Introduction to Reproducible Research

Jonathan Gilligan

January 23, 2018

Prefatory Note

This document is adapted from an introduction to reproducible research that I wrote for the laboratory section of my Global Climate Change class, and thus it has a lot of emphasis on climate science.

Why Do We Need Reproducible Research?

In the spring of 2012, Bruno Iksil, a securities trader at the investment bank JPMorgan, Chase, & Company who was also known by the nickname "The London Whale" for his aggressive trades, made a series of costly mistakes that cost JPM-Chase $6.2 billion. (Hurtado 2015) Iksil was attempting to manage the financial risk of a portfolio of investments. However, an analyst on Iksil's team had calculated the volatility (a measure of financial riskiness) of his portfolio using an Excel spreadsheet and made a subtle error in a mathematical formula, dividing by the sum of two numbers instead of the average. That error caused Iksill to underestimate risk by a factor of 2, and thus to expose JPM-Chase to far more risk than he realized with his enormous trades. (Kwak 2013)

Two years earlier, Carmen Reinhart and Kenneth Rogoff, two highly respected academic economists published an influential research paper on the effect of government debt on economic growth. (Reinhart and Rogoff 2010) This paper concluded that when government debt exceeds 90% of GDP, the country's economic growth is likely to abruptly come to a halt, and even slide into recession. This paper was used to justify harsh austerity measures throughout Europe, where nations were struggling to recover from the 2008 global economic meltdown, and was cited by Paul Ryan as justification for his proposals to dramatically cut federal spending in the U.S.

Thomas Herndon, a graduate student in economics at the University of Massachusetts, was skeptical about this research, but Reinhart and Rogoff's paper did not explain all the details of their data and analysis. Finally, in 2013, Reinhart and Rogoff gave Herndon copies of the spreadsheets they had used in their analysis. Herndon found three glaring errors in the spreadsheet, and after he fixed the errors, there was no sudden slowdown of economic growth. (Krudy 2013; Bailey and Borwein 2013; Kwak 2013)

Such errors are not unique to economic research. In the past two years, errors in spreadsheet formulas led to the retraction of papers in prominent journals of environmental science, medicine, and biology. (Stern 2017; Palus 2016; Ferguson 2015)

In 2015, a major study of air pollution associated with "fracking" in natural gas wells was retracted when the authors discovered a major error in an Excel spreadsheet that they used for data analysis.

(Chawla 2016)

Spreadsheets are not the only source of major errors in scientific publications. Poor statistical practices have led to what has come to be called a "crisis of replication" in psychology and medicine and concern that many published scientific results are incorrect.

Of particular concern is the fact that major problems are being discovered in clinical medical research. Once recent review of 5,000 papers in eight top medical journals found that almost 100 had major inaccuracies.

Scientific Errors in Climate Science

Scientific errors have had significant impact in climate science In the 1990s, John Christy and Roy Spencer, a pair of prominent climate scientists at the University of Alabama at Huntsville (UAH), were analyzing satellite measurements of microwave emissions from the earth's atmosphere and using them to calculate the temperature of different layers of the atmosphere. They reported the surprising result that whereas measurements of temperatures at the earth's surface, made with thermometers at meteorological monitoring stations, consistently showed a large warming trend, the satellite measurements found that the lower troposphere was cooling off. Christy and Spencer claimed that their satellite measurements were more accurate than the thermometer measurements taken at the surface and challenged scientific findings that global warming was taking place. Controversy about the disagreement between the satellite measurements and the surface measurements raged for years, and was the subject of a book-length report from the National Academy of Sciences. (National Research Council 2000)

A rival team of scientists at the company Remote Sensing Systems (RSS) looked into the matter and conducted its own analysis of the satellite temperature records. Christy and Spencer would not release the computer code they used to analyze the satellite record to the public, so other scientists could not check it for errors. However, the independent analysis by scientists at RSS revealed a number of serious errors in the UAH analysis, including a place where the UAH team mistakenly added two numbers instead of subtracting one from the other (an easy mistake to make in programming, and one that might have been caught years earlier if the code had been available for public inspection). (Wentz and Schabel 1998; Mears and Wentz 2005; Christy et al. 2007) In the end, the UAH team released corrections to their satellite temperature measurements, and the corrected record agreed well with other measurements, including those by thermometers at the surface and weather balloons. The years of controversy over whether the lower troposphere was cooling turned out to be mostly the result of computer programming errors.

