Chev. Monte Carlo	5,104	22	2	2.0	16	3,220	200	41	200	2.73	Domestic
18	Chev. Monza	3,667	24	2	2.0	7	2,750	179	40	151	2.73	Domestic
19	Chev. Nova	3,955	19	3	3.5	13	3,430	197	43	250	2.56	Domestic
20	Dodge Colt	3,984	30	5	2.0	8	2,120	163	35	98	3.54	Domestic
21	Dodge Diplomat	4,010	18	2	4.0	17	3,600	206	46	318	2.47	Domestic
22	Dodge Magnum	5,886	16		4.0	17	3,600	206	46	318	2.47	Domestic
23	Dodge St. Regis	6,342	17		4.5	21	3,740	220	46	225	2.94	Domestic
24	Ford Fiesta	4,389	28		1.5	9	1,800	147	33	98	3.15	Domestic

Stata/SE 15.0 - C:\Program Files (x86)\Stata15\ado\base\a\auto.dta

File Edit Data Graphics Statistics User Window Help

```
. mean price

Mean estimation                   Number of obs    =         74

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
       price |   6165.257   342.8719      5481.914      6848.
--------------------------------------------------------------

.
. ** Output appears is copied to the log file
.
```

```
. mean price

Mean estimation                   Number of obs    =         74

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
       price |   6165.257   342.8719      5481.914      6848.6
--------------------------------------------------------------

. * output appears here in the main Stata (black) window
```

Opening a .do file…

Stata/SE 15.0

File Edit Data Graphics Statistics User Window Help

New Do-file Editor

```
  ___  ____  ____  ____  ____ (R)
 /__    /   ____/   /   ____/
___/   /___/   /___/   /___/   15.0   Copyright 1985-2017 StataCorp LLC
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
     Special Edition                  College Station, Texas 77845 USA
                                      800-STATA-PC        http://www.stata.com
                                      979-696-4600        stata@stata.com
                                      979-696-4601 (fax)

Unlimited-user Stata network license expires 14 Sep 2018:
       Serial number:  401509006065
         Licensed to:  Vernon Gayle
                       University of Edinburgh

Notes:
```

Do-file Editor - Untitled2.do
File
Edit
View
Project
Tools
New
Open
Close
Close project
Save Ctrl+S
Save as... Ctrl+Shift+S
Save all
Save project
Print Ctrl+P
Print preview
Print setup...
Recent files
Recent projects
Exit
Open Ctrl+O
Project...

Stata/SE 15.0

File Edit Data Graphics Statistics User Window Help

```
                                                  (R)
  ___  ____  ____  ____  ____
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   15.0   Copyri
  Statistics/Data Analysis            StataC
                                      4905 L
     Special Edition                  Colleg
                                      800-ST
                                      979-69
                                      979-69

Unlimited-user Stata network license expires
       Serial number:  401509006065
         Licensed to:  Vernon Gayle
                       University of Edinbur

Notes:
      1.  Unicode is supported; see help uni
      2.  Maximum number of variables is set
```

Do-file Editor - workshop_wp2_20180611_vg_v1

File Edit View Project Tools

workshop_wp2_20180611_vg_v1

```
STOP

/**

AQMEN Applied Quantitative Methods Network /


An Introduction to Statistical Concepts for Data Analysis

Stata Workshop (July 2018)

A two day hands-on workshop led by Professor Vernon Gayle,
University of Edinburgh.

        ********************************************
        * IT IS IMPORTANT THAT YOU READ THIS HANDOUT *
        * AND FOLLOW THE STATA.DO FILE LINE BY LINE! *
        ********************************************

Topics:

The course introduces participants to fundamental concepts in
rudimentary data analysis techniques and how to interpret resu
These skills are critical for the successful analysis of data.
```

Do-file Editor - workshop_wp2_20180611_vg_v1

File Edit View Project Tools

workshop_wp2_20180611_vg_v1

Execute selection (do)

```
187
188
189 ********************************************************************************
190 *
191 *                    Setting Up Stata
192 *
193 ********************************************************************************
194
195 * This section is about organising preliminary settings in Stata *
196
197 * clear the computer memory *
198
199 clear
200
201 /** More causes Stata to display --more-- and pause until any key is pressed.
202     It is usually more convenient to have this function switched off **/
203
204 set more off
205
206 * keep a log file containing your output *
207
208 * close any log files that might already be running *
209
210 capture log close
```

