Indiana University

Reproducible Results and the Workflow of Data Analysis

Scott Long

Departments of Sociology and Statistics

www.indiana.edu/~jslsoc/ftp/

Workshop in Methods | September 2019

Roadmap

- Open science, replication, and reproducible results
- What is a workflow for reproducible results?
- Criteria for selecting a workflow
- Workflow tasks: planning, organization, documentation
- Workflow for computing
- Data cleaning, analysis, and presentation
- Preserving files
- Collaboration
- Developing your workflow

The reproducible results movement

- Open Science: transparancy and accessibility
- Integrity in research
- NAS Committee on Reproducibility and Replicability in Science

Changing expectations

- Journals require submission of data and analysis files
- Funding agencies strengthen requirements for data access
- Haverford College students post reproducible results on Dataverse

Accountability

- Retraction Watch: Tracking retractions as a window into the scientific process

Retraction due to coding error

RETRACTED: In Sickness and in Health? Physical Illness as a Risk Factor for Marital Dissolution in Later Life

[redacted]

Abstract

The health consequences of marital dissolution are well known, but little work has examined the impact of health on the risk of marital dissolution. In this study we use a sample of 2,701 marriages from the Health and Retirement Study (1992–2010) to examine the role of serious physical illness onset (i.e., cancer, heart problems, lung disease, and/or stroke) in subsequent marital dissolution due to either divorce or widowhood. We use a series of discrete-time event history models with competing risks to estimate the impact of husband's and wife's physical illness onset on risk of divorce and widowhood. We find that only wife's illness onset is associated with elevated risk of divorce, while either husband's or wife's illness onset is associated with elevated risk of widowhood. These findings suggest the importance of health as a determinant of marital dissolution in later life via both biological and gendered social pathways.

Keywords

aging, chronic disease, gender, mortality, physical health

A large body of literature has identified marital status as a strong predictor of health and well-being. Not only are the married healthier than the unmarried (e.g., Lillard and Waite 1995; Emberson 1992), but studies find that both divorce and widowhood are precursors to declines in physical and mental health (e.g., Hughes and Waite 2009; Williams and Umberson 2004). Less attention, however, has been paid to how health may be a determinant of marital status. Work in this area has tended to focus on the positive selection of the healthier into marriage (e.g., Byrne et al. 1989; Smith and Smith 2010), but poor health may be an equally important force for selection, and the work has examined the impact of health on the risk of marital dissolution. In this study we use a sample of 2,701 marriages from the Health and Retirement Study (1992–2010) to examine the role of serious physical illness onset (i.e., cancer, heart problems, lung disease, and/or stroke) in subsequent marital dissolution due to either divorce or widowhood. We use a series of discrete-time event history models with competing risks to estimate the impact of husband's and wife's physical illness onset on risk of divorce and widowhood. We find that only wife's illness onset is associated with elevated risk of divorce, while either husband's or wife's illness onset is associated with elevated risk of widowhood. These findings suggest the importance of health as a determinant of marital dissolution in later life via both biological and gendered social pathways.

Booth, and Johnson 2008). Illness may initiate changes to spouses' roles—in particular, increasing caregiving responsibilities for the healthy spouse—which can tax marital relationship dynamics (Wolff and Kasper 2006). Illness may also decrease household income due to the inability of one or both spouses to work (Teachman 2010), which may increase marital strain.

Only a few studies have examined the role of poor health in subsequent divorce, and these studies are mixed in their findings, with some finding

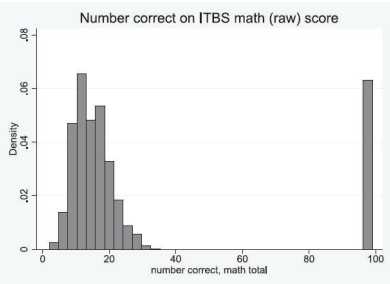
Incorrect data in published research

Distributional Analysis in Educational Evaluation: A Case Study from the New York City Voucher Program

Marianne Bitler, Thurston Domina, and Emily Penner
University of California, Irvine, Irvine, California, USA

Hilary Hoynes
University of California, Berkeley, Berkeley, California, USA

Abstract: We use quantile treatment effects estimation to examine the consequenc assignment New York City School Choice Scholarship Program across the distrib achievement. Our analyses suggest that the program had negligible and statistic effects across the skill distribution. In addition to contributing to the literature o the article illustrates several ways in which distributional effects estimation can er research: First, we demonstrate that moving beyond a focus on mean effects esti possible to generate and test new hypotheses about the heterogeneity of educational that speak to the justification for many interventions. Second, we demonstrate t effects can uncover issues even with well-studied data sets by forcing analysts to v new ways. Finally, such estimates highlight where in the overall national achievemen scores of children exposed to particular interventions lie; this is important for explo validity of the intervention's effects.