The Move toward Open Climate Science

In 2009, a leak of thousands of emails by a number of climate scientists led to a scandal known as "Climategate," in which prominent climate scientists were accused of doctoring their data and analysis. Investigations conclusively cleared them of any misdeeds, and it turned out that the things they were accused of doing in secret had in fact been clearly reported in published papers years

before the scandal. However, the damage to the reputation of the scientists and the public's trust in climate science taught the climate science community the importance of being completely open with data, methods, and computer code.

Since then, climate scientists have moved significantly toward adopting principles of openness. Today, pretty much all major climate data sets are available for free on the internet. Computer code used for important analysis, including the source code to many of the major global climate models, is publicly available (although much of it is not much use unless you have a supercomputer to run it on).

Making all the data and code available helps win trust by convincing the public that climate scientists do not have anything to hide. It also facilitates faster scientific progress by allowing scientists more easily to build on one another's work, and it makes it easier to find and correct errors. A good list of major sources of climate data and computer code is available at http://www.realclimate.org/index.php/data-sources/. The R Open Science Project (https://ropensci.org/) maintains a number of sophisticated open-source scientific projects that cover many fields, such as biology and climate science. A retired engineer and amateur climate scientist, D. Kelly O'Day, maintains a blog (https://rclimate.wordpress.com/) where he shares R scripts to download, analyze, and graph climate data.

The Big Picture

For the most part, science works. Advances in all fields of science have led to deep understanding of nature, and have led to technological breakthroughs that drive our economy, enable us to live much longer and healthier lives, and otherwise improve the quality of our lives.

Nonetheless, even if only a few percent of major scientific research papers are wrong, this has potential to mislead us about which medicines or medical procedures are safe and which are dangerous, about which government policies are likely to be effective, and in the private sector, can lead companies to make financially disastrous mistakes.

Reproducible Research

Two important principles in science, which should prevent these errors, are that research should be *transparent* and *reproducible*: Research reports should describe the procedures clearly and in enough detail that other scientists know exactly what was done. And scientists who repeat the research procedures, as described in the reports, should find similar results, within the limits of experimental uncertainty.

However, as the anecdotes above, and hundreds of similar reports of problems in research reveal, too often even well-meaning scientists fall short of providing enough detail about their methods for other scientists to understand their work and catch errors, and it is often difficult to truly reproduce previously published research.

To address these problems, the scientific community is increasingly embracing the principles of what has come to be called **reproducible research**.

Federal funding agencies, scientific journals, and scientific societies now call for authors to reveal all the details of their experiments and analysis, and must share the data and computer codes they used to perform the analyses described in their publications.

Whether you are doing research in basic science, such as quantum physics, conducting clinical trials to assess the effectiveness and safety of new drugs and medical procedures, investigating climate change, analyzing economic policy, or working for a private company to study financial risks and opportunities, it will be important for you to be able to do your research accurately, to communicate the details clearly with your co-workers and your bosses, and to be able to return to your old research reports and vouch for all the details of what you did.

Whether you find yourself working in pure academic research, in public policy, or for private industry, the tools of reproducible research will help you do these things effectively.

What is Reproducible Research?

Reproducible research seeks to make scientific research completely reproducible by documenting every decision a researcher made in the course of collecting and analyzing data. At the simplest level, this would mean that when a scientist submits a paper to a research journal, she would include all the data and a clear description of the analysis.

However, a written description of the analysis process might inadvertently omit crucial steps, or the researcher might describe what she thought she did, but might have made errors in her actual analysis.

In the example of Reinhart and Rogoff's paper on debt and economic growth, the two economists described what they thought they had done, but they were unaware that their spreadsheet contained errors. For three years after they published the paper, the errors remained buried in their spreadsheets but other economists only knew the written descriptions of the analysis that appeared in their paper and could not examine the spreadsheet for themselves.

Thus, reproducible research calls for researchers to share not only their data, but also any spreadsheets, computer programs, or scripts they used to perform the analysis. This will allow other researchers to catch errors where the actual analysis procedure does not match the description in the published report, just as Thomas Herndon was able to do when he obtained Reinhart and Rogoff's spreadsheets.

Scripts versus Spreadsheets

In principle, this should suffice, but in practice it turns out that auditing a spreadsheet is very difficult. When you open a spreadsheet in Excel, you see a grid of text and numbers, but the formulas used in calculating the values of certain cells from other cells are largely invisible and it is difficult to read and audit every formula in a spreadsheet that contains thousands of cells.

