

Transparency and Reproducibility Tools

Paul Hofman

Transparency and Reproducibility workshop, 3 December 2018

Overview

- ① Automatic Version Control
- ② Automation
- ③ Good Code
- ④ Latex
- ⑤ R Markdown

Version Control

- Tracks changes in (text) documents
- Access to old versions
- Easily compare changes with previous versions

Several ways:

- Manually with file names
- Version Control System (Git, Subversion)

Git

- Developed by Linus Torvalds (in 2005) to control development of Linux
- *Distributed* revision-control system
- Aimed at non-linear development and collaboration
- Track changes in text files
- Internet access not required (!)
- Free! But...

Easiest to use are commercial offerings:

- Github
 - Free public folders
 - Free private folders for researchers
 - Great desktop app
- Gitlab
 - Netherlands-based
 - WUR pays for own server (we have free access!)

Git workflow

- 1 Clone Repository ('Repo', or research folder)
- 2 Work normally
- 3 Commit changes
- 4 Upload commit

Git Demo

Demonstration

Git Summary

Advantages:

- Very robust and oft-used version control system
- Easy to learn (especially with desktop app)
- Requires (positive) change in workflow: conscious changes ('commits')
- Can automatically merge conflicting files

Disadvantages:

- No automatic support for files over 100 MB
- Advanced usage requires command line or website

Automation

- Research involves a lot of steps
 - Research Design
 - Collect data
 - Clean data
 - Analyse data
 - Output tables/graphs
 - Write Paper
- Many of these steps can be automated
- Most of us are already doing this partially (do-files, Rscripts)
- But more is better

Advantages of Automation

- Less prone to errors
- Better reproducibility (Clear path from 'raw' data to paper)
- Easier to make changes

How to do it

- Find every manual step in your process, and try to eliminate it
- Output tables to .rtf, .tex or .csv, and automatically include them in your paper
- Tie your same-language scripts together (easy in Stata, less so in Python and R)
- More advanced: tie different-language scripts together (rundirectory.py)
- Some stuff on my website: www.hofmanpaul.com/automation

rundirectory.py

```
72 # The first step is to append the 2015 data. Each village has its own e
73 run_python("01_Data/03_survey2015/01_addaxhh/append2015.py")
74
75 # Next, we clean and collate the individual yeardata together, with Sta
76 run_stata("01_Data/06_Code/01_2010 Clean and Collate/clean2010.do")
77 run_stata("01_Data/06_Code/02_2012 Clean and Collate/clean2012.do")
78 run_stata("01_Data/06_Code/03_2015 Clean and Collate/clean2015.do")
79
80 # Next, several preparatory steps for the analysis
81 run_stata("01_Data/06_Code/04_Merge/01_AddVillagecodes.do")
82 run_stata("01_Data/06_Code/04_Merge/02_MergePrep.do")
83 run_stata("01_Data/06_Code/04_Merge/03_Merge.do")
84 run_stata("01_Data/06_Code/04_Merge/04_CleaningPostMerge.do")
85
86 # Now we use R to extract raster data from .tif files, which are used i
87 run_R("06_maps/extract_rasterHansen.R")
88 run_R("06_maps/extract_rasterEVI.R")
89
90 # Then, we run the analysis files
91 run_stata("01_Data/06_Code/05_Analysis/02_Paper/06_Analysis.do") #There
92 run_stata("01_Data/06_Code/05_Analysis/02_Paper/ExtraAnalyses.do") # Th
93
94 # And finally make the Latex file
95 run_latex("08_Paper/Paper Addax.tex")
```

How to do it

- Find every manual step in your process, and try to eliminate it
- Output tables to .rtf, .tex or .csv, and automatically include them in your paper
- Tie your same-language scripts together (easy in Stata, less so in Python and R)
- More advanced: tie different-language scripts together (rundirectory.py)
- Some stuff on my website: www.hofmanpaul.com/automation

Good Code

- We spend a lot of time writing code
 - To communicate with the computer: do this
 - But we also communicate with ourselves: how did I make Figure 2 in research Project X 3 years ago?
- Good Code is a necessary skill (or art?) for the modern economist
- Guido van Rossum's (Python) key insight: computer code is more often read than written