Fragility of published results

Measurement, methods, and divergent patterns: Reassessing the effects of same-sex parents ☆

Simon Cheng [a,1], Brian Powell [b,1]

[a] 344 Mansfield Rd., Department of Sociology, University of Connecticut, Storrs, CT 06269, United States
[b] 744 Ballantine Hall, 1020 E. Kirkwood Ave., Department of Sociology, Indiana University, Bloomington, IN 47405-7103, United States

ARTICLE INFO

Article history:
Received 8 October 2013
Revised 24 March 2015
Accepted 8 April 2015
Available online 23 April 2015

Keywords:
Children
Family structure
Methodology
Same-sex parenting
Sexuality

ABSTRACT

Scholars have noted that survey analysis of small subsamples—for example, same-sex parent families—is sensitive to researchers' analytical decisions, and even small differences in coding can profoundly shape empirical patterns. As an illustration, we reassess the findings of a recent article by Regnerus regarding the implications of being raised by gay and lesbian parents. Taking a close look at the New Family Structures Study (NFSS), we demonstrate the potential for misclassifying a non-negligible number of respondents as having been raised by parents who had a same-sex romantic relationship. We assess the implications of these possible misclassifications, along with other methodological considerations, by reanalyzing the NFSS in seven steps. The reanalysis offers evidence that the empirical patterns showcased in the original [redacted] article are fragile—so fragile that they appear largely a function of these possible misclassifications and other methodological choices. Our replication and reanalysis of [redacted] study offer a cautionary illustration of the importance of double checking and critically assessing the implications of measurement and other methodological decisions in our and others' research.

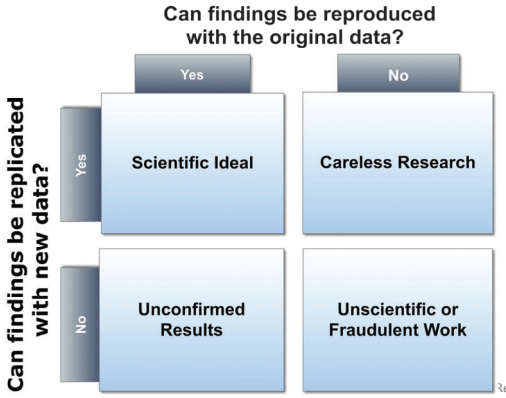
Lost in the flood

A chemical engineer … who claims his supporting data were wiped out in a flood has notched his ninth retraction, seven from a single journal, for suspicious images and related issues. — retractionwatch.com 20190903

Replication and reproduction of results

Reproducibility: identical results with the *same data*.

Replicability: confirmation of results with *new data*.



Challenges to replicability

Fraud

Sample driven analyses

Decisions based on unique characteristics of the sample.

- Data mining portrayed as theory testing
- Post analysis hypothesis construction
- Undocumented specification searches and p-hacking
- "Cherry picking" the sample

Using a sample to select a model for diabetes

1. Consider six random sub-samples.
2. Stepwise regression selects four different models
   - Does being female significanty affect diabetes?

Variable	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
bmi	1.067***	1.066***	1.004	1.074***	1.101***	0.971
white	0.518***	0.547***	0.521***	0.543***	0.505***	0.562***
age	1.262***	1.351***	1.324***	1.288***	1.282***	1.341***
agesq	0.999***	0.998***	0.998***	0.998***	0.998***	0.998***
hsdegree	0.720***	0.680***	0.662***	0.749***	0.780***	0.650***
weight	1.006***	1.006***	1.016***	1.004**	-----	1.022***
height	-----	-----	0.936**	-----	-----	0.909***
female	-----	-----	-----	0.854*	0.733***	-----

Legend: p<.1; ** p<.05; *** p<.01

Model variability versus sampling variability

Young and Holsteen. 2015. Model Uncertainty and Robustness. *SMR*.

- Point estimates capture ''one ad-hoc route through the thicket of possible models'' (Leamer 1985:308).
- For example, do higher income tax rates cause taxpayers to ''vote with their feet'' and migrate to states with lower taxes?

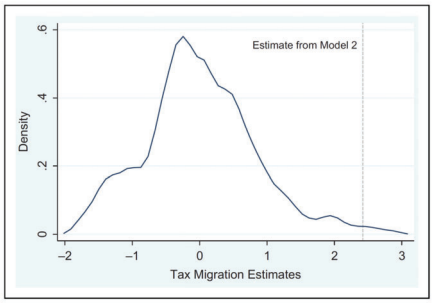


Figure 4. Modeling distribution of tax migration estimates.
Note: Kernel density graph of estimates from 24,576 models.

Reproducibility with same data

Can you show me exactly how you got your results?