Do-file Editor - workshop_wp2_20180611_vg_v1

File Edit View Project Tools

workshop_wp2_20180611_vg_v1

Execute selection (do)

```
187
188
189 ********************************************************************************
190 *
191 *                    Setting Up Stata
192 *
193 ********************************************************************************
194
195 * This section is about organising preliminary settings in Stata *
196
197 * clear the computer memory *
198
199 clear
200
201 /** More causes Stata to display --more-- and pause until any key is pressed.
202     It is usually more convenient to have this function switched off **/
203
204 set more off
205
206 * keep a log file containing your output *
207
208 * close any log files that might already be running *
209
210 capture log close
```

Techies might like to use ctrl d

Installing Packages

- *findit* command can be used to search the Stata site and other sites for information

- *findit qvgraph*

- *ssc install* followed by the new programme's name

Estimating Work Time…

Clint Eastwood
The Enforcer

GOOD LUCK!

Aim for GOLD and we might end up with BRONZE

Wednesday 25/07/2018 @ 16:45:

- Proficient in enabling raw data for analysis in Stata
- Knowledgeable in the core principles and practices of data wrangling
- Awareness of open source tools that support reproducible analyses
- Infused with a passion for data management…

Don't quit.
Suffer now
and live the
rest of your
life as a
champion.
Muhammad Ali -

Part III

Extensions

Open Source

Stata is a comprehensive and easy-to-learn statistical software package, however not suitable for every task/situation:

1. Proprietary

2. Likes variable-by-case data structures

3. Necessary but not sufficient for reproducible analyses

Open Source

Alternatives:

1. Gretl, R
2. JSON, XML best handled in Python
3. Jupyter Notebooks

Gretl

Gnu Regression, Econometrics and Time-series Library

gretl

Is a cross-platform software package for econometric analysis, written in the C programming language. It is free, open-source software. You may redistribute it and/or modify it under the terms of the GNU General Public License (GPL) as published by the Free Software Foundation.

gretl conference 2017

The fifth biennial gretl conference took place in Athens on June 2-3, 2017; details at gretlconference.org. Previous conferences took place in Berlin (2015), Oklahoma City (2013), Toruń (2011) and Bilbao (2009).

Features

- Easy intuitive interface (now in French, Italian, Spanish, Polish, German, Basque, Catalan, Galician, Portuguese, Russian, Ukrainian, Turkish, Czech, Traditional Chinese, Albanian, Bulgarian, Greek, Japanese and Romanian as well as English)
- A wide variety of estimators: least squares, maximum likelihood, GMM; single-equation and system methods
- Time series methods: ARIMA, a wide variety of univariate GARCH-type models, VARs and VECMs (including structural VARs), unit-root and cointegration tests, Kalman filter, etc.
- Limited dependent variables: logit, probit, tobit, sample selection, interval regression, models for count and duration data, etc.
- Panel-data estimators, including instrumental variables, probit and GMM-based dynamic panel models
- Output models as LaTeX files, in tabular or equation format
- Integrated powerful scripting language (known as hansl), with a wide range of programming tools and matrix operations
- GUI controller for fine-tuning Gnuplot graphs
- An expanding range of contributed function packages, written in hansl
- Facilities for easy exchange of data and results with GNU R, GNU Octave, Python, Ox and Stata

R

[Home]

Download

CRAN

R Project

About R
Logo
Contributors
What's New?
Reporting Bugs
Development Site
Conferences
Search

R Foundation

Foundation
Board
Members
Donors
Donate

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please choose your preferred CRAN mirror.

If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

News

- The R Foundation has been awarded the Personality/Organization of the year 2018 award by the professional association of German market and social researchers.
- **R version 3.5.1 (Feather Spray) prerelease versions** will appear starting Friday 2018-06-22. Final release is scheduled for Monday 2018-07-02.
- **R version 3.5.0 (Joy in Playing)** has been released on 2018-04-23.

News via Twitter

The R Foundation
@_R_Foundation
Useful advice for package maintainers using tidyverse.

