Real-time Regression Analysis of Streaming Health Datasets

Lan Luo

This work is supervised by Professor Peter X.K. Song.

UNIVERSITY OF MICHIGAN

March 25, 2019

Motivation

Challenges:

- storage for cumulatively growing datasets;
- recomputation when new data batch arrives.

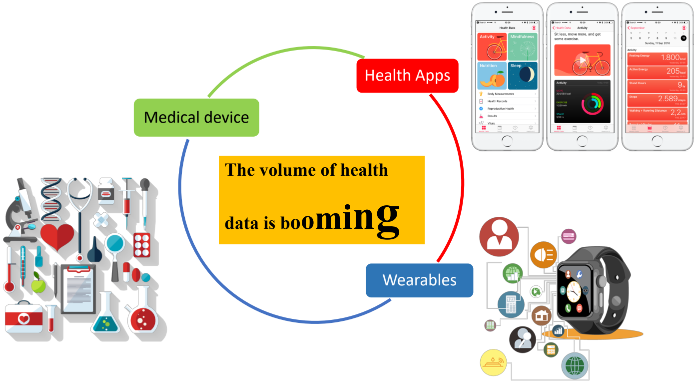
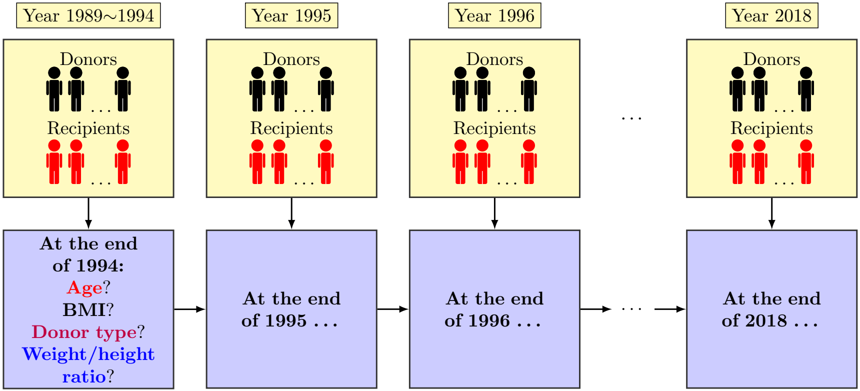


Image sources: https://unixtitan.net/explore/see-clipart-smart-watch/, https://www.idownloadblog.com/2018/01/11/health-app-data-murder-investigaation/, http://medtechasia.in/new-medical-devices-being-regulated-under-the-drugs-and-cosmetics-act/.

Motivation



Objective: updating the association between 5-year graft failure and demographic features of donors/recipients based on yearly updated kidney transplant data.

Introduction

HMMcopy.pdf

1. Rule: Updating the objective statistics without historical raw data but only summary statistics.
2. A simple example:

$$M(D_1, D_2) = \frac{1}{n_1 + n_2} \left(\sum_{i=1}^{n_1} x_{1i} + \sum_{i=1}^{n_2} x_{2i} \right) = \frac{1}{n_1 + n_2} \left(n_1 M(D_1) + \sum_{i=1}^{n_2} x_{2i} \right).$$

3. Question: Can the maximum likelihood estimation (MLE) in generalized linear models (GLM) be updated sequentially like the sample mean?

Model and Notations

- GLM: continuous, binary and count data;
- Suppose $(y_i, \boldsymbol{x}_i) \sim f(y, \boldsymbol{x}; \beta_0, \phi_0)$ independently for $i = 1, \ldots, N_B$;
- Goal: fit a regression model with mean $\mathbb{E}(y_i \mid \boldsymbol{x}_i) = g(\boldsymbol{x}_i^T \beta)$.

notation.pdf

Existing Methods

Methods	Pros	Cons
Oracle MLE	(i) asymptotically unbiased (ii) statistical efficient	(i) store large data (ii) refit model
AI-SGD	online point estimation and bootstrap inference	large bias and standard error
OLSE	online point estimation and inference	linear model only
CEE/CUEE	online point estimation and inference	restriction: $B \ll n_b$

Table: AI-SGD denotes the averaged implicit stochastic gradient descent (Toulis & Airoldi, 2015); OLSE, CEE and CUEE are three online meta-type estimation methods (Schifano *et al.* , 2016).