Emerging expectations for reproducibility

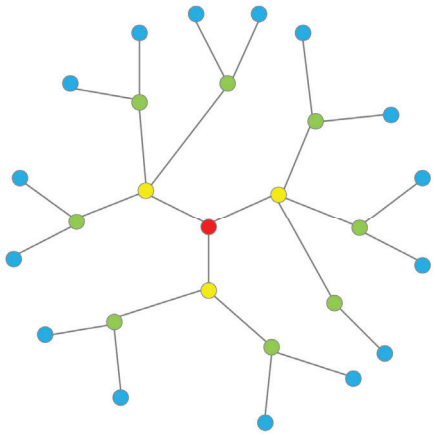
- AJPS requires verification of results before a paper is reviewed.
  : Five of 200 submissions succeeded.
- Some jounals require data and script files be submitted.
  : Sometimes with submission
  : Sometimes with acceptance
- Why do this even if not required? An embarassing example…

Why results are hard to reproduce

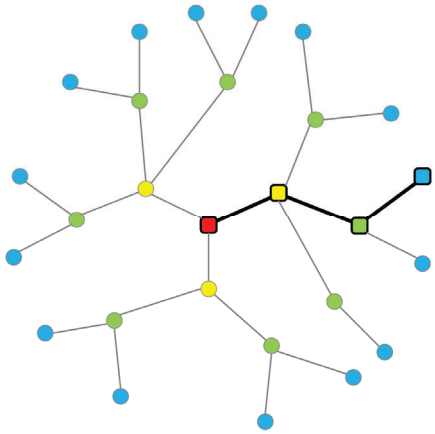
1. The curse of dimensionality: Research involves many decisions.
   - Where to truncate a variable?
   - What seed for the RN generator?
   - How to scale with partially missing data?
   - Which cases to keep for analysis?
   - How to code education?
   - What values to assign to income greater than $200,000?
   - And so on...

With only 10 choices, there are 1,024 combinations.

Decisions in the path to analysis: the choices that could be made



Decisions in the path to analysis: the choices made



Why are results hard to reproduce? (continued)

2. Missing documentation so you can't retrace your path.
3. Newer software can produce different results.
   - A colleague spent weeks trying to reproduce results that differed because of new software.
4. Missing files make reproducibility impossible.
   - Retractions because of "lost" data.

Reproducibility and workflow

- Reproducibility requires a systematic workflow.
- My talk considers this topic.

What is a workflow?

A workflow is *coordinated* procedures for all aspects of data management, analysis, and presentation.

- Planning research
- Organizing and documenting
- Importing and cleaning data
- Analyzing data
- Presenting and publishing results
- Revising results
- Preserving files

You have a workflow

1. Your workflow might be:
   - Planned
   - Ad hoc
   - Planned in an ad hoc way
2. You can improve your workflow with a modest investment of time.
   - Thinking about WF makes it better.
   - The less experience you have, the easier it is to improve.
   - It takes time to learn, but saves much more time.
   - It prevents errors.
   - It makes you a better data analyst.

Origins of the workflow project

1. A dissertation delayed 18 months to determine provenance.
2. A paper's single 743 line do-file that didn't reproduce any results.
3. Conflicting results from the "same" dataset.

   "The datasets are exactly the same except for the married variable."
4. The wrong variable used in analyses for NAS report.
5. Mislabeled gene in a study of alcoholism.
6. Misleading output such as...

Definitel a problem

```
. tabulate female sdchild_v1

      R is |      Q15 Would let X care for children
   female? | Defintel   Probably   Probably   Definitel |     Total
-----------+--------------------------------------------+----------
      Male |        41         99        155        197 |       492
    Female |        73         98        156        215 |       542
-----------+--------------------------------------------+----------
     Total |       114        197        311        412 |     1,034
```


How important is it to…

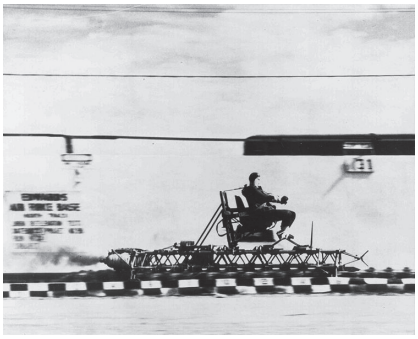
```
. codebook tc1*, compact

Variable    Obs Unique      Mean Min  Max  Label
-------------------------------------------------------------------------------
tc1doc     1074     10 8.714153   1   10  Q46 How important is it to go to ...
tc1fam     1074     10 8.755121   1   10  Q43 How important is it to turn t...
tc1friend  1073     10 7.799627   1   10  Q44 How important is it to turn t...
tc1mhprof  1045     10  7.58756   1   10  Q48 How important is it to go to ...
tc1psy     1050     10 7.567619   1   10  Q47 How important is it to go to ...
tc1relig   1039     10  5.66025   1   10  Q45 How important is it to turn t...
-------------------------------------------------------------------------------
```