Some Guidelines

- Be consistent
- Readability is more important than succinctness
- Use meaningful but short names
 - no: for i in X
 - yes: for var in editvars
- Variable names: lower_case_underscores or CamelCase (etc)
- Loop as much as you can
- Indent in loops and rarely otherwise
- Limit line length (80-100 characters)
 - To fit two files on one screen
 - For Github track changes
- Write functions for things you do often (more advanced)

Commenting

- 'Block comments': brief explanation of next section of code. Always do this
- 'Inline comments': use sparingly. The code should communicate what happens
- But, always comment choices ('Winsorize income at 90%')
- Assume that your reader knows the language (if they don't they shouldn't look at your code)
- Be careful: code cannot be out of date, but comments can be. Inconsistency between code and comments is the worst

Latex

- Latex is a typesetting system (alternative to Microsoft Word)
- Created by Leslie Lamport in 1983
- Designed for scientific documenting
- Gives you more control over your documents
- Based on the Tex typesetting engine created in 1978 by computer programming legend Donald Knuth
 - Tex is a 'finished' product

Advantages of Latex

- Superior typesetting quality
- Equations are easy
- Git can track changes and version history
- Separate writing and typesetting (can reduce distraction)
- Fantastic for scientific presentations
- Great for automation: automatically insert figures and tables
- Good support for citations (BibTex)
- Nerds love it (= Great online support)
- Makes you look smart

Typesetting quality

Microsoft Word 2008

Call me Ishmael. Some years ago – never mind how long precisely – having little or no money in my purse, and nothing particular to interest me on shore, I thought I would sail about a little and see the watery part of the world. It is a way I have of driving off the spleen, and regulating the circulation. Whenever I find myself growing grim about the mouth; whenever it is a damp, drizzly November in my soul; whenever I find myself involuntarily pausing before coffin warehouses, and bringing up the rear of every funeral I meet; and especially whenever my hypos get such an upper hand of me, that it requires a strong moral principle to prevent me from deliberately stepping into the street, and methodically knocking people's hats off – then, I account

Adobe InDesign CS4

Call me Ishmael. Some years ago – never mind how long precisely – having little or no money in my purse, and nothing particular to interest me on shore, I thought I would sail about a little and see the watery part of the world. It is a way I have of driving off the spleen, and regulating the circulation. Whenever I find myself growing grim about the mouth; whenever it is a damp, drizzly November in my soul; whenever I find myself involuntarily pausing before coffin warehouses, and bringing up the rear of every funeral I meet; and especially whenever my hypos get such an upper hand of me, that it requires a strong moral principle to prevent me from deliberately stepping into the street, and methodically knocking people's hats off – then, I account it high time to get to sea

pdf-LaTeX 3.1415926

Call me Ishmael. Some years ago – never mind how long precisely – having little or no money in my purse, and nothing particular to interest me on shore, I thought I would sail about a little and see the watery part of the world. It is a way I have of driving off the spleen, and regulating the circulation. Whenever I find myself growing grim about the mouth; whenever it is a damp, drizzly November in my soul; whenever I find myself involuntarily pausing before coffin warehouses, and bringing up the rear of every funeral I meet; and especially whenever my hypos get such an upper hand of me, that it requires a strong moral principle to prevent me from deliberately stepping into the street, and methodically knocking people's hats off – then, I account it high time to get to

Typesetting quality

Hyphenation and inter-word spacing statistics

	Word	InDesign	pdf-Latex
Number of hyphenations	9	10	4
SD of IWS (pt)	2.26	1.94	1.42
Maximum IWS (pt)	14.4	13.2	9.0
Number of lines with IWS > 9 pt	5	2	0

SD: standard deviation; IWS: inter-word spacing

Advantages of Latex













- Superior typesetting quality
- Equations are easy
- Git can track changes and version history
- Separate writing and typesetting (can reduce distraction)
- Fantastic for scientific presentations
- Great for automation: automatically insert figures and tables
- Good support for citations (BibTeX)
- Nerds love it (= Great online support)
- Makes you look smart

Git and Latex

Worked on Maarten's comments

 Paul Hofman committed  bef4462  13 changed files

Next, continue comments from fig 6 (EVI) onwards)

