Good practices: coding practices, debugging, and reproducible research

Chris Paciorek

2022-08-29

Table of contents

PDF

Sources:

- Chambers
- Hadley Wickham's advanced R notes on debugging.
- Roger Peng's notes on debugging in R
- Murrell, Introduction to Data Technologies, Ch. 2
- Journal of Statistical Software vol. 42: 19 Ways of Looking at Statistical Software
- Wilson et at., Best practices for scientific computing, ArXiv:1210:0530
- Gentzkow and Shapiro tutorial for social scientists
- Millman and Perez article about reproducible research
- Chapter 11 of Transparent and Reproducible Social Science Research

This unit covers good coding/software development practices, debugging (and practices for avoiding bugs), and doing reproducible research. As in later units of the course, the material is generally not specific to R, but some details and the examples are in R.

1. Good coding practices

Some of these tips apply more to software development and some more to analyses done for specific projects; hopefully it will be clear in most cases.

Editors

Use an editor that supports the language you are using (e.g., *Atom*, *Emacs/Aquamacs*, *Sublime*, *vim*, *VSCode*, *TextMate*, *WinEdt*, or the built-in editor in *RStudio*). Some advantages of this can include: (1) helpful color coding of different types of syntax and of strings, (2) automatic indentation and spacing, (3) code can often be run or compiled from within the editor, (4) parenthesis matching, (5) line numbering (good for finding bugs).

Coding syntax

The files goodCode.R and badCode.R in the *units* directory of the class repository provide examples of code written such that it does and does not conform to the suggestions listed in this section.

Here are some R-related style guides:

- Adler has style tips.
- The tidyverse style guide or its offshoot Google R style.
- A empirical style guide based on the code of R Core and key package developers.
- This R journal article summarizes the state of naming styles on CRAN.

And here's a summary of my own thoughts:

- Header information: put metainfo on the code into the first few lines of the file as comments. Include who, when, what, how the code fits within a larger program (if appropriate), possibly the versions of R and key packages that you used.
- Indentation: do this systematically (your editor can help here). This helps you and others to read and understand the code and can help in detecting errors in your code because it can expose lack of symmetry.
- Whitespace: use a lot of it. Some places where it is good to have it are
  - around operators (assignment and arithmetic);
  - between function arguments;
  - between list elements; and
  - between matrix/array indices, in particular for missing indices.
- Use blank lines to separate blocks of code and comments to say what the block does
- Split long lines at meaningful places.
- Use parentheses for clarity even if not needed for order of operations. For example, `a/y*x` will work but is not easy to read and you can easily induce a bug if you forget the order of ops.
- Documentation - add lots of comments (but don't belabor the obvious). Remember that in a few months, you may not follow your own code any better than a stranger. Some key things to document: (1) summarizing a block of code, (2) explaining a very complicated piece of code - recall our complicated regular expressions, (3) explaining arbitrary constant values.
- For software development, break code into separate files (2000-3000 lines per file) with meaningful file names and related functions grouped within a file.

- Choose a consistent naming style for objects and functions: e.g. *nIts* vs. *n.its* vs *numberOfIts* vs. *n_its*
  - This R journal article summarizes the state of naming styles on CRAN.
  - In object-oriented languages such as Python and Java, periods are used in the context of object-oriented programming, so I recommend not using periods in the names of your objects.
- Try to have the names be informative without being overly long.
- Don't overwrite names of objects/functions that already exist in R. E.g., don't use 'lm'. That said, the namespace system helps with the unavoidable cases where there are name conflicts.
- Use active names for functions (e.g., *calcLogLik*, *calc_logLik* rather than *logLik* or *logLikCalc*). The idea is that a function in a programming language is like a verb in regular language (a function *does* something), so use a verb in naming it.
- Learn from others' code

This semester, someone will be reading your code - the GSI and and me when we look at your assignments. So to help us in understanding your code and develop good habits, put these ideas into practice in your assignments.

Coding style

This is particularly focused on software development, but some of the ideas are useful for data analysis as well.

- Break down tasks into core units
- Write reusable code for core functionality and keep a single copy of the code (w/ backups of course or ideally version control) so you only need to make changes to a piece of code in one place
- Smaller functions are easier to debug, easier to understand, and can be combined in a modular fashion (like the UNIX utilities)
- Write functions that take data as an argument and not lines of code that operate on specific data objects. Why? Functions allow us to reuse blocks of code easily for later use and for recreating an analysis (reproducible research). It's more transparent than sourcing a file of code because the inputs and outputs are specified formally, so you don't have to read through the code to figure out what it does.
- Functions should:
  - be modular (having a single task);
  - have meaningful name; and
  - have a comment describing their purpose, inputs and outputs (see the help file for an R function of your choice for how this is done in that context).
- Write tests for each function (i.e., unit tests)
- Don't hard code numbers - use variables (e.g., number of iterations, parameter values in simulations), even if you don't expect to change the value, as this makes the code more readable. For example, the speed of light is a constant in a scientific sense, but best to make it a variable in code: `speedOfLight <- 3e8`
- Use R lists to keep disparate parts of related data together
- Practice defensive programming (see also the discussion below on assertions)
  - check function inputs and warn users if the code will do something they might not expect

    or makes particular choices;
  - check inputs to *if* and the ranges in *for* loops:
    * use `seq_len()` and `seq_along()` instead of `1:n` in setting up for loops
    * use `if(isTRUE(condition))` in if statements in case the condition is NA (which would otherwise cause an error)
  - provide reasonable default arguments;
  - document the range of valid inputs;
  - check that the output produced is valid; and
  - stop execution based on checks and give an informative error message.
- Try to avoid system-dependent code that only runs on a specific version of an OS or specific OS
- Learn from others' code
- Consider rewriting your code once you know all the settings and conditions; often analyses and projects meander as we do our work and the initial plan for the code no longer makes sense and the code is no longer designed specifically for the job being done.