Confusing output

```
. tab occ ed, row
           |                                        Years of education
Occupation |         3          6          7          8          9         10         11
12         |     Total
-----------+-------------------------------------------------------------------------------------------------
-----------+----------
    Menial |         0          2          0          0          3          1          3
12         |         2 |        31
           |      0.00       6.45       0.00       0.00       9.68       3.23       9.68
38.71      |      6.45 |    100.00
-----------+-------------------------------------------------------------------------------------------------
-----------+----------
   BlueCol |         1          3          1          7          4          6          5
26         |         7 |        69
           |      1.45       4.35       1.45      10.14       5.80       8.70       7.25
37.68      |     10.14 |    100.00
-----------+-------------------------------------------------------------------------------------------------
-----------+----------
     Craft |         0          3          2          3          2          2          7
39         |         7 |        84
           |      0.00       3.57       2.38       3.57       2.38       2.38       8.33
46.43      |      8.33 |    100.00
-----------+-------------------------------------------------------------------------------------------------
-----------+----------
  WhiteCol |         0          0          0          1          0          1          2
19         |         4 |        41
           |      0.00       0.00       0.00       2.44       0.00       2.44       4.88
46.34      |      9.76 |    100.00
-----------+-------------------------------------------------------------------------------------------------
-----------+----------
```


The foundation of workflow is ironical *optimism*

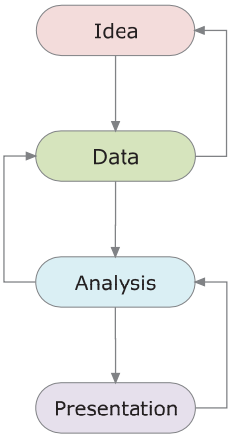
The *universal aptitude for ineptitude* makes any human accomplishment an incredible miracle. – *John Paul Stapp*



From 0 to 995mph and back in 3 seconds…

"I was fine, only blind for a few days."

Stages in your workflow



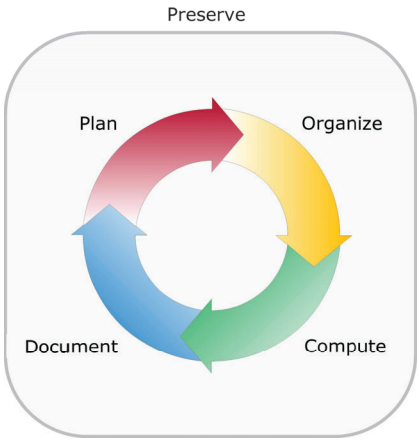
Substantive goals for your research

90% of the work unless you hurry…

Often the simplest step

Writing and revising

Tasks within each stage



Criteria for choosing your workflow

- To be effective, a workflow is explicit and planned.
- Your workflow should strive to meet these inter-related, sometimes conflicting, criteria.

Reproducibility

You must be able to reproduce your results.

Core criteria

Critically, procedures should be:

- Accurate
- Efficient
- Coordinated

Supporting criteria

Core criteria are supported by procedures that are:

- Standardized
- Automated
- Scalable

Criteria for usability

To use your workflow, it needs to be:

- Simple
- Congenial to how you work
- Transferable

Planning

The ideal

Blau and Duncan's *The American Occupational Structure*

- Analyses were specified 9 months before output was received.
- Book was written from a single set of output.

What to plan

- Project timeline
- What to publish, where and when
- Division of labor
- Procedures to document and organize research
- File naming
- Variable names, labels and metadata
- Analyses
- Preserving files

A plan is a reminder to stay on track, finish, and publish results.

Work. Finish. Publish. – Michael Faraday

Plan on different levels

- Grand plan: what is your research program
- Big plan: keys steps in project
- Middle plan: tasks within each step
- Small plan: nitty gritty details for execution

Make time to plan

- Give yourself uninterrupted time to plan -- deep work
- Turn off devices

Organizing

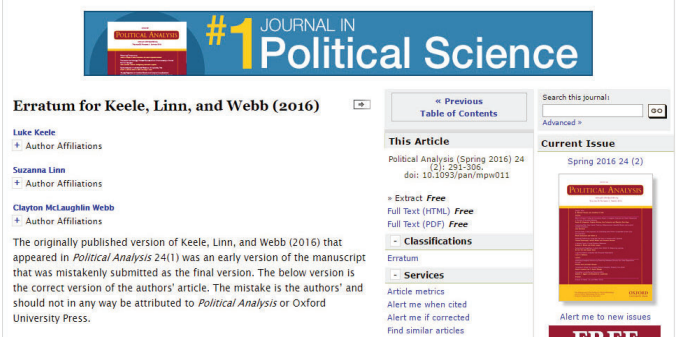
1. Organization has two goals
   - Finding things
   - Avoiding duplication
2. Organization
   - Lets you work faster
   - Rewards consistency and uniformity
   - Is contagious — so is disorganization
   - Requires *maintainance* to overcome entropy

Signs of poor organization

1. *Can't find a file* and worry you deleted it.
2. *Multiple versions of a file* and you don't know which is which.
   - You and a co-author edit different versions of a paper, leading to inconsistent drafts.
   - You need the file for draft submitted for review. Is **`FinalReportV16.docx`** the final draft?
3. *Multiple copies of the same reprint*.
4. A student at ICPSR showed me this text:
   - URGENT: don't analyze **`final.dta`**, use **`lastversion.dta`** for presentation tomorrow.

