Looping Structures

Samuel P Callisto & Ashwin Karanam

June 28, 2019

Contents

Part 2: Looping Structures

Great reference: 'R for Data Science: Iteration'

Iteration is an important aspect of coding because it allows you to repeat operations multiple times without copy-pasting sections of your code. There are many looping structures available in R, but the most commonly used is the for-loop.

For-loops

Three essential components for a for-loop

- Output
- Sequence
- Body

```
## create example dataset
df <- data.frame(
  a = rnorm(10),
  b = rnorm(10),
  c = rnorm(10),
  d = rnorm(10)
)

output <- vector("double", 4)      # 1. output
for (i in 1:ncol(df)) {            # 2. sequence
  output[i] <- median(df[,i])      # 3. body
}
output
```

```
## [1] -0.11937503 -0.08361914  0.40557951  0.23181049
```

About the output container

It may be tempting to skip this step and grow the size of your output variable with each iteration. However, this uses much more memory and can cause your computer to crash on very large datasets ("computationally expensive"). Best practice is to always create a container for your output prior to generating it, such as a vector, matrix, list, or data.frame. If you are unable to know the length of your output before starting the loop, save your output in a list data type (see 'R for Data Science: Unknown Output Length').

Slightly fancier method

- calculating number of required spaces from dataset rather than hard-coding 4; this allows for more flexible input
- seq_along() is a wrapper function for length() which avoids zero-length vector errors
- can use element selection [[]] for both vectors and data.frames rather than using different syntax

```
output <- vector("double", ncol(df))   # 1. output
for (i in seq_along(df)) {             # 2. sequence
  output[[i]] <- median(df[[i]])       # 3. body
}
output
```

```
## [1] -0.11937503 -0.08361914  0.40557951  0.23181049
```

Alternate for-loop structures

The basic for-loop uses the variable i as an index through a vector or data.frame. In some cases it might be advantageous to access values directly rather than by using an index. There are two alternate methods for iteration using for-loops: 1: n in names(xs) 2: x in xs

```
## create empty vector for output with meaningful name
mtcarsMeans <- vector("numeric", ncol(mtcars))

## add column names to output vector
names(mtcarsMeans) <- names(mtcars)

## loop through columns to calculate mean for each
for(i in names(mtcars)){
  mtcarsMeans[[i]] <- mean(mtcars[[i]])
}

## display output
mtcarsMeans
```

```
##        mpg        cyl       disp         hp       drat         wt
##  20.090625   6.187500 230.721875 146.687500   3.596563   3.217250
##       qsec         vs         am       gear       carb
##  17.848750   0.437500   0.406250   3.687500   2.812500
```

Breaking out of loops

Break

Sometimes there will be a case in which you want to stop (break) or skip (next) when you encounter an element.

```
lettersBeforeP <- ""                                   # output
for(i in LETTERS){                                     # sequence
  lettersBeforeP <- paste(lettersBeforeP,i,sep= " ")   # body
  if(i == "P"){
    break
  }
}
lettersBeforeP
```

```
## [1] " A B C D E F G H I J K L M N O P"
```

Next

Using break will cause the loop to end, but you can use next to skip to the next iteration and continue looping

```
alphaWithoutSAM <- ""                                # output
for(i in letters){                                   # sequence
  if(i == "s" | i == "a" | i == "m") next            # body
  alphaWithoutSAM <- paste(alphaWithoutSAM, i, " ")
}
alphaWithoutSAM
```

```
## [1] " b   c   d   e   f   g   h   i   j   k   l   n   o   p   q   r   t   u   v   w   x   y   z  "
```

Nesting for-loops

Sometimes you will be utilizing multi-level data that varies by multiple factors. In these cases we can utilize multiple for-loops nested within each other.

