Logistic Regression with Differential Privacy

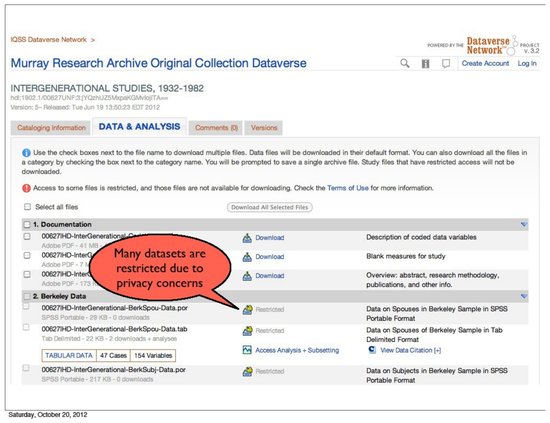
Paul Handorff

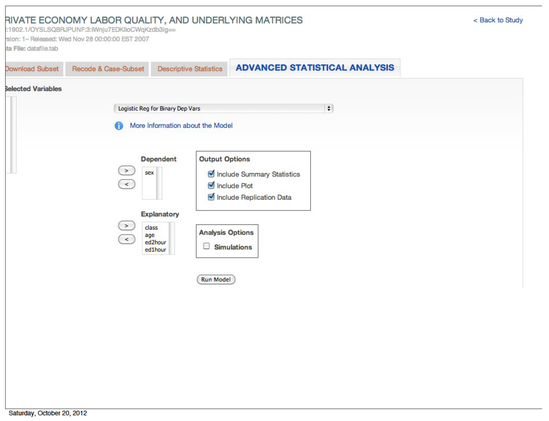
Harvard College ’14, supported by HCRP

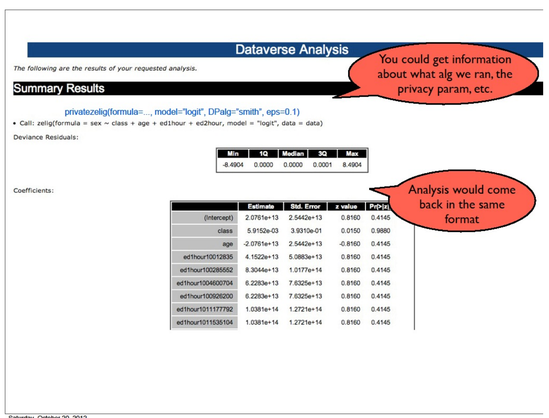
Privacy Tools for Sharing Research Data

A National Science Foundation Secure and Trustworthy Cyberspace Project

PRIVATE ZELIG WORKFLOW





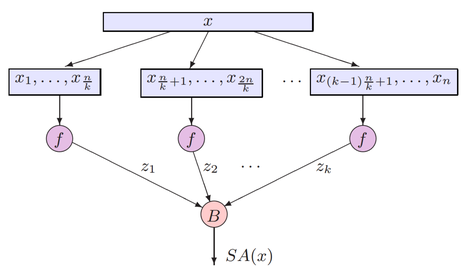


OBJECTIVES

- Gather data on error introduced by differentially private algorithms for statistical estimation on real-world datasets.
- Determine which steps of differentially private algorithms introduce the most error, in order to guide future theory work.
- Investigate whether black-box methods are well-suited for implementing Private Zelig.
- Allow for statisticians and quantitative social scientists to provide feedback on how they could integrate differentially private statistical estimation techniques into their workflows.

BACKGROUND

- The simplest differentially private algorithms add noise based on the *global sensitivity* of the function being computed.
- For real-world datasets, this worst-case assumption is often too pessimistic; better to add *instance-based* noise.
- First attempt: Add noise based on the *local sensitivity* of the function being computed at the particular database used.
- **Problem:** For many functions of interest, it is hard to compute the local sensitivity at a particular database.
- Second attempt: Add noise based on the *smooth sensitivity*, a smooth upper bound on the local sensitivity. This allows computation of many quantities of interest with acceptable levels of noise where using global sensitivity would “drown out” signal: minimum, maximum, MST cost, etc.
- **Problem:** There are still commonly-used functions for which no efficient algorithm to compute the smooth sensitivity is known, including finding k-means cluster centers and learning mixtures of Gaussians.
- **Sample and Aggregate Framework**: Split input database into k different subsamples, then compute f on each one. Combine the k inputs using an aggregation function. (See figure below.)
- Changing a row of the database x changes only one of the subsamples, so we can bound the smooth sensitivity at x by the smooth sensitivity of the aggregation function.
- Introduced by [NRS07].