   Surely this is a rooky mistake….

The final paper



Organization: the curse of cheap storage

1. It is easier to create a file than to find a file.
2. It is easier to find a file than to know what is in a file.
3. It is easy to create lots of files.
   - I have 742,098 files on Dropbox

Files scattered across multiple locations

: Office computer	: Home computer
: Laptop	: LAN
: Dropbox	: Box
: USB sticks	: External drives
: Old laptop	: Mom's computer

Operating systems focus on entertainment

Win	**Mac**
Desktop	Desktop
Music	Music
Pictures	Pictures
Videos	Movies
Documents	Documents

Digital asset management (DAM)

How important is this?

For most people, this is a critical first step for an efficient and reproducible workflow.

How to manage files

1. Name files carefully and systematically.
2. Create a planned directory structure so that every file has only one place it belongs.

Metadata

With planned names and directories,

a file's name and location documents the file.

File naming

Writing

- groups 2017-11-07 jsl.docx : draft by JS Long on 2017-11-07
- groups 2018-01-17 sam.docx : draft by SA Mustillo on 2018-01-17

PDF reprint files all located in /Bookshelf

- Long 1978 ASR productivity position.PDF

Datasets

- groups-hrs1.dta : not final1.dta
- groups-hrs2.dta : not final2.dta

Script files

- groups-data03-recoding.do
- groups-data04-scales.do

Primary directories

\- To shelve	Files to put in the correct directory
\Active	Active projects
\Admin	Administration and service
\Bookshelf	Books, articles, reprints, etc.
\Inactive	Projects that are on hold
\Shared	Files shared with others on the cloud
\Teaching	Teaching materials
\Templates	Templates of documents and commands
\Vault	Files that will *never* change

A structure for projects in \Active, \Inactive and \Teaching

1. Each project has its own directory .
2. Each project directory has the same structure.

\Active\GroupDifferences

\- To shelve	Files to put in the right place
\Admin	Administrative documents
\Posted	Shared produces of the research
\Resources	PDF articles, codebooks, etc.
\Work	Default location for statistics programs
\Write	Documents being written.

Uniform formats for robust script files

```
capture log close
log using wftalk01-example, replace text
version 15.1
clear all
macro drop _all
set linesize 80

// project: introduction to workflow
// program: wftalk01.do
// date:    2018-08-23
// author:  Scott Long

// #1 describe task

// #2 describe task

log close
exit
```

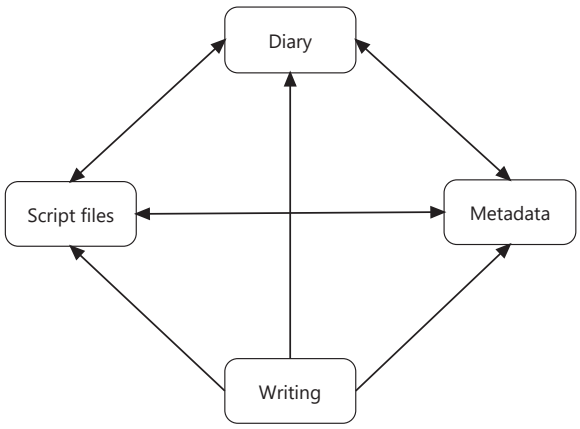

Documentation

1. Without documentation,
   - Reproduction is much more difficult.
   - Mistakes are more likely.
   - Work takes longer.
   - Revisions take much longer.

Suggestions for documentation

1. Write it today using full dates and names.
2. Check it next week. Add new and delete the irrelevant.
3. Review documentation at key stages of your work, like finishing a draft.
4. Use reinforcing, non-redundant forms of documentation.

Reinforcing forms of documentation



Execution and computing

Execution is carrying out tasks within each step.

Cornell 1975: the entire computing infrastructure

IBM 370 with 240K memory

Winchester drives with 3MB storage

Cost of computing $1,000,000

Mean time to degree 7.6 years

Laptop 2009

Laptop with 2GB memory

: 10,000 times more memory

1TB drive

: 350,000 times more

Cost of computing $400

Mean time to degree 7.6 years

A thought experiment

1. Divide graduate students randomly into two groups.

   Computers can compute any time they want.