Example: creating a times table from one through ten

```
## create output container
timesTable <- matrix(-99,nrow=10, ncol=10)

## iterate through row dimension
for(i in 1:10){
  ## iterate through column dimension
  for(j in 1:10){
    timesTable[i,j] = i*j
  }
}
timesTable
```

```
##       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
##  [1,]    1    2    3    4    5    6    7    8    9    10
##  [2,]    2    4    6    8   10   12   14   16   18    20
##  [3,]    3    6    9   12   15   18   21   24   27    30
##  [4,]    4    8   12   16   20   24   28   32   36    40
##  [5,]    5   10   15   20   25   30   35   40   45    50
##  [6,]    6   12   18   24   30   36   42   48   54    60
##  [7,]    7   14   21   28   35   42   49   56   63    70
##  [8,]    8   16   24   32   40   48   56   64   72    80
##  [9,]    9   18   27   36   45   54   63   72   81    90
## [10,]   10   20   30   40   50   60   70   80   90   100
```

Example: closer to home

Let's create a dataset we're more familiar with. Let's say you have a study where subjects have multiple visits and you have demographic information on each subject at each visit. Let's try to create a dataset for this study.

```
set.seed(123456)
ID <- sort(rep(seq(1,140,by=1),times=10),decreasing = FALSE)
VISIT <- rep(c(1,1,2,2,3,3,4,4,5,5),times=140)
DRUG <- rep(c("A","B"),times=700)
DOSE <- round(rnorm(1400,mean=1000,sd=200),-2)
DOSE <- ifelse(DOSE < 0, 0,DOSE)

stdx <- data.frame(ID,VISIT,DRUG,DOSE)
head(stdx)
```

```
##   ID VISIT DRUG DOSE
## 1  1     1    A 1200
## 2  1     1    B  900
## 3  1     2    A  900
## 4  1     2    B 1000
## 5  1     3    A 1500
## 6  1     3    B 1200
```

Let's say you want to calculate percent change in dose for each individual for each drug at each visit relative to last visit. Ideas? Think of how many loops you have to write. Start with the top-most loop. Which one is it?

```r
## Loop over individuals and then loop over

OUTPUT <- data.frame(NULL)
numid <- unique(stdx$ID)
for (i in numid) {
  temp1 <- subset(stdx,ID==i)
  numdrug <- unique(temp1$DRUG)

  for (j in numdrug) {
    temp2 <- subset(temp1,DRUG == j)
    perch.rel <- (temp2$DOSE - dplyr::lag(temp2$DOSE))*100/dplyr::lag(temp2$DOSE)
    temp3 <- cbind(temp2,perch.rel)
    OUTPUT <- rbind(OUTPUT, temp3)
  }
}

head(OUTPUT)
```

```
##   ID VISIT DRUG DOSE  perch.rel
## 1  1     1    A 1200         NA
## 3  1     2    A  900 -25.000000
## 5  1     3    A 1500  66.666667
## 7  1     4    A 1300 -13.333333
## 9  1     5    A 1200  -7.692308
## 2  1     1    B  900         NA
```

While writing loops for such problems will be the most intuitive solution, remeber in coding the most intuitive isn't always the best. Can you accomplish this same not using loops? Answer discussed at the end!

While-loops

While

This type of looping structure is useful for situations when the number of iterations necessary for the task to finish is unknown. It should be noted that all for-loops can be re-written as as a while-loop, but the opposite is not true.

```r
countToTen <- ""
i <- 1
while(i <= 10){
  countToTen <- paste(countToTen, i, sep=" ")
  i <- i + 1
}
```

```
countToTen
```

```
## [1] " 1 2 3 4 5 6 7 8 9 10"
```

Repeat

A modifed version of the while loop is repeat, which will keep running until it reaches a break statement.

```
countToTen <- ""
i <- 1
repeat{
  countToTen <- paste(countToTen, i, sep=" ")
  if(i==10){
    break
  }else{
    i <- i + 1
  }
}
countToTen
```

```
## [1] " 1 2 3 4 5 6 7 8 9 10"
```