01_Data/06_Code/05_Analy.../06_Analysis.do		178	179	~~~~~
01_Data/08_Images/incproportions.png		179		-\begin{figure}[!htbp]\centering
01_Data/08_Images/trends_evi.png			180	+\begin{figure}[!htbp]\centering %This figure is made in QGIS
06_maps/EVI_correlations.R		180	181	\caption{Village Locations}
06_maps/EVIoutput.csv		181	182	\includegraphics[width=\textwidth]{Villages Map nosatellite.png}
06_maps/Hansen_GFC...N_020W.tif.aux.xml		182	183	\label{fig:vilmap}
06_maps/Villages Map nosatellite.png			184	+ \floatfoot{Shows location of all villages examined in this study. Data comes from GPS coordinates that were collected during data collection in 2010 and 2012.}
06_maps/Villages Map.qgs		183	185	\end{figure}
06_maps/Villages Map.qgs~		184	186	\begin{figure}[!htbp]\centering
06_maps/extract_rasterEVI.R		185	187	\caption{Income Proportions (2010-2012)}
08_Paper/Paper Addax.pdf		186	188	\includegraphics[width=\textwidth]{incproportions.png}
08_Paper/Paper Addax.tex		187	189	\label{fig:incproportions}
		188	190	+ \floatfoot{Shows proportions of traditional income (that is, excluding 'new' income sources like land lease payments and salaried income). Other income includes remittances, self-declared other revenues and pension income.}
			191	\end{figure}
		189	192	-\begin{figure}[!htbp]\centering
		190	193	+\begin{figure}[!htbp]\centering %Made in QGIS
		191	194	\caption{Forest Loss Map}
		192	195	-\includegraphics[width=\textwidth]{Parallel trends 2.png}
		193	196	+\includegraphics[width=\textwidth]{hansen_bw.png}
		194	197	\label{fig:trendsmap}
		195	198	-\floatfoot{Source: Hansen/UMD/Google/USGS/NASA}
		196	199	+\floatfoot{This map shows forest loss over time around the sample villages (re

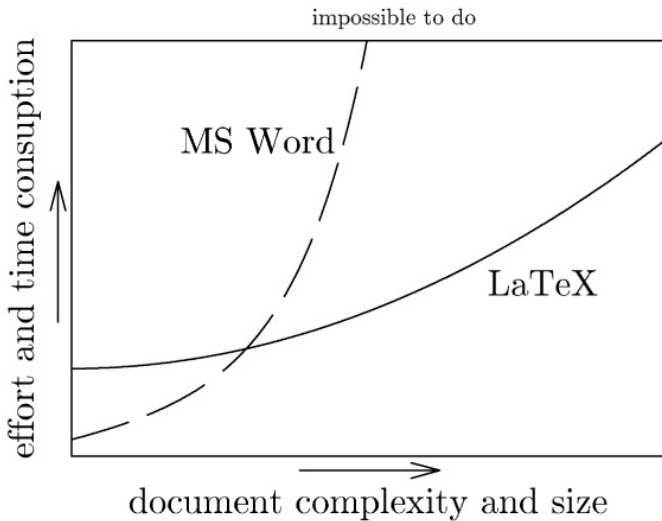
Advantages of Latex

- Superior typesetting quality
- Equations are easy
- Git can track changes and version history
- Separate writing and typesetting (can reduce distraction)
- Fantastic for scientific presentations
- Great for automation: automatically insert figures and tables
- Good support for citations (BibTex)
- Nerds love it (= Great online support)
- Makes you look smart

Disadvantages of Latex

- Steep learning curve
 - Presentations can be a good introduction
- Syntax is cumbersome
- Your co-authors don't know how to use it
- Word is sometimes a requirement
 - There are programs that convert (pandoc)
- 'Building' your document can take a while

Word vs Latex



R Markdown

- Markdown is a 'lightweight' version of Latex: popular for websites (my website uses it)
- Created by tech blogger John Gruber in 2004
- **R** Markdown is the statistical suite R, integrated with Markdown
- Allows you to import data, run analyses and automatically put them in a table - all in one document