   Planners only get to ompute 12 hours a week.

2. Who will finish their dissertation first?

Computation: Critical rules

1. Compute less but more thoughtfully.
2. Compute with a plan.
3. Spend more time on data management than statistical modeling.
4. New content gets a new name—always!
5. Shared results are never changed.

A computing workflow

This includes four components:

1. Robust and legible script files
2. Posting files
3. Dual workflow for data management and analysis
4. Run order naming of scripts

Robust and legible script files

1. Robust programs run on another computer with *no* changes.
   - To tell if a program is robust, run them on another computer.
2. Careful comments of what is being done.
   - Revisions are easier.
   - Errors are found when documenting what you are doing.
   - Others will understand what you are doing.
3. Consistent formatting for legibility.
   - Errors are easier to spot.
   - Others can more easily understand your work.

The *essential* posting principle

If you never plan to publish or present your findings, ignore this.

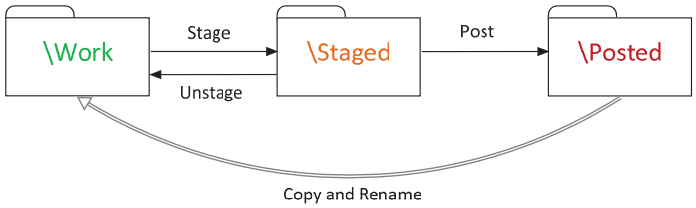
Two simple rules define posting

1. If you share results, always post the file used to produce those results.
2. If you post a file, never change that file.

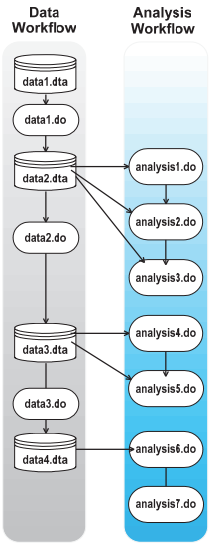
Why is it essential?

1. Posting ensures you have the files that produced your results.
2. Without posting, you might change a critical file and be unable to reproduce earlier results.
3. Posting prevents you and a collaborator from having the "same file" with different content.
4. And other, similar issues.

How posting works



Dual workflow and run order naming



Data cleaning, including names and labels

Poor variable and value labels lead to errors

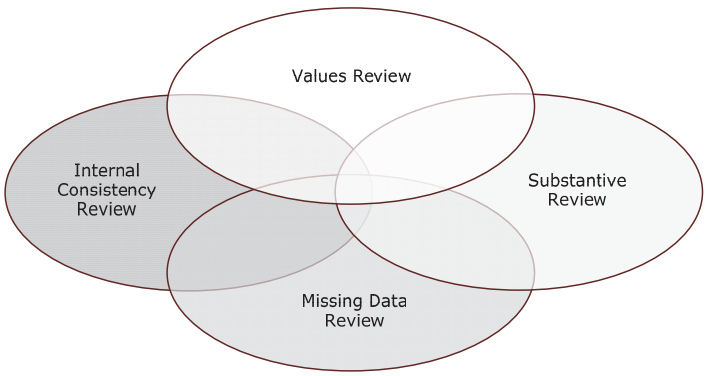
```
. codebook tc1*, compact
Variable     Obs Unique      Mean Min  Max  Label
-------------------------------------------------------------------------------
tc1doc      1074     10 8.714153   1   10  Q46 How important is it to go to ...
tc1fam      1074     10 8.755121   1   10  Q43 How important is it to turn t...
tc1friend   1073     10 7.799627   1   10  Q44 How important is it to turn t...
tc1mhprof   1045     10  7.58756   1   10  Q48 How important is it to go to ...
tc1psy      1050     10 7.567619   1   10  Q47 How important is it to go to ...
tc1relig    1039     10  5.66025   1   10  Q45 How important is it to turn t...

. tabulate female sdchild_v1

      R is |     Q15 Would let X care for children
   female? | Defintel  Probably   Probably   Definitel   |     Total
-----------+--------------------------------------------+----------
      Male |       41         99        155        197   |       492
    Female |       73         98        156        215   |       542
-----------+--------------------------------------------+----------
     Total |      114        197        311        412   |     1,034
```


Careless names

1. Confusion between **`ownsex`** and **`ownsexu`** caused weeks of delay.
2. Do you want **`R003189`** or **`R001389`**?
3. Is **`timetophd`** elapsed time or enrolled time?

Data cleaning and preventing retractions

Statistical analysis assumes the variables are clean.



