COMPUTATIONAL ALGORITHMS FOR DISTRIBUTED REGRESSION ANALYSIS BASED ON SAS SOFTWARE

Sentinel

Qoua L Her, PharmD, MSc, Yury Vilk, PhD, Jessica Young, PhD, Zilu Zhang, MSc, Jessica Malenfant, MPH, Sarah Malek, MPPA, Sengwee Toh, ScD

Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA, USA

BACKGROUND & OBJECTIVES

Background:

- Distributed regression analysis (DRA) is a privacy-protecting analytic method that performs regression analysis with only summary-level data from participating sites **(Figure 1)**
- Feasibility and utility of DRA have been well documented [1]
- No DRA applications in SAS, the statistical software used by several national distributed data networks (DDNs), are available for routine use
- SAS/IML can be used to perform DRA computations, but not all data partners in national DDNs have access to SAS/IML, as it is licensed separately from SAS

Objective: To develop a DRA application using only BASE SAS and SAS/STAT modules for use in national DDNs

METHODS

We used a distributed iteratively reweighted least squares (IRLS) algorithm to perform distributed linear and logistic regression analysis and a distributed Newton-Raphson (NR) algorithm to perform distributed Cox proportional hazards regression analysis

- Algorithms were implemented using only BASE SAS and SAS/STAT modules
- The main steps in the algorithms include:
  - Compute summary data at each data partner **(Figure 2)**
  - Combine site-specific summary data at the analysis center
  - Execute PROC REG with SSCP-type input to solve the IRLS/NR system of equations
- A simulated horizontally partitioned DDN of three data partners and an analysis center was created to test the algorithms **(Figure 3)**
- PopMedNet, a secure distributed data sharing software, was used to transfer the summary data in the simulated DDN [1]

We used two different datasets to test the DRA application

- **Distributed linear and logistic regression:** "Boston Housing data," included 506 observations of medium housing prices and neighborhood characteristics [2]
  - Data was randomly partitioned among data partners (n_1 = 172, n_2 = 182, n_3 = 152)
  - Outcome: continuous housing price and dichotomized housing price (below or above median)
  - Covariates: crime, industrialization, and distance to employment centers
- **Distributed Cox proportional hazards regression:** "Maryland State Prison data," included 432 convicts followed for one year post release and baseline characteristics [3]
  - Data randomly partitioned among data partners (n_1 = 134, n_2 = 149, n_3 = 149)
  - Outcome: time to re-incarceration (weeks)
  - Covariates: financial aid, age, and number of prior convictions