Sample and Aggregate Framework [S ‘11]

METHODS

- Using Adult dataset from UCI Machine Learning Repository. (~30,000 records, 15 dimensions including both numerical and categorical fields)
- We can choose a random subset of the Adult dataset to test performance on smaller datasets.
- Perform logistic regression using Zelig’s built-in model.
- Two ways to measure performance:
  (1) Logistic Regression is a maximization problem (maximizing log likelihood); we can compare the maximum log likelihood achieved with non-private logistic regression to the maximum achieved with sample and aggregate.
  (2) Adult dataset has a prediction task: Given age, education, occupation, hours worked per week, and other such fields, predict whether an individual’s income is at least $50,000. We can compare the accuracy of the predictions resulting from private and non-private logistic regression.
- We can test different choices of aggregation function:
  Widened Winsorized Mean: Proposed by [S11]. Privately estimate the interquartile range, use it to omit outliers, then output the mean with additive noise based on the interquartile range, size of dataset, and privacy parameter.
  Median: The median of a dataset can be privately estimated in a number of different ways, including by the *exponential mechanism*, *smooth sensitivity* and *propose-test-release*.
- Currently: Requires the user to manually select a bound (on coefficients) to be used in differentially private algorithms.

PRELIMINARY RESULTS

- Sample and aggregate performs well using the *median* aggregation function under ideal conditions: ~30,000 records, 3 dimensions. Log likelihood of private version within 1-3% of the optimal log likelihood found by non-private logistic regression.
- Even with these ideal conditions, *Widened Winsorized Mean* performs fairly poorly.
- Error becomes more significant with fewer records (below ~5,000) or with more dimensions.
- The general technique of converting a categorical variable with l levels into l different binary (indicator) variables is not well-suited to differential privacy.

FUTURE WORK

- Develop an automated testing framework to gather more experimental data, and use this framework to better characterize the deterioration in error rates in databases with fewer records or more dimensions.
- Test different choices of aggregation functions on multiple datasets.
- Test a version of this procedure which does not require the user to input a bound on the regression coefficients introduced by [S11]; this procedure involves first privately estimating a bound, then using that bound in the original algorithms.
- Develop an “integrated unbounded procedure” using methods from [DL09] which may achieve better performance by combining the bound-estimation and statistical estimation steps described above.
- Isolate the error introduced by the subsampling and aggregation steps, in order to find “bottlenecks” in the process and guide future theory work.

REFERENCES

[NRS07]
Kobbi Nissim, Sofya Raskhodnikova, Adam Smith. *Smooth Sensitivity and Sampling in Private Data Analysis*. STOC ‘07.

[DL09]
Cynthia Dwork and Jing Lei. *Differential Privacy and Robust Statistics*. STOC ‘09.

[S11]
Adam Smith. *Privacy-Preserving Statistical Estimation with Optimal Convergence Rates*. STOC ‘11.

UCI Machine Learning Repository. *Adult Data Set*.
Available at http://archive.ics.uci.edu/ml/datasets/Adult.

ACKNOWLEDGEMENTS

I would like to thank my advisers, Prof. Salil Vadhan and Jonathan Ullman, for their support throughout the summer and this fall; the Harvard College Research Program, for financial support during the summer; and the National Science Foundation, for supporting many of the group meetings and other events of the *Privacy Tools for Sharing Research Data* project which I was fortunate enough to attend last summer.

CRCS Center for Research on Computation and Society at Harvard School of Engineering and Applied Sciences

BERKMAN CENTER FOR INTERNET & SOCIETY AT HARVARD UNIVERSITY

The Institute *for* Quantitative Social Science at Harvard University

DATA PRIVACY LAB