Advanced Topics in R: Parallel Processing and Bootstrapping

Ian Crandell

April 22nd, 2015

Outline

1. Bootstrap Introduction
2. Bootstrap Applications
3. Parallel Processing Using Snowfall

Bootstrap Introduction: What is the bootstrap?

- The bootstrap is a way to estimate the uncertainty of a statistic without making distributional assumptions such as normality.
- The main idea is to *resample* from your original data set with replacement to create a new dataset.
- This works because, for large N, the sample distribution is very close to the population distribution.
- Remember the analogy: the population is to the sample as the sample is to the bootstrap samples.

Bootstrap Introduction: Procedure

Suppose we have a sample of size N and we wish to estimate its mean along with an uncertainty interval for the mean. The bootstrap proceeds thus:

1. Sample N data points *with replacement* from the original data set. This is the bootstrapped data set.
2. Compute the mean of this data set.
3. Repeat this procedure B times creating a "sample" of B means.
4. The mean of these means is the bootstrap estimate of the true mean.
5. Pick the appropriate quantiles of this sample as the end points of your bootstrap uncertainty interval for the original mean.

This procedure is known as the *non-parametric bootstrap*. Results from Efron and Tibsherani indicate that using B = 2,000 is almost always sufficient

Bootstrap Introduction: Advantages

- The bootstrap makes no assumptions about the shape of the underlying data.
- *Any* uncertainty interval for any sample statistic can be estimated this way, even when we have no idea what the underlying distribution is.

Bootstrap Applications

In the accompanying R script we cover the following examples:

1. Bootstrap interval for normal and χ^2_3 data vs normal theory confidence interval.
2. Bootstrap interval for the median of a χ^2_3 random variable.
3. Bootstrap interval for the standard deviation of a χ^2_3 random variable.

Bootstrap Applications: Linear regression

1. Typical inference in linear models relies on assumptions about the residuals: that they be normally distributed with equal variance.
2. When these assumptions are not met, inference can be misleading.
3. The bootstrap can be used to create point estimates and uncertainty intervals for the parameters without making assumptions about the residuals.

Bootstrap Applications: Linear regression proceure

1. Consider a data set with response Y and regressors X_i for $i = 1, \ldots, p$. Let $Z = [Y, X_1, \ldots, X_p]$
2. Resample with replacement from the rows of Z
3. Run the regression B times on the resampled rows, save the coefficient estimates.
4. These saved coefficient estimates are bootstrapped samples from the joint distribution of the coefficient vector.

This is called the *random X* bootstrap. A fixed version is also available, but it is less robust to violations of linear regression assumptions (homoskedasticity, appropriateness of the model).

Parallel Processing Using Snowfall: Introduction

- Almost all modern computers have at least two processing cores (CPUs). By default, however, only one CPU is typically used to do computations.
- For a certain class of procedures, we can take advantage of all of a computer's CPUs with minimal hassle. The bootstrap is one such procedure.
- Rather than do B resamples on 1 core, we can do $\frac{B}{c}$ resamples on c cores. This will produce the same number of resamples and *usually* be considerably quicker.

Parallel Processing Using Snowfall: Runtime

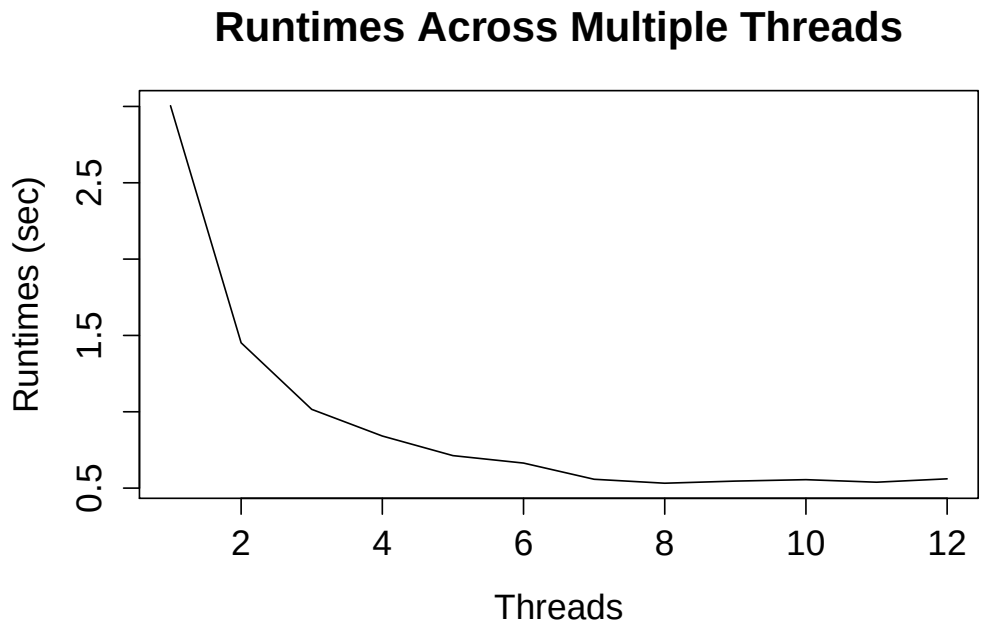

Figure: Runtimes on my laptop with 2.6GHz quad core. Times do not include cluster set up.

Parallel Processing Using Snowfall: Initializing code

- install.packages("snowfall")
- library(snowfall)
- sfInit(parallel = TRUE, cpus = 4)
- sfLibrary(...)
- sfExport(...)
- *sfSapply(...)*
- sfStop()

Parallel Processing Using Snowfall: sfLapply

- All snowfall computing functions are derived from the apply suite of functions in the R base package. We'll focus on sfLapply.
- *sfLapply(X, FUN, ...)*
- X is a list of numbers to which FUN, some function, will be applied to in sequence. The ellipsis indicates additional arguments which can be passed to FUN.
- For the bootstrap, X will represent the iteration numbers for our bootstrap samples. We'll need to use a dummy variable in our function to accomplish this.

References

1. "Bootstrapping Regression Models" John Fox, http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-bootstrapping.pdf
2. "Introduction to the Bootstrap" Efron and Tibshirani (1993)