Working with "Big Data"

Due: Tuesday, June 2nd, 4pm.
Printed version of report in Nick Ulle's mailbox.
Send the electronic version to dtemplelang@ucdavis.edu along with the SSH URL for the git repository on bitbucket (having given access to the duncantl and nick-ulle logins).

In this assignment, you are to work with the New York taxi data. These are available http://www.andresmh.com/nyctaxitrips/ and there is a description of the fields at http://publish.illinois.edu/dbwork/open-data/.

Two of the data files are available from eeyore:

- trip_data_1.csv.zip
- trip_fare_1.csv.zip

Your tasks are:

1. Compute the deciles of the *total amount* less the tolls.
2. Fit a linear regression predicting *total amount* less the tolls using *trip time* as the predictor.
3. **BONUS MARKS**: Use multiple regression and add the surchage as a regressor, estimating the resulting coefficients.

Implement these tasks in (at least) two distinct ways. Possible approaches include:

1. High-level languages such as R, Python, or MATLAB. Use whatever additional packages you want to make the computations more efficient.
2. Shell commands.
3. Low-level languages such as C or C++.
4. Random sampling. Use the Bag of Little Bootstraps to compute standard errors of the estimates.
5. A relational database.
6. Parallel processing.
7. Other technologies such as Hadoop, Spark, Mahout, ...

You may use R or other languages to supplement your work on any of these. In other words, the approaches are not mutually exclusive. For example, you could use R with a database, and do the computations on chunks of data in parallel.

If doing the computations by individual row or blocks of row, make certain to account for numerical imprecision, e.g., when computing the mean by maintaining the total and the sample size. For example, consider Welford's method described in John Cook's blog entry. Also, see Wikipedia. Welford's paper is available via JSTOR and is only 2 pages with some algebra to illustrate/derive the updating formula.

Get your approaches working on one of the files, then extend the computations to the other files.

You might be interested in exploring other packages such as bigmemory, data.table and its fread() function.

Report

Discuss and compare the approaches you used. Address the following questions:

- Were the results the same? If there are differences, explain why.
- How much computational time did each approach require?
- How much programming time did each approach require? Describe any challenges you encountered.
- What are the benefits and drawbacks of each approach? Under what circumstances is one approach more appropriate than another? and why?

When writing your report:

- Include the URL of your Bitbucket repository.
- Put margins on every page, including source code.
- Use a white or light-colored background for source code.
- Format tables neatly (no raw output).
- Put your files in the Assignment5 directory of your repository.
- Write 3 - 6 pages, at least half text.

Duncan Temple Lang *<duncan@r-project.org>*

Last modified: Mon May 19 01:11:00 PDT 2015