

KANG MINSU

Data Enthusiast

Master of Science
Dept. Statistics and Actuarial Science
Soongsil University

Phone : 010-3103-7944
E-Mail : kms950216@gmail.com
Github : <https://github.com/minsu0216>

INTRODUCTION



Profile

강민수 KANG MINSU

1995.02.16

서울특별시 금천구 금하로 816

송실대학교 일반대학원 정보통계보험수리학과 **이학석사**
(2019.09 - 2021.08)

송실대학교 자연과학대학 정보통계보험수리학과 **학사**
(2013.03 - 2019.08)

Research Interest

Predictive Modeling, Big Data Analytics, Data Visualization

CONTENTS

1

PCA, Clustering, Prediction, Visualization

Validity Evaluation and Modeling for Colorimetric Sensor Array

2

NRI, IDI, cNRI, dAUC, Bootstrap Confidence Intervals, Type I Error, Power

Comparison of Evaluation Index for improving model performance using Bootstrap Confidence Intervals

3

Linear Regression, Normal Equation, QR Factorization, Numerical Stability, RHadoop

Numerically Stable Algorithms estimating Linear Regression Coefficients in RHadoop

Validity Evaluation and Modeling for Colorimetric Sensor Array

Purpose

- 개발된 비색센서 어레이의 소화기암 관련 VOC 검출 가능 여부 검증
- 주어진 자료를 통해 예측 모형 구축
- 분석 결과와 새로운 자료에 대한 분석 및 예측 결과 시각화 툴 개발

Data

- 총 90개의 VOC를 어레이에 각각 5번 반복하여 노출시켰을 때, 어레이의 35개 spot의 색 변화 데이터
- VOC 종류(1-90), spot 번호(1-35), 색(R,G,B), 반복 회차(1-5)

$r_{ijk}, g_{ijk}, b_{ijk}$
 $i = 1, 2, \dots, 5$ (number of iteration)
 $j = 1, 2, \dots, 35$ (number of spot)
 $k = 1, 2, \dots, 90$ (VOC type)

Issues

- 1) VOC 노출과 무관한 체계적인 편향에 의한 RGB 값의 변화 관측 (특정 spot에서 큰 변동 관측, 밝기의 정도가 동일하지 않음)
- 2) 원 데이터는 실제 환자의 날숨이 아닌 VOC를 각각 노출 시켰을 때의 데이터
- 3) 예측 모형의 목적에 따라 성능 평가 지표의 중요도 차이 존재

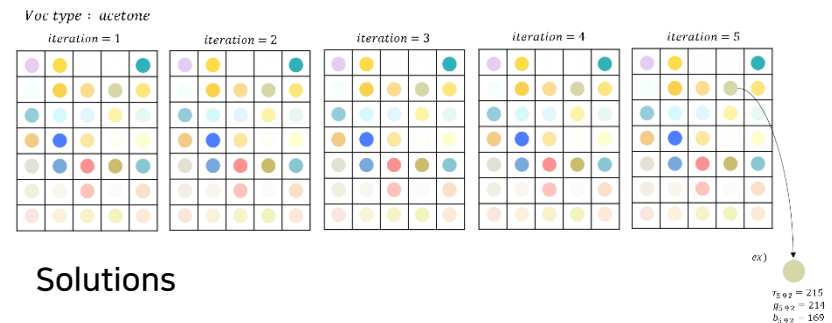
Methodology

- PCA 후 k-means, 계층적 군집분석
- LDA, Logistic Regression, SVM, XGBoost 모형 이용
- R Shiny 이용한 시각화 웹페이지 개발 후 Docker로 배포

Environment

R 3.6.1

shiny; shinydashboard; shinymanager; plotly; dendextend
 RColorBrewer; MASS; xgboost; e1071

