506 Problem Set 6

Xiaohan Liu

Table of contents

GitHub repository:

```
library(nycflights13)
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
library(parallel)
library(future)
```

```
Warning: package 'future' was built under R version 4.3.2
```

```
library(furrr)
```

```
Warning: package 'furrr' was built under R version 4.3.2
```

```r
#set.seed(506)

#' Function to perform bootstrap sampling
#' @param data data to sample from
#' @param n_bootstrap number of bootstrap samples to take
#' @return list of data frames with mean air time for each origin
bootstrap_means <- function(data, n_bootstrap = 1000) {
  boot_means <- replicate(n_bootstrap, {
    sample_data <- data %>% group_by(dest) %>%
      sample_n(size = n(), replace = TRUE)
    mean_data <- sample_data %>% group_by(origin) %>%
      summarize(mean_air_time = mean(air_time, na.rm = TRUE))
    return(mean_data)
  }, simplify = FALSE)
  return(boot_means)
}

## Select necessary columns
flights_data <- flights %>% select(origin, dest, air_time)
```

Non-Parallel

```r
## Measure time for non-parallel bootstrap sampling
system.time({
  boot_results <- bootstrap_means(flights_data)
  ## produce a table including the estimates and CI for each origin
  final_results <- bind_rows(boot_results) %>%
    group_by(origin) %>%
    summarize(
      Mean = mean(mean_air_time),
      Lower_CI = quantile(mean_air_time, probs = 0.025),
      Upper_CI = quantile(mean_air_time, probs = 0.975)
    )
})
```

```
 user  system elapsed
75.53    5.40   88.26
```

```r
print(final_results)
```

```
# A tibble: 3 x 4
  origin  Mean Lower_CI Upper_CI
  <chr>  <dbl>    <dbl>    <dbl>
1 EWR     153.     153.     154.
2 JFK     178.     178.     179.
3 LGA     118.     118.     118.
```

Parallel

```r
## Parallel version using 'parallel' package
system.time({
  n_cores <- 8
  cl <- makeCluster(n_cores)
  clusterExport(cl, varlist = c("bootstrap_means", "flights_data", "n_cores"))
  clusterEvalQ(cl, {library(dplyr)})

  boot_results_parallel <- parLapply(cl, 1:n_cores,
                                     function(x)
                                       bootstrap_means(flights_data,
                                                       n_bootstrap = 1000/n_cores))
  boot_results2 <- do.call("rbind", boot_results_parallel)

  ## produce a table including the estimates and CI for each origin
  final_results_parallel <- bind_rows(boot_results2) %>%
    group_by(origin) %>%
    summarize(
      Mean = mean(mean_air_time),
      Lower_CI = quantile(mean_air_time, probs = 0.025),
      Upper_CI = quantile(mean_air_time, probs = 0.975)
    )

  print(final_results_parallel)

  stopCluster(cl)
})
```

```
# A tibble: 3 x 4
  origin  Mean Lower_CI Upper_CI
  <chr>  <dbl>    <dbl>    <dbl>
1 EWR     153.     153.     154.
2 JFK     178.     178.     179.
```

```
3 LGA     118.     118.     118.
```

```
   user  system elapsed
   0.24    0.18   29.97
```

Future

```
## Parallel version using 'future' packages
plan(multisession, workers = detectCores() - 1)

system.time({
  ## Perform bootstrap sampling using 'future' package
  boot_results_future <- future_map(1:1000,
                                    ~bootstrap_means(flights_data,
                                                     n_bootstrap = 1),
                                    .options = furrr_options(seed = TRUE))

  ## produce a table including the estimates and CI for each origin
  final_results_future <- bind_rows(boot_results_future) %>%
    group_by(origin) %>%
    summarize(
      Mean = mean(mean_air_time),
      Lower_CI = quantile(mean_air_time, probs = 0.025),
      Upper_CI = quantile(mean_air_time, probs = 0.975)
    )

  print(final_results_future)
})
```

```
# A tibble: 3 x 4
  origin  Mean Lower_CI Upper_CI
  <chr>  <dbl>    <dbl>    <dbl>
1 EWR     153.     153.     154.
2 JFK     178.     178.     179.
3 LGA     118.     118.     118.
```

```
   user  system elapsed
   0.87    0.14   35.97
```