Statistical Issues in Large Database Research

Timothy T. Houle, PhD

Disclosures

- NIH Grants
- No relationships with industry

Overview

- Estimation
  - Big data, big challenges
- Interpretation
  - Performance of statistical tests
- Common errors
  - Bias reduction fallacy
  - Garden of forking paths
- Best practices

When does data become “Big”?

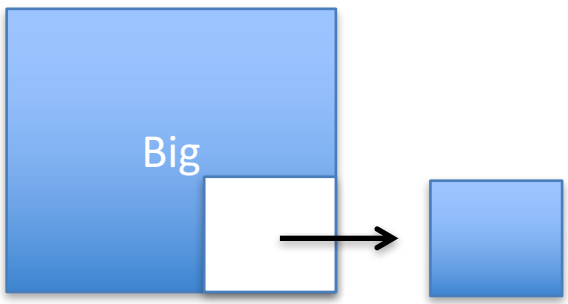
- No one definition
- Depends somewhat on available hardware/software
  - And the objects created during the process (e.g., 10,000 rows produces ~ 50,000,000 distances in clustering algorithms)
- Fun definitions: You have big data if…
  - you can get coffee while waiting for the procedure to run
  - it takes more time to estimate an analysis than to design it.

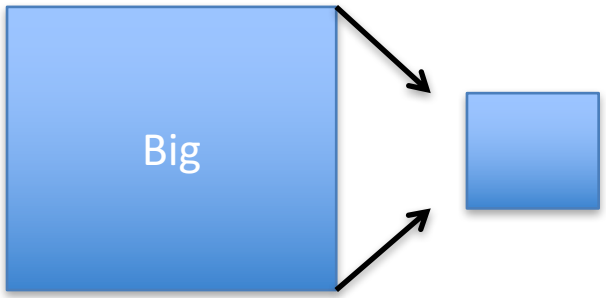
Pragmatic definitions

- In-memory versus disk
  - Plan and index computations with on-disk data
- One computer to many computers
  - Distributed environment

Three types of big data estimation problems

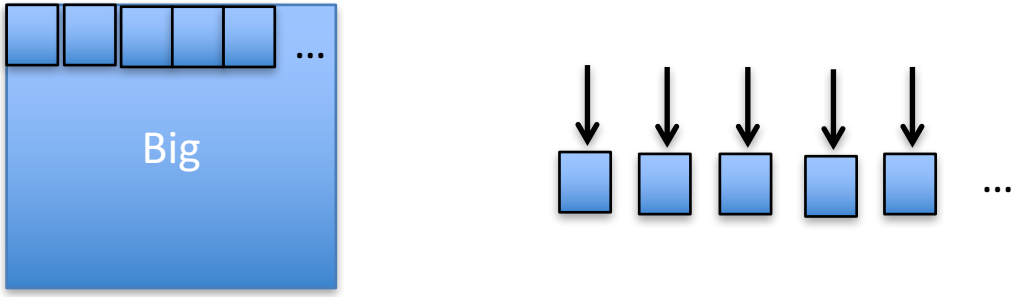
1. Big data problems that can be made to be small data using subsets or a summary
   - 90% of big data problems fall into this category.
   - Tools:
     - distributed database (e.g., hive, impala, teradata)
     - dplyr to subset or summarize





Three types of big data estimation problems

2. Big data problems that are actually lots and lots of small data problems (e.g., fit one model per individual for N = 5,000,000).
   - 9% of big data problems fall into this category
   - This is parallelism (e.g., “embarrassingly parallel”) Tools: foreach package (R), Hadoop, Spark, distributed computation



Three types of big data estimation problems

3. Irretrievably big data problems where you must have all the data (e.g., complex model with rare interactions).
   - 1% of problems probably fall into this category
   - Tools: de novo systems that are specifically designed to solve a particular problem. Techniques involve deferred computation, just-in-time computation, parrallelism

https://peadarcoyle.wordpress.com/2015/08/02/interview-with-a-data-scientist-hadley-wickham/

Rules of thumb for estimation in R

- Jan Wijffels (useR!-Conference) rules of thumb:
  - One million records can easily be analyzed with standard R.
  - One million to one billion records can also be processed in R, but need special attention.
  - More than one billion records need to be analyzed by map reduce algorithms (Hadoop, Spark)

https://www.r-bloggers.com/five-ways-to-handle-big-data-in-r/

Interpretation of large database analyses

- Modeling versus hypothesis testing
  - Simple hypothesis tests (e.g., t-test) are uncommon
- Databases not designed to examine the exact hypothesis/inference
- Sophisticated models are often needed
  - Isolate an effect of interest (confounders, covariates)
  - Reduce a source of bias (e.g., indication bias)
  - Strengthen causal inferences (e.g., quasi-randomization, propensity, mediation)

Statistical hypothesis testing

- As N gets very large, statistical power gets very high
  - Even miniscule differences can be found to be statistically significant
  - With independent samples of N = 10 million (each), power = 99.98% to detect a difference between 20.0% and 20.1%

- Bayesian methods are little better as priors are overwhelmed with the data (i.e., the prior probability must be very strong to have an influence on the parameter estimates)

Performance of common statistical tests

- Many common tests must be interpreted with caution in large samples
  - Hosmer-Lemeshow (PMID: 22833304)
  - Goodness of fit tests
  - Control charts (CI bounds)
  - Many others

- Must focus on effect sizes and clinical significance of the statistical inferences.
  - Especially using marginal effects

Common Errors

- Ignoring sampling issues
  - All statistical analyses are exercises in observational selection
- Unreliable measurement
  - The association between X and Y is attenuated by the product of their unreliabilities

Common Errors

- Machine learning is not the answer to all of the problems on Earth
  - Overfitting is still a problem in large databases (especially for rare events)
- Lack of internal validation
  - Some large database efforts effectively have infinite data (e.g., massive new data sets posted each day)

Common Errors

- Garden of forking paths (Andrew Gelman)
  - Many, many analyses could have been conducted (and many actually were)
  - Given the high degree of statistical power, the null distribution should be carefully considered
  - Surprising findings should be further vetted

- "Unregistered databases are like weapons of mass destruction"
  - My misrepresentation of quote from John Ionnadis

http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Best Practices

- Ask for help
  - Not just statistical issues, also exercises in computer science
- Think about the sampling, think about the measurements, think about the hypotheses, think about the threats to the interpretations
- Define the effect sizes that would have clinical/pragmatic significance BEFORE conducting the analysis
- Register the plan of analysis, either on a external site or with an internal entity
  - Reduce the threat of the garden of forking paths

Thank You