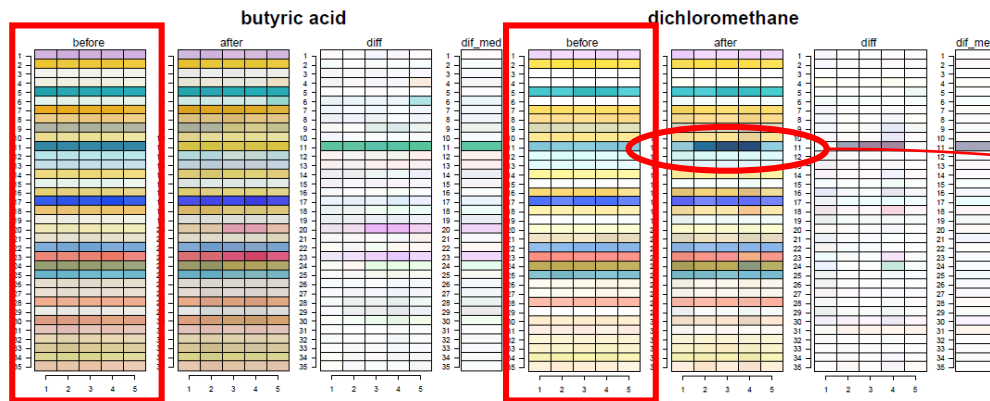


Solutions

- 1) 문제가 있는 spot의 이미지를 읽어와 전체 spot에서의 RGB 값의 평균값으로 보정(품질 관리), 밝기를 보정한 RGB 값으로 변환(RGB 정규화)
- 2) 암 환자와 일반인의 호기가스에서 나오는 VOC의 상대 비율을 이용해 암 환자와 일반인 데이터 생성
- 3) 진단의 목적인 경우, 양성, 음성이라고 예측한 사람들 중 실제로 양성, 음성인 사람의 비율이 중요하다고 판단(PPV, NPV)

Validity Evaluation and Modeling for Colorimetric Sensor Array

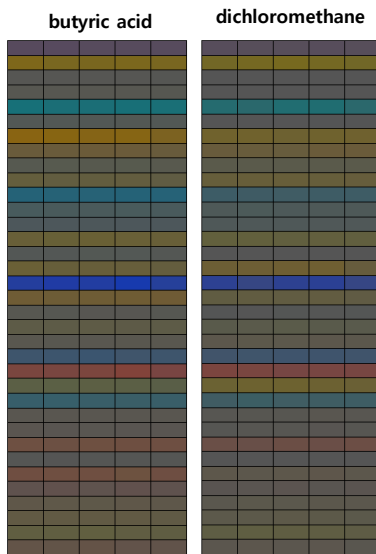
1) 소화기암 관련 VOC 검출 가능 여부 검증



반복 관측치마다 변동이 큰 spot 보정

X 축 : number of iteration ($i = 1, 2, \dots, 5$)
 before : 5번의 반복실험에서의 노출 전
 after : 5번의 반복실험에서의 노출 후

Y 축 : number of spot ($j = 1, 2, \dots, 35$)
 diff : 5번의 반복실험에서의 노출 전/ 후의 차이
 diff_med : 5번의 반복실험에서의 노출 전/ 후 중앙값의 차이



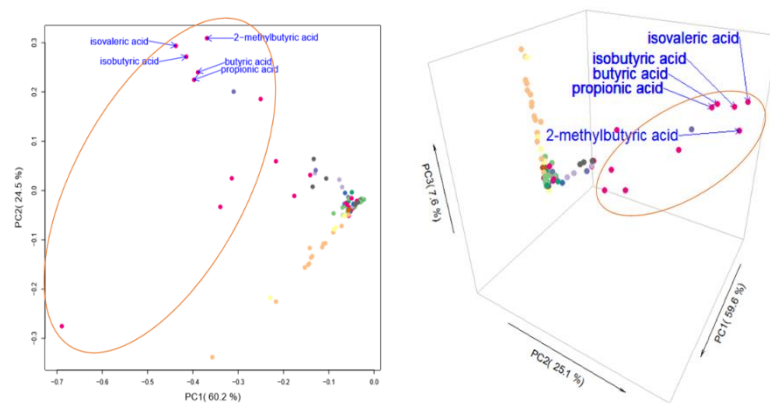
	iteration 1	iteration 2	iteration 3	iteration 4	iteration 5
실제 이미지					
보정 전					
보정 후					