Alternative Data Structures

{"help": "https://data.gov.au/api/3/action/help_show?name=datastore_search", "success": true, "result": {"resource_id": "eb1e6be4-5b13-4feb-b28e-388bf7c26f9
{"type": "text", "id": "Operates_in_NSW"}, {"type": "text", "id": "Operates_in_NT"}, {"type": "text", "id": "Operates_in_QLD"}, {"type": "text", "id": "Opera
g_social_or_public_welfare"}, {"type": "text", "id": "Advancing_security_or_safety_of_Australia_or_Australian_public"}, {"type": "text", "id": "Another_purpo
"text", "id": "Women"}, {"type": "text", "id": "Youth"}], "records": [{"Victims_of_crime": "", "Operating_Countries": "", "Charity_Legal_Name": "Sydney Missi
hed": "", "Address_Line_2": "", "Address_Line_3": "", "Address_Line_1": "43 Badminton Road", "Charity_Website": "www.smbc.com.au", "Number_of_Responsible_Per
on": "", "Operates_in_VIC": "", "Aboriginal_or_TSI": "", "Women": "", "State": "NSW", "Town_City": "Sydney Olympic Park", "Operates_in_NSW": "Y", "Advancing_
umber_of_Responsible_Persons": "14", "Veterans_or_their_families": "", "Country": "AUSTRALIA", "Registration_Date": "03-12-2012", "Operates_in_QLD": "", "ABN
nal_or_TSI": "", "Women": "", "State": "QLD", "Town_City": "Mount Gravatt", "Operates_in_NSW": "", "Advancing_social_or_public_welfare": "", "Unemployed_Pers
03-12-2012", "Operates_in_QLD": "Y", "ABN": "11000761571", "Youth": "", "Operates_in_SA": "", "General_Community_in_Australia": "Y", "Another_purpose_benefic
_welfare": "", "Unemployed_Persons": "", "People_with_Disabilities": "", "Operates_in_TAS": "", "Advancing_security_or_safety_of_Australia_or_Australian_publ
ther_purpose_beneficial_to_the_community": "", "People_at_risk_of_homelessness": "", "HPC": "", "Pre_Post_Release_Offenders": "", "PBI": "", "Financial_Year_
rposes_beneficial_to_ther_general_public_and_other_analogous": "", "Operates_in_NT": "", "Advancing_Education": "", "Men": "", "Promote_or_oppose_a_change_to
, "Advancing_Health": "", "Other_Organisation_Names": "", "Operates_in_WA": "", "Ethnic_Groups": "", "Operates_in_ACT": "", "Preventing_or_relieving_sufferin
ing_or_protecting_human_rights": "", "Victims_of_Disasters": "", "Promoting_reconciliation__mutual_respect_and_tolerance": "", "Communities_Overseas": "", "P
, "Advancing_natual_environment": "", "Aged_Persons": "", "Postcode": "3168", "Charity_Size": "Small", "_id": 6, "Children": ""}, {"Victims_of_crime": "", "O
, "Gay__Lesbian__Bisexual": "", "Date_Organisation_Established": "01-01-1979", "Address_Line_2": "", "Address_Line_3": "", "Address_Line_1": "106-114 Mount D
"", "Charity_Legal_Name": "Pacific Hills Christian Education Limited", "Advancing_Religion": "Y", "Operates_in_VIC": "", "Aboriginal_or_TSI": "", "Women": "
site": "", "Number_of_Responsible_Persons": "16", "Veterans_or_their_families": "", "Country": "AUSTRALIA", "Registration_Date": "03-12-2012", "Operates_in_Q
d": 8, "Children": "Y"}, {"Victims_of_crime": "", "Operating_Countries": "", "Charity_Legal_Name": "Burrum Benevolent Society Inc", "Advancing_Religion": "",
ine_3": "", "Address_Line_1": "1 Prosper Street", "Charity_Website": "", "Number_of_Responsible_Persons": "6", "Veterans_or_their_families": "", "Country": "
SI": "Y", "Women": "", "State": "NSW", "Town_City": "Armidale", "Operates_in_NSW": "Y", "Advancing_social_or_public_welfare": "", "Unemployed_Persons": "",
tration_Date": "03-12-2012", "Operates_in_QLD": "", "ABN": "11002389746", "Youth": "", "Operates_in_SA": "", "General_Community_in_Australia": "", "Another_p
"", "People_with_Disabilities": "", "Operates_in_TAS": "", "Advancing_security_or_safety_of_Australia_or_Australian_public": "", "Migrants__Refugees_or_Asylu
e_beneficial_to_the_community": "", "People_at_risk_of_homelessness": "", "HPC": "", "Pre_Post_Release_Offenders": "", "PBI": "", "Financial_Year_End": "31/1
eneficial_to_ther_general_public_and_other_analogous": "", "Operates_in_NT": "", "Advancing_Education": "", "Men": "", "Promote_or_oppose_a_change_to_law__go
dvancing_Health": "", "Other_Organisation_Names": "", "Operates_in_WA": "", "Ethnic_Groups": "", "Operates_in_ACT": "", "Preventing_or_relieving_suffering_of
cting_human_rights": "", "Victims_of_Disasters": "", "Promoting_reconciliation__mutual_respect_and_tolerance": "", "Communities_Overseas": "", "People_with_O
ulture": "", "Advancing_natual_environment": "", "Aged_Persons": "", "Postcode": "3166", "Charity_Size": "Small", "_id": 13, "Children": ""}, {"Victims_of_cr
, "Gay__Lesbian__Bisexual": "", "Date_Organisation_Established": "01-01-1950", "Address_Line_2": "431 St Kilda Road", "Address_Line_3": "", "Address_Line_1":
ting_Countries": "", "Charity_Legal_Name": "Austin Medical Research Foundation", "Advancing_Religion": "", "Operates_in_VIC": "Y", "Aboriginal_or_TSI": "",
alth", "Charity_Website": "www.austinmrf.org.au", "Number_of_Responsible_Persons": "15", "Veterans_or_their_families": "", "Country": "AUSTRALIA", "Registrat
State": "VIC", "Town_City": "Woodend", "Operates_in_NSW": "", "Advancing_social_or_public_welfare": "", "Unemployed_Persons": "", "People_with_Disabilities":
D": "", "ABN": "11005151771", "Youth": "Y", "Operates_in_SA": "", "General_Community_in_Australia": "", "Another_purpose_beneficial_to_the_community": "", "P
perates_in_TAS": "", "Advancing_security_or_safety_of_Australia_or_Australian_public": "", "Migrants__Refugees_or_Asylum_Seekers": "", "Purposes_beneficial_t
": "", "People_at_risk_of_homelessness": "", "HPC": "", "Pre_Post_Release_Offenders": "", "PBI": "", "Financial_Year_End": "30/06", "Address_Type": "Business