A two-way table would have detected the problem

RETRACTED: In Sickness and in Health? Physical Illness as a Risk Factor for Marital Dissolution in Later Life

[redacted]

Abstract

The health consequences of marital dissolution are well known, but little work has examined the impact of health on the risk of marital dissolution. In this study we use a sample of 2,701 marriages from the Health and Retirement Study (1992–2010) to examine the role of dangerous physical illness onset (i.e., cancer, heart problems, lung disease, and/or stroke) in subsequent marital dissolution due to either divorce or widowhood. We use a series of discrete-time event history models with competing risks to estimate the impact of husband's and wife's physical illness onset on risk of divorce and widowhood. We find that only wife's illness onset is associated with elevated risk of divorce, while either husband's or wife's illness onset is associated with elevated risk of widowhood. These findings suggest the importance of health as a determinant of marital dissolution in later life via both biological and gendered social pathways.

Keywords

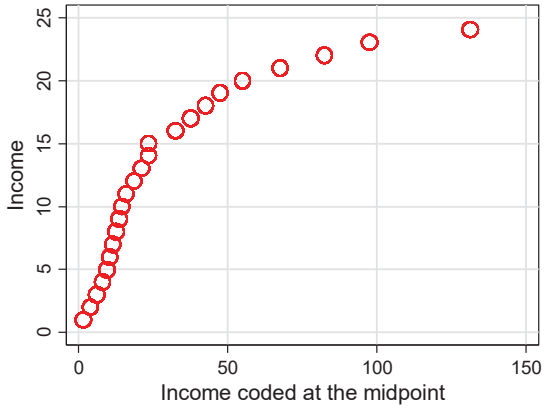
aging, chronic disease, gender, marriage, physical health

A large body of literature has identified marital status as a strong predictor of health and well-being. Not only are the married healthier than the unmarried (e.g., Lillard and Waite 1995; Umberson 1992), but studies find that both divorce and widowhood are precursors to declines in physical and mental health (e.g., Hughes and Waite 2009; Williams and Umberson 2004). Less attention, however, has been paid to how health may be a determinant of marital status. Work within this area has tended to focus on the positive selection of the healthier into marriage (e.g., Byrne et al. 1989; Smith and Smith 2010), but poor health may be an equally important force for selec-

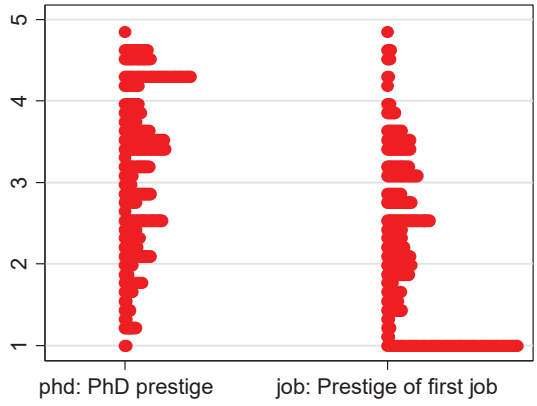
Booth, and Johnson 2008). Illness may initiate changes to spouses' roles—in particular, increasing caregiving responsibilities for the healthy spouse—which can tax marital relationship dynamics (Wolff and Kasper 2006). Illness may also decrease household income due to the inability of one or both spouses to work (Teachman 2010), which may increase marital strain.

Only a few studies have examined the role of poor health in subsequent divorce, and these studies are mixed in their findings, with some finding

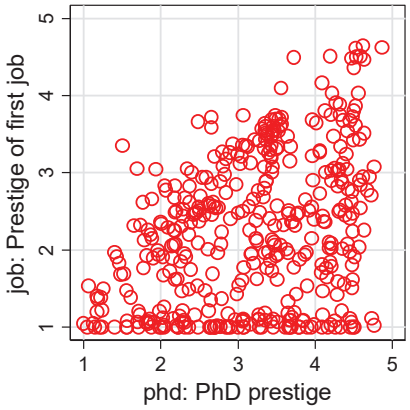
Use graphs to find errors



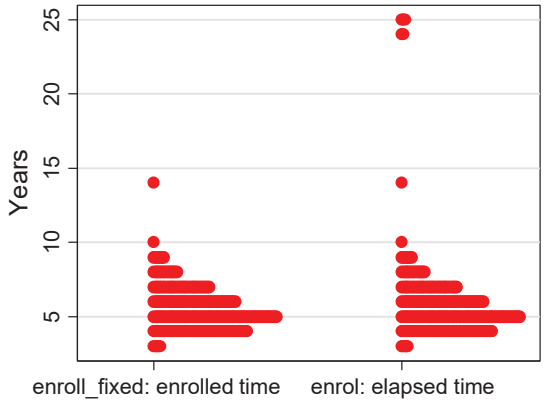
Graphs highlight forgotten coding decisions



Locating outliers and gaining substantive insights



Avoiding expensive mistakes from misread documentation



Statistical analysis

This can be the simplest part of the project.