Figure 1: Distributed regression analysis

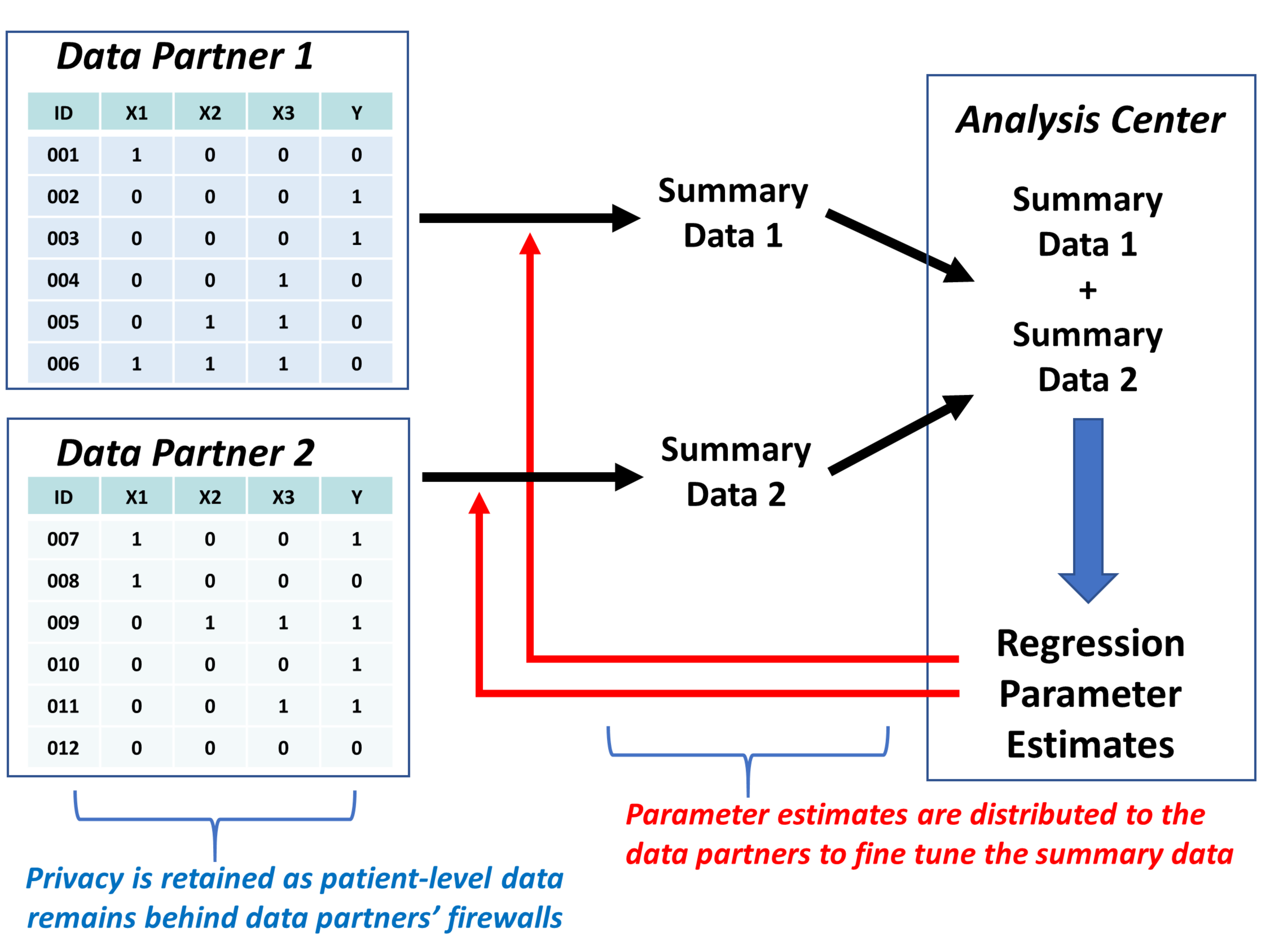

Table 1: Distributed Regression Analysis (DRA) vs. Pooled Patient-Level Regression Analysis

Linear Regression (Boston Housing data)

Covariates	DRA Estimates	DRA Standard Errors	Pooled Patient-Level Estimates	Pooled Patient-Level Standard Errors	Differences in Parameter Estimates	Differences in Standard Errors
Intercept	35.50548	1.57690	35.50548	1.57690	-8.38E-13	2.26E-14
Crime	-0.27283	0.04401	-0.27283	0.04401	4.44E-16	9.92E-16
Distance	-1.01582	0.23259	-1.01582	0.23259	1.09E-13	3.22E-15
Industry	-0.73017	0.07229	-0.73017	0.07229	3.54E-14	1.32E-15

Logistic Regression (Boston Housing data)

Covariates	DRA Estimates	DRA Standard Errors	Pooled Patient-Level Estimates	Pooled Patient-Level Standard Errors	Differences in Parameter Estimates	Differences in Standard Errors
Intercept	2.49660	0.49057	2.49660	0.49060	1.33E-15	9.99E-16
Crime	-0.14465	0.03686	-0.14460	0.03690	2.04E-13	-2.97E-14
Distance	-0.14105	0.06976	-0.14100	0.06980	1.38E-14	-2.22E-16
Industry	-0.13889	0.02376	-0.13890	0.02380	2.42E-14	1.94E-09

Cox Proportional Hazards Regression (Maryland State Prison data)

Covariates	DRA Estimates	DRA Standard Errors	Pooled Patient-Level Estimates	Pooled Patient-Level Standard Errors	Differences in Parameter Estimates	Differences in Standard Errors
Age	-0.06692	0.02084	-0.06692	0.02084	-1.39E-16	2.78E-17
Financial Aid	-0.34644	0.19024	-0.34644	0.19024	2.22E-16	-2.78E-17
Prior Arrest	0.09653	0.02724	0.09653	0.02724	-1.80E-16	1.73E-17

RESULTS

- The DRA SAS application produced regression parameter and standard error estimates within machine precision to the corresponding pooled patient-level data analyses produced by standard SAS procedures **(Table 1)**

CONCLUSION

- We successfully developed a DRA application using only SAS BASE and SAS/STAT modules
- The application may facilitate the adoption of DRA in national DDNs

REFERENCES

1. Her, QL., JM. Malenfant, S. Malek, Y. Vilk, J. Young, L. Li, J. Brown, and S. Toh. 2018. "A query workflow design to perform automatable distributed regression analysis in large distributed data networks." *EGEMS (Wash DC)* no. 6 (1):11
2. Harrison, David, and Daniel L Rubinfeld. 1978. "Hedonic housing prices and the demand for clean air." *Journal of environmental economics and management* no. 5 (1):81-102.
3. Rossi, Peter H, and J Patrick Henry. 1980. "Seriousness: A measure for all purposes." *Handbook of criminal justice evaluation*:489-505.