Voila

auscharities_20180424 - Excel

A1 | Victims_of_crime

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Victims_of	Operating	Charity_Le	Advancing	Operates_	Aboriginal	Women	State	Town_City	Operates_	Advancing	Unemploy	People_wi	Operates_	Advancing	Migrants_	Purposes_
2			Sydney Mi	Y	Y			NSW	Croydon	Y				Y			
3			Integricare					NSW	Sydney Olympic Park	Y	Y						Y
4		BRA,KHM,	Australian	Y				QLD	Mount Gravatt								
5		GRC	The Greek	Y				NSW	HAMILTON	Y							
6			Wangarang Industries Limited					NSW	Orange	Y							
7		IDN,ESP	Subud Aus	Y	Y			VIC	CLAYTON	Y				Y			
8			Bethel Christian School Limited					NSW	Mount Druitt	Y							
9			Pacific Hill	Y				NSW	DURAL	Y			Y				
10			Burrum Benevolent Society Inc				Y	QLD	Howard			Y	Y				
11			St Peters Pre-School Ltd			Y		NSW	Armidale	Y							
12			Chrysalis School For Rudolf Steiner Education Ltd					NSW	Bellingen	Y							
13			Lasercraft Australia Ltd					NSW	West Gosford	Y	Y						
14			ISCAST LTI	Y	Y			VIC	Oakleigh	Y				Y			Y
15			Nurses Memorial Cei	Y				VIC	MELBOURNE								
16			Austin Medical Rese	Y				VIC	Heidelberg								
17			Braemar College Lim	Y				VIC	Woodend								
18			Buddha Sa	Y				NSW	MEDLOW BATH	Y							
19			Odyssey House Victo	Y				VIC	RICHMOND								
20			The Dla Fc	Y	Y			VIC	Hawthorn	Y				Y			
21			St George Family Support Services Inc.							Y	Y						
22			Alphington Grammai	Y				VIC	Alphington		Y						

auscharities_20180424

Ready Scroll Lock 100%

12:04 25/06/2018

Jupyter Notebooks

Summary

Rationale for the workshop:

"Many organizations can barely find a way to use their R/Python programmers on reasonable datasets."

"The piece missing from the data science movement right now is really simple: intelligent application of data science tools."

https://www.linkedin.com/pulse/data-science-dead-5-years-less-justin-b-dickerson-phd-mba-pstat- accessed 16.07.2018.

Plenary

Have at it:

- Outstanding queries
- Potential applications of Stata and data wrangling in your organisation
- Unmet needs

Data Wrangling & Munging: Organising, Managing and Enabling Data for Analysis©

Prof. John MacInnes & Dr Diarmuid McDonnell

University of Edinburgh

July 2018

© Diarmuid McDonnell

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Administrative Data Research Centre Scotland

AQMeN Applied Quantitative Methods Network

THE UNIVERSITY of EDINBURGH