밝기를 보정하기 위해 RGB 정규화

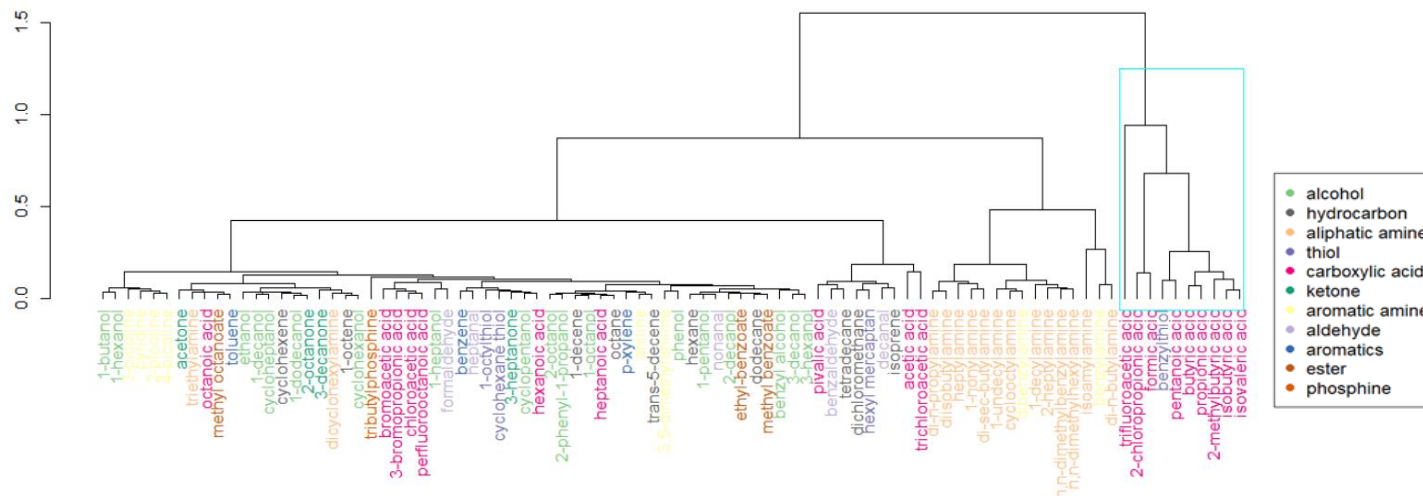
$$\left(r' = \frac{r}{r+g+b}, \quad g' = \frac{g}{r+g+b}, \quad b' = \frac{b}{r+g+b} \right)$$

Validity Evaluation and Modeling for Colorimetric Sensor Array

1) 소화기암 관련 VOC 검출 가능 여부 검증



주성분 분석 후 2개의 주성분과 3개의 주성분을 사용했을 때의 산점도



세개의 주성분을 이용하여 계층적 군집분석을 수행한 결과
소화기 암과 관련된 VOC가 하나의 군집으로 묶이는 것을 볼 수 있다.

Validity Evaluation and Modeling for Colorimetric Sensor Array

2) 임의 데이터 생성 후 모델링

① 자료 생성

$$r^*_{jk} = \tilde{r}_{jk} + normal(0, \hat{\sigma}_{r_{jk}c})$$

$$g^*_{jk} = \tilde{g}_{jk} + normal(0, \hat{\sigma}_{g_{jk}c})$$

$$b^*_{jk} = \tilde{b}_{jk} + normal(0, \hat{\sigma}_{b_{jk}c})$$

$j = 1, \dots, 35$ (number of spot), $k = 1, \dots, 10$ (voc type), $c = 1, 2, 3$ (cluster)

$normal(\mu, \sigma)$: 평균이 μ , 표준편차가 σ 인 정규분포에서 난수 생성

② 호기 가스 상대 비율과 선형 결합

암환자와 일반인을 구분하기 위해 사람의 호기 가스에서 나오는 VOC별 선형 결합

$\underline{w}_n = (w_{1n}, w_{2n}, \dots, w_{10n})'$: 일반인의 호기 가스 상대 비율 n = 일반인

$\underline{w}_c = (w_{1c}, w_{2c}, \dots, w_{10c})'$: 암환자의 호기 가스 상대 비율 c = 암환자

voc.name	w_n	w_c
Isoprene	0.7414	0.7414
Acetone	1	1
Dichloromethane	0.0138	0.0138
ocatane	0.069	0.069
Hexane	0.069	0.069
toluene	0.0862	0.0862
Propionic acid	0.0017	0.4828
Butyric acid	0	0.0345
Isovaleric acid	0.0017	0.0016
2-Methylbutyric acid	0	0.0138