Renewable Estimation

- A simple scenario with two data batches D_1 and D_2, we aim to solve $\hat{\beta}_2^\star$ satisfying

$$\boldsymbol{U}_1(D_1; \hat{\beta}_2^\star) + \boldsymbol{U}_2(D_2; \hat{\beta}_2^\star) = \mathbf{0},$$

- Taking the first-order Taylor expansion of the first term around $\hat{\beta}_1$ leads to

$$\color{red}{\boldsymbol{U}_1(D_1; \hat{\beta}_1)} + \boldsymbol{J}_1(D_1; \hat{\beta}_1)(\hat{\beta}_1 - \hat{\beta}_2^\star) + \boldsymbol{U}_2(D_2; \hat{\beta}_2^\star) + \color{red}{O_p(\|\hat{\beta}_2^\star - \hat{\beta}_1\|^2)} = \mathbf{0}.$$

- The error term may be asymptotically ignored. We propose a new estimator $\tilde{\beta}_2$ satisfies the following equation:

$$\boldsymbol{J}_1(D_1; \hat{\beta}_1)(\hat{\beta}_1 - \tilde{\beta}_2) + \boldsymbol{U}_2(D_2; \tilde{\beta}_2) = \mathbf{0}.$$

- $\tilde{\beta}_2$ is a renewable estimator of β_0, and the above equation is termed as an incremental estimating equation.

Renewable Estimation

- Generalizing to an arbitrary time point B, $\tilde{\boldsymbol{\beta}}_B$ is the solution to the following incremental estimating equation:

$$\sum_{b=1}^{B-1} \boldsymbol{J}_b(D_b; \tilde{\boldsymbol{\beta}}_b)(\tilde{\boldsymbol{\beta}}_{B-1} - \tilde{\boldsymbol{\beta}}_B) + \boldsymbol{U}_B(D_B; \tilde{\boldsymbol{\beta}}_B) = \mathbf{0}$$

- Solving via incremental updating algorithm:

$$\begin{aligned}\tilde{\boldsymbol{\beta}}_B^{(r+1)} &= \tilde{\boldsymbol{\beta}}_B^{(r)} + \left\{\sum_{b=1}^{B-1} \boldsymbol{J}_b(D_b; \tilde{\boldsymbol{\beta}}_b) + \boldsymbol{J}_B(D_B; \tilde{\boldsymbol{\beta}}_{B-1})\right\}^{-1} \\ &\quad \times \left\{\sum_{b=1}^{B-1} \boldsymbol{J}_b(D_b; \tilde{\boldsymbol{\beta}}_b)(\tilde{\boldsymbol{\beta}}_{B-1} - \tilde{\boldsymbol{\beta}}_B^{(r)}) + \boldsymbol{U}_B(D_B, \tilde{\boldsymbol{\beta}}_B^{(r)})\right\} \\ &= \tilde{\boldsymbol{\beta}}_B^{(r)} + \left\{\tilde{\boldsymbol{J}}_{B-1} + \boldsymbol{J}_B(D_B; \tilde{\boldsymbol{\beta}}_{B-1})\right\}^{-1} \tilde{\boldsymbol{U}}_B^{(r)}\end{aligned}$$

- Key components from historical data: $\left\{\tilde{\boldsymbol{\beta}}_{B-1}, \tilde{\boldsymbol{J}}_{B-1}\right\}$.

Theoretic Justification

Main Theorem

Under some regularity conditions, as $N_B = \sum_{b=1}^{B} n_b \to \infty$,

1. consistency: $\tilde{\boldsymbol{\beta}}_B \xrightarrow{p} \boldsymbol{\beta}_0$;
2. asymptotic efficiency: $\sqrt{N_B}(\tilde{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathcal{I}^{-1}(\boldsymbol{\beta}_0))$;
3. ignorable error to the oracle MLE: $\|\tilde{\boldsymbol{\beta}}_B - \hat{\boldsymbol{\beta}}_B^{\star}\|_2 = \mathcal{O}_p(1/N_B)$.

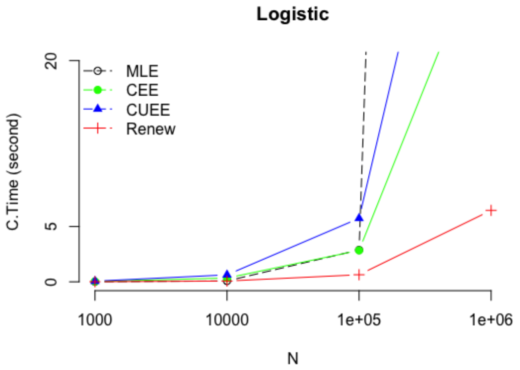
Implementation

- No historical data storage;
- Real-time estimation and inference.

