Programming Best Practices in R

Eugene Katsevich

2024-08-21

Contents

When programming, one often wishes to get to the answer as quickly as possible. However, programming quickly jeopardizes the correctness and reproducibility of the code that is written. This document lays out a number of programming best practices to promote the correctness and reproducibility of code, which may take additional time to master at first but which save time in the long run. This document is geared towards R programming, though many of the principles presented here are language-agnostic.

1 Code readability

Code readability refers to how easily a programmer can understand and interpret a piece of code. It is an essential aspect of software development because readable code is easier to maintain, faster to debug, more collaborative, and less prone to errors. Code readability can be broadly categorized into two main aspects: code style and code transparency. *Code style* pertains primarily to the format and presentation of the code. *Code transparency* delves deeper into the logic and structure of the code. It refers to how straightforwardly the functionality, logic, or operations are conveyed.

1.1 Example

Consider the following two examples:

Poor code readability:

```
a = sum(dat[dat[,1]>5 & dat[,2]<10,3])
b = mean(dat[dat[,1]>5 & dat[,2]<10,3])
print(paste('Sum =',a,', Mean =',b))
```

Good code readability:

```r
# Library for data manipulation
library(dplyr)

# Analysis parameters
MIN_AGE <- 5
MAX_SCORE <- 10

# Extract summary performance metrics for subset of observations
summary_data <- dat |>
  filter(age >= MIN_AGE & score <= MAX_SCORE) |>
  summarize(
    sum_performance = sum(performance_metric),
    mean_performance = mean(performance_metric)
  )

# Print the computed results
cat(sprintf("Sum of Performance: %f | Mean of Performance: %f",
            summary_data$sum_performance,
            summary_data$mean_performance))
```

In this section, we will discuss how to write code like that in the second snippet above.

1.2 Code style

Code should adhere to the tidyverse style guide, which includes guidance on the following aspects of code:

- Naming conventions
- Spacing and indentation
- Commenting

You can use the `styler` package to automatically conform your spacing and indentation to the tidyverse style guide.

1.3 Code transparency

To write transparent code, follow these guidelines:

- Use tidyverse paradigms as much as possible (e.g. `dplyr` summaries instead of apply operations)
- Use names, rather than indices, for subsetting (e.g. `results["mse", "lasso"]` versus `results[2,4]`)
- Use named arguments in function calls, especially with more than one argument (e.g. `rbinom(3, 1, 0.5)` versus `rbinom(n = 3, size = 1, prob = 0.5)`)
- Put logically related chunks of code together into code blocks, with a comment describing the thrust of that code block.

2 Automation and reproducibility

To ensure the robustness and reliability of your analyses, strive for automation and reproducibility. This approach ensures your work remains consistent and easily repeatable.

2.1 Automate manual operations

Manual operations, including moving files, creating directories, and saving figures, are inefficient and error-prone. As many operations as possible should be automated instead.

```r
# Instead of manually downloading a file, use R to do it programmatically:
download.file(url = "https://example.com/data.csv", destfile = "data/data.csv")
```

```r
# Instead of manually saving figures, use ggsave():
ggsave("plots/my_plot.png", plot = p)
```

2.2 Embrace modularity

Avoid repetitive code. Repetition not only lengthens your script but also increases the chance for mistakes.

```r
# Bad practice: Repetitive code
mean1 <- mean(df1$weight, na.rm = TRUE)
mean2 <- mean(df2$weight, na.rm = TRUE)
mean3 <- mean(df3$weight, na.rm = TRUE)

# Good practice: Create a function
calculate_mean <- function(df) {
  mean(df$weight, na.rm = TRUE)
}

mean1 <- calculate_mean(df1, "weight")
mean2 <- calculate_mean(df2, "weight")
mean3 <- calculate_mean(df3, "weight")

# Even better: Reduce repetition further via sapply
means <- sapply(list(df1, df2, df3), calculate_mean)
```

2.3 Name constants in your scripts

"Magic numbers" are unexplained numbers in your scripts:

```r
# Bad practice: Magic number
if (age > 30) ...

# Good practice: Using a named constant
MAX_AGE <- 30
if (age > MAX_AGE) ...
```

Descriptive constant names provide clarity. Furthermore, especially if these constants are used in multiple places throughout your script, updating them becomes as simple as changing one line of code. It is also advisable to put all such constants together, near the top of the script.

2.4 Produce results by running entire scripts

All results should be produced by writing scripts and then executing those scripts in their entirety. While pieces of the script can be run manually during development, run the entire script on the data when the time comes to actually produce a result.

3 Code speed

When optimizing the execution speed of your code, it's essential to strike a balance between code readability and efficiency. However, as Donald Knuth famously stated, 'premature optimization is the root of all evil.' Focus on writing clean and functional code first. Once your code works correctly, you can then consider optimizing the most computationally intensive parts if necessary.

3.1 Vectorization

R is a vectorized language, which means that operations can be performed on entire vectors rather than looping over individual elements. For example, instead of using a loop to square each element of a vector, you can simply square the vector directly:

```r
numbers <- c(1, 2, 3, 4, 5)

# Non-vectorized operation using a loop
squared_numbers <- vector("numeric", length(numbers))
for (i in seq_along(numbers)) {
  squared_numbers[i] <- numbers[i]^2
}

# Example of vectorized operation
squared_numbers <- numbers^2
```

3.2 Factoring code out of loops

Often, parts of the code inside a loop don't depend on the loop variable and can be taken outside the loop, leading to efficiency gains. For example, if you're repeatedly computing something within a loop that doesn't change, compute it once outside the loop.

```r
# Inefficient loop: Compute mean of the entire dataset in each iteration
for (i in 1:n_bootstrap) {
  sample <- sample(data_points, size = length(data_points), replace = TRUE)
  mu_data <- mean(data_points) # This is unnecessary in the loop
  bootstrap_means[i] <- mean(sample) - mu_data
}

# Optimized loop: Factor out the mean computation of the dataset
bootstrap_means_optimized <- numeric(n_bootstrap)
mu_data <- mean(data_points)
for (i in 1:n_bootstrap) {
  sample <- sample(data_points, size = length(data_points), replace = TRUE)
  bootstrap_means[i] <- mean(sample) - mu_data
}
```