R Markdown example

```
188 ## Outcomes
189 We are interested in a number of outcomes related to marketing, production, income, and savings. `r
    table_nums('outcomes', display = 'cite')` shows the descriptives of the outcome variables of interest coming from a
    survey conducted in 2015 in the Kenema district.

190
191 ```{r 'outcomes', echo = F, results = 'asis'}
192 dfOutcomes = dfOutcomes[,1:5]
193 dfOutcomes = dfOutcomes %>% mutate_if(is.numeric, round, 3)
194 kable(dfOutcomes, format = 'latex', booktabs = T, caption = 'Mean, standard deviation, and intra cluster correlation
    for outcome variables in the region') %>%
195 footnote("Sample comes from early 2016 data collection over the previous season in Kenema District. Sample is
    restricted to cocoa farmers (about 50% of total sample). ICC is the Intra-Cluster Correlation, WCV the
    Within-Cluster Variation. Types of buyers are traders, cooperatives and motorbike buyers. Locations are Kenema, town
    and own village", threeparttable = T)

196
197
198
199 ## Power calculations
200 With the available data we can do some rigorous power calculations in order to determine the necessary sample size.
    We do two separate power calculations for the two treatments and calculate sample or cluster size for each outcome
    variable. `r table_nums('powercalc1', display = 'cite')` shows the power calculations for the credit intervention,
    which is at the individual level. The columns give the necessary sample size to achieve power of 0.8 to detect
    effect sizes of 30, 20, and 10% of the mean respectively. `r table_nums('powercalc2', display = 'cite')` calculates
    the number of clusters (villages) in order to achieve power of 0.8 to detect effect sizes of 30, 20, and 10% of the
    mean respectively.
```

R Markdown example

Table 2: Mean, standard deviation, and intra cluster correlation for outcome variables in the region

Variable	N	Mean	SD	ICC
Price received for 1 kg of cocoa (in 1'000 Le)	511	10.042	10.656	0.040
Earnings from cocoa in 2015 (in 1'000 Le)	659	536.627	846.111	0.101
Total cocoa production in 2015 (Kg)	609	101.656	123.353	0.103
Total losses to black pod in 2015 (Kg)	683	26.290	42.138	0.043
Cocoa farm size (Ha)	702	1.590	1.809	0.062
# of seedlings expanded with	454	275.542	359.810	0.051
# Days worked on cocoa farm in 2015	642	63.005	68.815	0.313
Used cocoa to repay a loan (1=yes)	552	0.056	0.230	0.016
Monthly expenditures	701	333.780	287.625	0.057
Assets score	699	25.216	14.198	0.058
House quality score (max 35)	703	11.773	3.270	0.240
Have savings (1=yes)	702	0.188	0.391	0.058
Amount saved (in 1'000 Le)	121	611.678	1028.278	0.000
No of types of buyers farmer sold to in 2015 (max 3)	552	1.116	0.378	0.035
No of types of locations farmers sold to in 2015 (max 3)	552	1.156	0.406	0.059

Note:

Sample comes from early 2016 data collection over the previous season in Kenema District. Sample is restricted to cocoa farmers (about 50% of total sample). ICC is the Intra-Cluster Correlation, WCV the Within-Cluster Variation. Types of buyers are traders, cooperatives and motorbike buyers. Locations are Kenema, town and own village

Outcomes

We are interested in a number of outcomes related to marketing, production, income, and savings. Table 2 shows the descriptives of the outcome variables of interest coming from a survey conducted in 2015 in the Kenema district.

Advantages of R Markdown

- Very simple syntax
- Can include blocks of Latex code for more complicated sections
- Your entire paper and the code for analysis in one document
- Good support for citations
- Fast

Disadvantages of R Markdown

- Only capable of producing simple documents
 - Though including Latex sections makes it capable of producing anything
- Including code makes the document very long (and a distraction)
- Manually making tables is pretty bad
- There are different 'implementations' of *Markdown*: sometimes uncertain how your document will look

Example Workflow

Paper that has Python, Stata, R and Latex code




- 1 Raw Data is in appropriate folders, some files downloaded from internet and put in correct folders
- 2 Python and Stata clean files
- 3 These clean files are picked up by Stata and R to produce analyses tables as .tex files and .png images
- 4 These .tex files are picked up by Latex and put in a .pdf
- 5 One file, rundirectory.py runs all these scripts in sequence

Version Control by Github

Assorted

- Use a good Text Editor (Sublime Text, Atom, Textmate)
 - I use Sublime Text for R, Latex and Python. Stata maybe soon

Further Reading

- Code and Data for the Social Sciences: A Practitioner's Guide - Matthew Gentzkow and Jesse Shapiro. Available here 
 - Extremely thorough guide by two economists covering almost all aspects covered here
- How to improve your relationship with your future self - Jake Bowers and Maarten Voors. Available here 
 - Focuses a lot on writing good code, and R Markdown
- PEP8 – Style guide for python code – Guido van Rossum Available here 
 - Really written for Python, but gives great guidelines on coding style