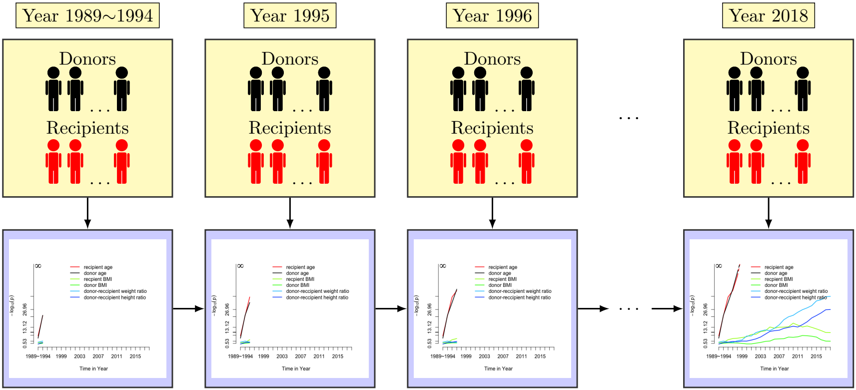
algorithmcol.pdf

Data Example

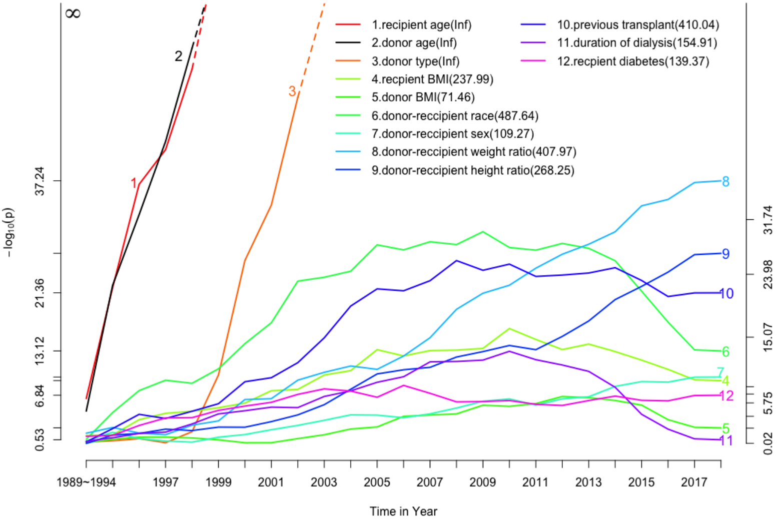
- Total sample size: 244,614 samples ($N > 10^5$);
- Model: logistic regression;
- Outcome: 1 for graft failure at the 5-th year after transplantation and 0 otherwise;
- Objective: update p-values based on yearly-arrived data batches.



Data Example



Data Example



Conclusions and Future Directions

1. Conclusions:
   - A **novel renewable estimation method** in the generalized linear model;
   - **Ignorable error** to the oracle MLE;
   - Our method is computationally efficient;
   - Real-time regression analysis without strong regularity conditions on relative scale of data batch size n_b and total number of batches B.
2. Future Directions:
   - Model homogeneity is assumed, and addressing partial homogeneous model is an undergoing direction.
3. **Questions?**

References

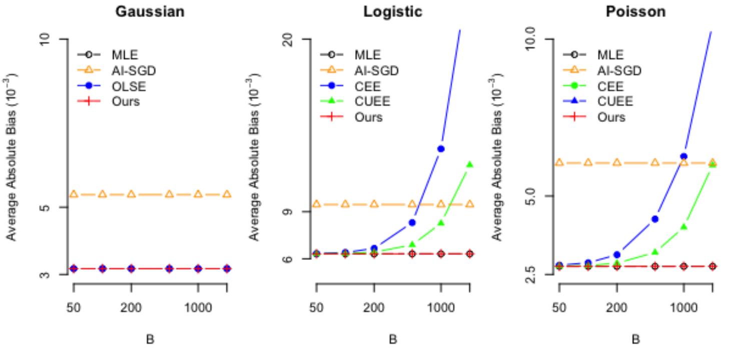
Schifano, Elizabeth D, Wu, Jing, Wang, Chun, Yan, Jun, & Chen, Ming-Hui. 2016. Online updating of statistical inference in the big data setting. *Technometrics*, **58**(3), 393–403.

Toulis, Panos, & Airoldi, Edoardo M. 2015. Scalable estimation strategies based on stochastic approximations: classical results and new insights. *Statistics and computing*, **25**(4), 781–795.