Thus, the scientific community has become increasingly mistrustful of spreadsheets and prefers data analysis tools that use scripts to conduct the analysis. Scripts (basically, short computer programs) are written in a textual form that is straightforward for a knowledgeable person to read and understand. Consistency checks to catch errors are much easier to implement in scripts than in spreadsheets.

From the Analysis to the Manuscript

Even when analysis is performed correctly, it can be difficult to transcribe every number correctly into the manuscript of a research report. When I (Professor Gilligan) was in graduate school, one of my professors told me a cautionary tale from early in his career: Three prominent physicists were attempting a very difficult calculation in quantum electrodynamics. To guard against errors, each of the three performed the calculation independently and then they compared their results. They were very excited to discover that their calculations agreed perfectly. One of them went to add up the different terms from the calculations and type the result into the manuscript of a paper they rushed into print to announce and share their accomplishment. Somewhere in the process, he transcribed a number incorrectly, so the published result was incorrect and the embarrassing error was not discovered until some time after it appeared in print.

Consider, too, what happens if a research report is almost complete and the researchers discover an error in their analysis scripts or in their raw data. After they make the correction, they must adjust every number in their final manuscript. This introduces additional risks of either mistyping numbers or of missing a number that needs to be changed.

With modern computing tools, it has become easy to integrate the analysis with the final report.

We won't get into it in this course, but there is a powerful tool called RMarkdown, which allows you to combine the text of your documents with scripts that perform data analysis and generate the figures and tables. Thus, any time you change a number in your data or change a line of code in your analysis script, the computer can automatically regenerate your document to update all of the numbers and figures accordingly. This approach is widely used both in scientific research and in business, where it can easily automate the generation of monthly or quarterly reports from data.

Elements of Reproducible Research

This section is adapted from the "Introduction to Reproducible Research" by the R Open Science Project, http://ropensci.github.io/reproducibility-guide/sections/introduction/

Kinds of Reproducibility

Reproducibility means different things in different scientific fields. One big distinction is between computational versus observational or empirical aspects of research.

- Computational reproducibility provides detailed information on exactly how computations (either calculations or simulations using computer models) were performed, and making it possible for others to exactly reproduce those computations or calculations. This includes providing source code for programs and scripts written by the researcher, together with detailed specifications of the software (including the specific versions used), and the hardware used (running the same software on different computer hardware can sometimes give different results).
- Empirical reproducibility provides detailed information about laboratory or field procedures that were used to acquire empirical data used in the analysis. In practice, this is often accomplished by providing the raw data together with details about how it was collected.

In this document, I will focus on *computational reproducibility*. I will not address all the details of computational reproducibility, but I will focus on three important aspects:

- **Literate computing and authoring:** Literate computing refers to mixing computer code with narrative description of what the code is doing. Stanford computer science professor Donald Knuth invented literate programming based on his experience with large software projects, where he found that if he wrote clear narrative descriptions of what his programming code was doing, he made fewer errors, and could find and correct those errors more quickly.
- **Automation:** Many of you may be used to so-called "point-and-click" software tools for statistical analysis. Examples include Excel, SPSS, and Stata. These tools let you perform analysis by reading in your data and then highlighting data with a mouse and selecting menu options. This approach makes it easy to get started with these software packages when you are a beginner, but make it difficult to track exactly what you did in your analysis, so when it comes time to write up your report, you may not remember exactly what you did, and in what order you did it. Automating your analysis using scripts means that your script contains the complete information about everything you did.

  Some programs, such as Stata, allow you to record your analysis and export a script (a `.do` file) that will allow you or others to reproduce your analysis. This is a valid form of reproducible research, but it is not the one we will use in this course.

  Automation is also important if you have to do the same operation on many different data sets. What seemed easy when you just had to create one graph or analysis can quickly become tedious as you have to drag the mouse and click on the same menu entries over and over again on a dozen different data sets. Automating your analysis with scripts makes it easy to run the same script on each of your different data sets.
- **Revision Control:** As you edit both your text and the R scripts you use for your analysis, it is valuable to be able to keep track of changes. For instance, if your analysis is working well, and then you edit something and it stops working it is useful to be able to go back and look at what changed between the time when it was working and when it stopped working.

  Revision control systems allow you to easily keep track of changes you make to your files.