ACKNOWLEDGEMENTS/DISCLOSURES

- This work was supported by the Office of the Assistant Secretary for Planning and Evaluation (ASPE) and the Food and Drug Administration (HHSF223201400030I/HHSF22301006T).
- The authors have no conflicts of interest to disclose

Figure 2: Summary data example (linear regression)

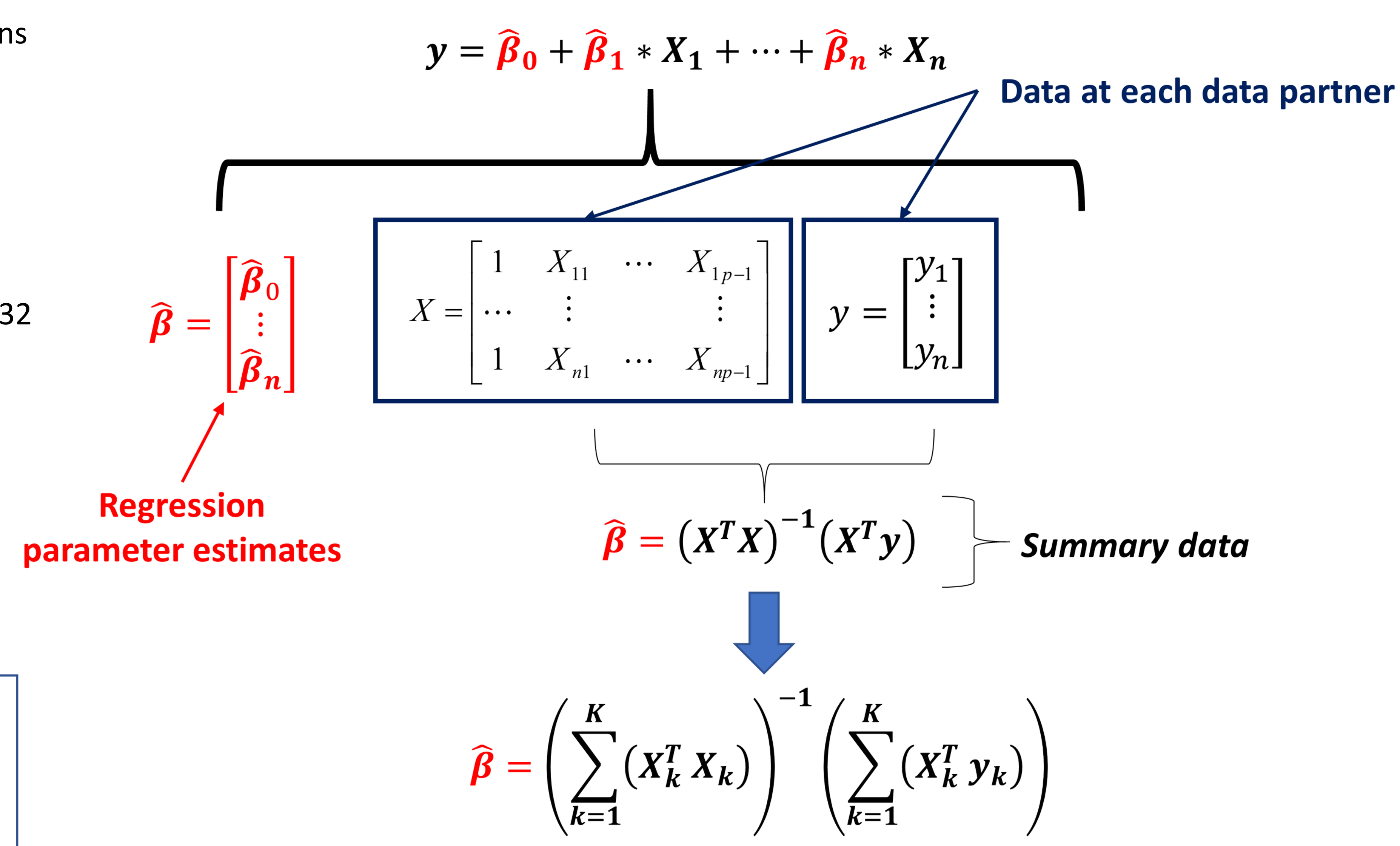

Figure 3: Simulated distributed data network