Simulation Design

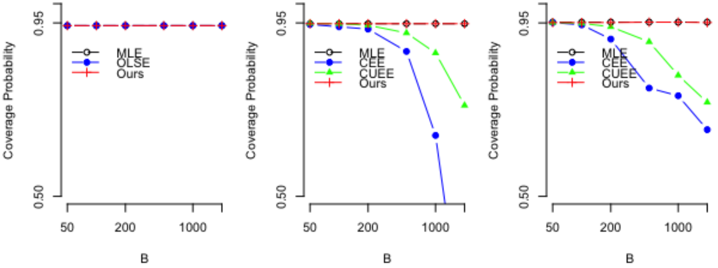
- B data streams with N_B independent observations $(y_i, \boldsymbol{x}_i)$ from GLMs with mean model $\mathbb{E}(y_i \mid \boldsymbol{x}_i) = g(\boldsymbol{x}_i^T \boldsymbol{\beta}_0)$;
- $\boldsymbol{\beta}_0 = (0.2, -0.2, 0.2, -0.2, 0.2)'$, $\boldsymbol{x}_{i[2:5]} \sim \mathcal{N}_4(\mathbf{0}, \boldsymbol{V}_4)$ where $\boldsymbol{V}_4$ is a compound symmetry covariance matrix with correlation 0.5.
- **Scenarios**:
  (1) divide a fixed $N_B = 10^5$ evenly into
  $B = 50, 100, 200, 500, 1000, 2000$ data streams;
  (2) fix sub-sample size $n_b = 100$, $b = 1, \ldots, B$ and let B increase from 10^5 to 10^6.

Simulation 1 - Bias



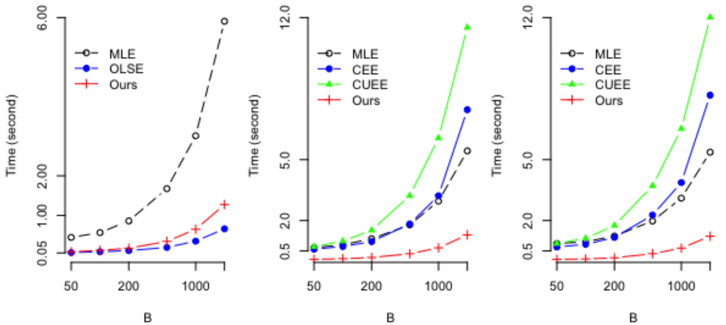
- Average bias in AI-SGD is the largest;
- In linear model, OLSE = Renew = MLE;
- In logistic or Poisson model, as B increases, CEE $>$ CUEE $\gg$ Renew $\approx$ MLE.

Simulation 1 - Coverage probability



- In linear model, OLSE = Renew = MLE;
- In logistic or Poisson model, 95% $\approx$ MLE $\approx$ Renew $\gg$ CUEE $>$CEE ≈ 0 as b increases;
- AI-SGD does not apply here.

Simulation 1 - Computation time



- Computation time = data loading time + processing time;
- In linear model, MLE > Renew > OLSE;
- In logistic or Poisson model, CUEE > CEE > MLE > Renew.

Simulation 2

Fix $n_b = 100$, increase B from 10^3 to 10^4:

	Logistic Model ($p = 5$, $n_b = 100$)				
	$B = 10^3$, $N_B = 10^5$				
	MLE	Renew	AI-SGD	CEE	CUEE
A.bias($\times 10^{-3}$)	5.023	5.020	**8.581**	**12.696**	6.218
Std. err.($\times 10^{-3}$)	6.538	6.539	-	6.576	6.581
Cov. prob.	0.956	0.958	-	**0.556**	**0.898**
C.Time (seconds)	2.951	0.663	-	3.022	5.763
R.Time (seconds)	0.471	0.487	0.165	2.846	5.587
	$B = 10^4$, $N_B = 10^6$				
A.bias($\times 10^{-3}$)	1.626	1.626	**8.581**	**12.978**	**4.157**
Std. err.($\times 10^{-3}$)	2.067	2.067	-	2.136	2.081
Cov. prob.	0.958	0.956	-	**0**	**0.644**
C.Time (seconds)	323.566	8.149	-	34.483	68.642
R.Time (seconds)	98.362	5.257	1.119	31.594	65.750

- Increasing N_B helped to reduce bias in MLE and Renew, but not in AI-SGD;
- CEE and CUEE failed because B is much larger than n_b.