  Another important use of revision control is not relevant to this course, but applies to larger research projects. I have computational models that are constantly under development and I publish papers based on them. Suppose that another scientist has a question about a paper

that I wrote two years ago, but I have made many changes to the model since then. How can I go back and recreate the version of the model that I used for that paper? Revision control systems make this very easy.

Finally, revision control systems are very useful for team projects because they allow a team of many researchers to coordinate their activities when they are all editing files (computer code and text) for a project at the same time.

We will be using a revision control tool called `git`. There is a web site called github.com, which allows people to share projects. My students and I use github to release software that we develop in our research, and there is an educational site connected to github, which we will use for computational assignments in this course.

Everyone should sign up for a free student account on github.

Walking the Walk

Over the last five years I have become increasingly convinced that reproducible research methods are both more efficient and also lead to higher quality research. I use these methods extensively in my own research.

Further Reading

If you are interested in learning more about reproducible research, I would recommend the following:

- Christopher Gandrud, *Reproducible Research with R and RStudio* (Second Edition) (CRC Press/Chapman & Hall, 2015). Gandrud is an economist and political scientist, who pioneered a lot of the methods that I use for reproducible research as part of his Ph.D. dissertation. This book is a comprehensive how-to guide to reproducible research, and all of the files necessary to reproduce the book are available online at https://github.com/christophergandrud/Rep-Res-Book
- The R Open Science Project, *Reproducibility in Science: A Guide to Enhancing Reproducibility in Scientific Results and Writing* http://ropensci.github.io/reproducibility-guide/

References

Bailey, David H., and Jonathan Borwein. 2013. "The Reinhart-Rogoff Error—or How Not to Excel at Economics. The Conversation." *The Conversation*, April. http://theconversation.com/the-reinhart-rogoff-error-or-how-not-to-excel-at-economics-13646.

Chawla, Dalmeet Singh. 2016. "Authors Retract Study That Found Pollution Near Fracking Sites. Retraction Watch." July 8. http://retractionwatch.com/2016/07/08/authors-retract-study-that-found-pollution-near-fracking-sites/.

Christy, John R., William B. Norris, Roy W. Spencer, and Justin J. Hnilo. 2007. "Tropospheric Temperature Change Since 1979 from Tropical Radiosonde and Satellite Measurements." *Journal of Geophysical Research: Atmospheres* 112 (D6): D06102. doi:10.1029/2005JD006881.

Ferguson, Cat. 2015. "Teflon Toxicity Paper Fails to Stick. Retraction Watch." April 16. http://retractionwatch.com/2015/04/16/teflon-toxicity-paper-fails-to-stick/.

Hurtado, Patricia. 2015. "The London Whale." *Bloomberg View*, April. https://www.bloomberg.com/view/quicktake/the-london-whale.

Krudy, Edward. 2013. "How a Student Took on Eminent Economists on Debt Issue—and Won." *Reuters*, April. https://www.reuters.com/article/us-global-economy-debt-herndon-idUSBRE93H0CV20130418.

Kwak, James. 2013. "The Importance of Excel. The Baseline Scenario." February 9. https://baselinescenario.com/2013/02/09/the-importance-of-excel/.

Mears, Carl A., and Frank J. Wentz. 2005. "The Effect of Diurnal Correction on Satellite-Derived Lower Tropospheric Temperature." *Science* 309 (5740): 1548–51. doi:10.1126/science.1114772.

National Research Council. 2000. *Reconciling Observations of Global Temperature Change*. https://www.nap.edu/catalog/9755/reconciling-observations-of-global-temperature-change.

Palus, Shannon. 2016. "Doing the Right Thing: Authors Share Data, Retract When Colleague Finds Error. Retraction Watch." May 17. http://retractionwatch.com/2016/05/17/doing-the-right-thing-authors-share-data-retract-when-colleague-finds-error/.

Reinhart, Carmen M., and Kenneth S. Rogoff. 2010. "Growth in a Time of Debt." *American Economic Review* 100 (2): 573–78. doi:10.1257/aer.100.2.573.

Stern, Victoria. 2017. "Think of the Unthinkable: JAMA Retraction Prompts Author to Urge Others to Share Data. Retraction Watch." April 19. http://retractionwatch.com/2017/04/19/think-unthinkable-jama-retraction-prompts-author-urge-others-share-data/.

Wentz, Frank J., and Matthias Schabel. 1998. "Effects of Orbital Decay on Satellite-Derived Lower-Tropospheric Temperature Trends." *Nature* 394 (6694): 661–64. doi:10.1038/29267.