1. Plan the analysis.
2. *Find exemplars* in the best journals.
3. *Use automation* and script files.
4. Maintain a *dual workflow* to prevent errors.

Papers and provenance

1. The *provenance of a result* is the script and dataset that produced it.
2. *Maintaining provenance* is critical for reproducibility.
   - If you don't know where a number came from, how do you reproduce it?
3. *Revisions are much easier* since you know exactly what needs to be changed.

Documenting provenance in a paper

1. The circled text contains results I may need to confirm later:

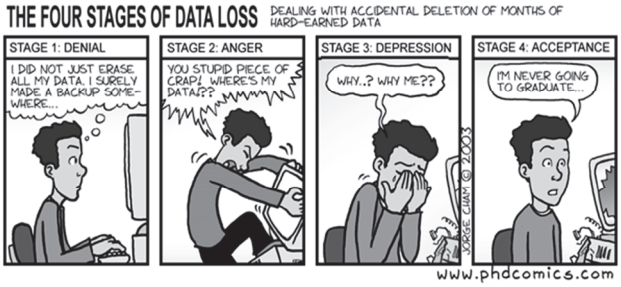
> 1922-1926 cohort, employed women have fewer limitations than those who are out for family reasons, (.48 and .73, respectively (z=2.55, p<.01).) However, this gap has disappeared for the 1943-1947 cohort and, indeed, employed women have slightly more limitations (.76 for non-

2. Turning on "show/hide ¶" reveals the provenance:

> 1922-1926 cohort, employed women have fewer limitations than those who are out for family reasons, (.48 and .73, respectively (z=2.55, p<.01 {cwhrr-fig03c-hrmemp4.do #4 jsl 17May06}).)
>
> However, this gap has disappeared for the 1943-1947 cohort and, indeed, employed women have

Preserving your files

- Expect things to go wrong, expect to delete the wrong file at the worst possible time, and expect a hose to be left on in the room above your computer, expect your partner to delete your files.
- If you expect the worst, you might prevent it.



Preserving files does *not* preserve content

Migrate formats as software changes

> "These files were saved six years ago as Gauss FMT files. We need to revise a paper and need the data in these files, but I can't open them. We have an old version of Gauss that doesn't run anymore. Any ideas?"

Media need to be migrated

Having the media doesn't mean you can read it

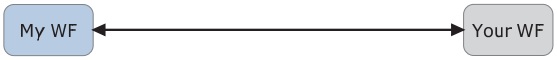
- Old tapes are costly to read
- Zip drives disappeared quickly

Collaboration and workflow

1. Collaboration makes it harder to have an effective workflow.
   - Why can't *they* be just like *me*?
2. Collaboration makes your workflow important.
   - Because your collaborators are not just like you.

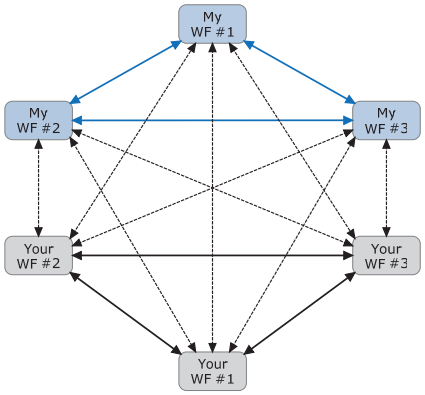
Why is workflow harder when you collaborate?

1. Assume one collaborator.
2. Ideally, you must coordinate two workflows:

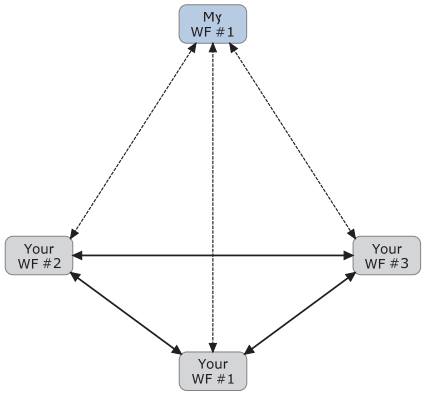


3. What if neither of you has a consistent workflow?

Two collaborators with ineffective workflows



Even if your collaborator doesn't cooperate



Conclusions

Changing your workflow

- Plan changex by assessing the greatest problems in your workflow
- Make changes slowly, systematically, thoughtfully.
- Finish the last 5% of each change.
- New tools are not as valuable as tools you have mastered.
- Do not make changes under a deadline

Your workflow or my workflow

- Being here improved your workflow.
- You don't need to use my workflow.
- You should evaluate how issues I raise are addressed by your workflow.

Resources

- The Workflow of Data Analysis Using Stata
- ICSPR Summer Program Workshop
- BITSS: Berkeley Initiative for Transparency in the Social Sciences
- Project TIER: Teaching Integrity in Emprical Research

Thank you