Assertions and testing

Both tests and assertions are critically important for writing robust code that is less likely to contain bugs.

Assertions are checks in your code that the state of the program is as you expect, including arguments provided by users. In addition to simple use of *stopifnot()* and using *if()* combined with *stop()* and *warning()*, the *assertthat* and *assertr* packages provide useful tools. The *checkmate* package specifically helps in checking arguments.

Tests evaluate whether your code operates correctly. This can include tests that your code provides correct and useful errors when something goes wrong (so that means that a test might be to see if problematic input correctly produces an error). *Unit tests* are intended to test the behavior of small pieces (units) of code, generally individual functions. Unit tests naturally work well with the ideas above of writing small, modular functions. *testthat* and other packages are designed to make it easier to write sets of good tests.

In Lab 2, we'll go over assertions and testing in detail.

Version control

- Use it! Even for projects that only you are working on.
- Use an issues tracker (e.g., the GitHub issues tracker is quite nice), or at least a simple to-do file, noting changes you'd like to make in the future.
- In addition to good commit messages, it's a good idea to keep good running notes documenting your projects.

We've already seen Git some and will see it in a lot more detail later in the semester, so I don't have more to say here.

2. Debugging and recommendations for avoiding bugs

The R debugging tutorial has information on

- basic debugging strategies,
- using R's interactive debugging tools,
- common causes of bugs,
- tips and tools for avoiding bugs and catching errors, and
- information on how to get help online.

In a future Lab, we'll go over debugging in detail, so you don't need to look through the R debugging tutorial at this time.

3. Tips for running analyses

Save your output at intermediate steps (including the random seed state) so you can restart if an error occurs or a computer fails. Using *save()* and *save.image()* to write to *.Rda* (*.RData*) files work well for this.

Run your code on a small subset of the problem before setting off a job that runs for hours or days. Make sure that the code works on the small subset and saves what you need properly at the end.

4. Reproducible research

The idea of "reproducible research" has gained a lot of attention in the last decade because of the increasing complexity of research projects, lack of details in the published literature, failures in being able to replicate or reproduce others' work, fraudulent research, and for other reasons.

We've seen a number of tools that can help with doing reproducible research, including version control systems such as git, the use of scripting such as bash and R scripts, and literate programming tools such as knitr and R Markdown.

Provenance is becoming increasingly important in science. It basically means being able to trace the steps of an analysis back to its origins. *Reproducibility* and *replicability* are related concepts:

Reproducibility - the idea is that a second person/group could get the exact same results as an existing analysis if they use the same input data, methods, and code. This can be surprisingly hard as time passes even if you're the one attempting to reproduce things.

Replicability - the idea is that a second/person could obtain results consistent with an existing analysis when using new data to answer the same scientific question.

Open question: What is required for something to be reproducible? What about replicable? What are the challenges in doing so?

Some basic strategies

- Have a directory for each project with subdirectories with meaningful and standardized names: e.g., *code*, *data*, *paper*. The Journal of the American Statistical Association (JASA) has a template GitHub repository with some suggestions.
- Have a file of code for pre-processing, one or more for analysis, and one for figure/table preparation.

  - The pre-processing may involve time-consuming steps. Save the output of the pre-processing as a file that can be read in to the analysis script.
  - You may want to name your files something like this, so there is an obvious ordering: "1-prep.R", "2-analysis.R", "3-figs.R".
  - Have the code file for the figures produce the **exact** manuscript/report figures, operating on a file (e.g., .Rda file) that contains all the objects necessary to run the figure-producing code; the code producing the .Rda file should be in your analysis code file (or somewhere else sensible).
  - Alternatively, use *knitr*, *R Markdown,* or *Jupyter notebooks* for your document preparation.
- Keep a document describing your running analysis with dates in a text file (i.e., a lab book).
- Note where data were obtained (and when, which can be helpful when publishing) and pre-processing steps in the lab book. Have data version numbers with a file describing the changes and dates (or in lab book). If possible, have all changes to data represented as code that processes the data relative to a fixed baseline dataset.
- Note what code files do what in the lab book.
- Keep track of the details of the system and software you are running your code under, e.g., operating system version, software (e.g., R, Python) versions, R or Python package versions, etc.
  - In R, *sessionInfo()* will report all this for you.

Formal tools

1. In some cases you may be able to carry out your complete workflow in a knitr/R Markdown document or in a Jupyter notebook.
2. You might consider workflow/pipeline management software such as Drake or other tools discussed in the CRAN Reproducible Research Task View. Alternatively, one can use the *make* tool, which is generally used for compiling code, as a tool for reproducible research: if interested, see the tutorial on Using make for workflows or this Journal of Statistical Software article for more details.
3. You might organize your workflow as an R package as described in this article.
4. Package management:
   - R: You can manage the versions of R packages (and dependent packages) used in your project using package management packages such as *renv*, *checkpoint*, and *packrat*.
   - Python: You can manage the versions of Python packages (and dependent packages) used in your project using Conda environments (or virtualenvs).
5. If your project uses multiple pieces of software (e.g., not just R or Python), you can set up a reproducible environment using *containers*, of which Docker containers are the best known. These provide something that is like a lightweight virtual machine in which you can install exactly the software (and versions) you want and then share with others. Docker container images are a key building block of various tools such as GitHub Actions and the Binder project
6. You can manage the configuration of a project using the *config* package. This allows you to set configuration values that control how your code/workflow runs, thereby enabling you to run the workflow under different scenarios (e.g., different input datasets, different models, different model configurations, etc.).