Apply and Purrr

Base R gives us the apply() family of functions, which can be useful alternatives to for-loops. Let's revisit the example from earlier of calculating the mean for all columns in the mtcars dataset

```
apply(mtcars,2,mean)
```

```
##        mpg        cyl       disp         hp       drat         wt 
##  20.090625   6.187500 230.721875 146.687500   3.596563   3.217250 
##       qsec         vs         am       gear       carb 
##  17.848750   0.437500   0.406250   3.687500   2.812500
```

If your function can easily adhere to the syntax of the apply() functions, it can give the same results as a for-loop much less code. * apply(): evaluate a function over the margins of a matrix or array * lapply(): evaluate a function on each element in a list, and return the results as a list * sapply(): evaluate a function on each element in a list, and return the results in a "simplified form" (not always predictable, but can be convenient) * vapply(): similar to sapply(), but with more consistent return types * tapply(): evaluate a function on subsets of a vector; alternative to group_by() for dealing with subsets

Another useful package in the tidyverse is 'purrr', which is similar to apply, but strives to be more consistent with argument structure and syntax.

The difference in run-time between all these different methods of iteration is fairly similar (though purrr is likely the fastest), so just choose whichever syntax you prefer.

Looping without loops?

the `dplyr` is extremely powerful if you know how to use it. The `group_by()` function can improve your coding by reducing dependency on `for` loops. Let's see how with our `stdx` dataset

```
stdx.2 <-
  stdx %>%
  group_by(ID,DRUG) %>%
  mutate(perch.rel = (DOSE-lag(DOSE))*100/lag(DOSE))
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

```
OUTPUT.2 <- stdx.2[order(ID,DRUG),]

head(OUTPUT.2)
```

```
## # A tibble: 6 x 5
## # Groups:   ID, DRUG [2]
##       ID VISIT DRUG   DOSE perch.rel
##    <dbl> <dbl> <fct> <dbl>     <dbl>
## 1  1.00  1.00 A      1200     NA
## 2  1.00  2.00 A       900    -25.0
## 3  1.00  3.00 A      1500     66.7
## 4  1.00  4.00 A      1300    -13.3
## 5  1.00  5.00 A      1200    - 7.69
## 6  1.00  1.00 B       900     NA
```

```r
sessionInfo()
```

```
## R version 3.4.3 (2017-11-30)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 17134)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] bindrcpp_0.2.2  forcats_0.3.0   stringr_1.3.1   dplyr_0.7.6
##  [5] purrr_0.2.5     readr_1.1.1     tidyr_0.8.1     tibble_1.4.2
##  [9] ggplot2_3.0.0   tidyverse_1.2.1
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.17     cellranger_1.1.0 pillar_1.1.0     compiler_3.4.3
##  [5] plyr_1.8.4       bindr_0.1.1      tools_3.4.3      digest_0.6.15
##  [9] lubridate_1.7.4  jsonlite_1.5     evaluate_0.10.1  nlme_3.1-131
## [13] gtable_0.2.0     lattice_0.20-35  pkgconfig_2.0.1  rlang_0.2.1
## [17] psych_1.7.8      cli_1.0.0        rstudioapi_0.7   yaml_2.1.16
## [21] parallel_3.4.3   haven_1.1.1      withr_2.1.1      xml2_1.2.0
## [25] httr_1.3.1       knitr_1.20       hms_0.4.1        rprojroot_1.3-2
## [29] grid_3.4.3       tidyselect_0.2.3 glue_1.2.0       R6_2.2.2
## [33] readxl_1.0.0     foreign_0.8-69   rmarkdown_1.8    modelr_0.1.1
## [37] reshape2_1.4.3   magrittr_1.5     backports_1.1.2  scales_1.0.0
## [41] htmltools_0.3.6  rvest_0.3.2      assertthat_0.2.0 mnormt_1.5-5
## [45] colorspace_1.3-2 utf8_1.1.3       stringi_1.1.7    lazyeval_0.2.1
## [49] munsell_0.5.0    broom_0.4.3      crayon_1.3.4
```