③ Euclidian distance로 변경

$$\underline{d}_n = (d_{1n}, d_{2n}, \dots, d_{35n})'$$

$$\underline{d}_c = (d_{1c}, d_{2c}, \dots, d_{35c})'$$

$$d_{jn} = \sqrt{R_{jn}^2 + G_{jn}^2 + B_{jn}^2} + normal(0, \xi)$$

$$d_{jc} = \sqrt{R_{jc}^2 + G_{jc}^2 + B_{jc}^2} + normal(0, \xi)$$

$j = 1, \dots, 35$ (number of spot), n = 일반인, c = 암환자

ξ = noise level (0,0.1,1,5,10)

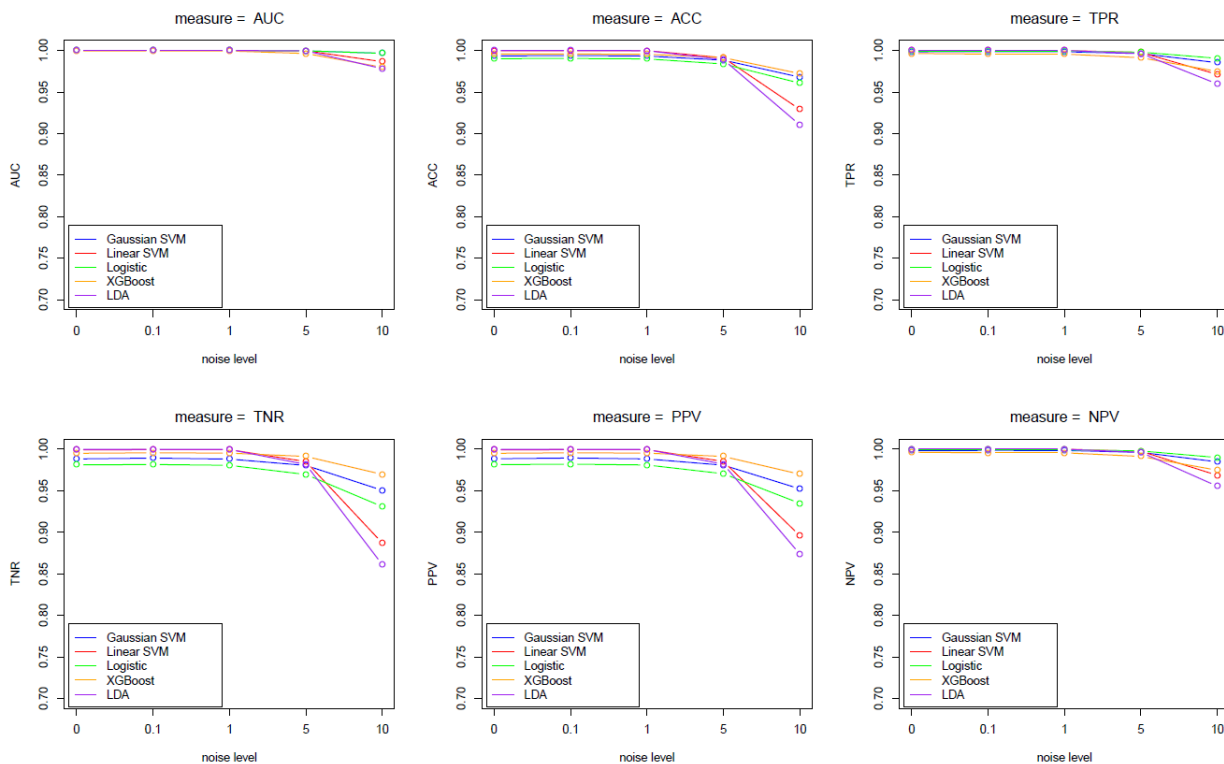
$$r^*_{j1} * w_{1n} + r^*_{j2} * w_{2n} + \dots + r^*_{j10} * w_{10n} \equiv R_{jn}$$

$$r^*_{j1} * w_{1c} + r^*_{j2} * w_{2c} + \dots + r^*_{j10} * w_{10c} \equiv R_{jc}$$

Validity Evaluation and Modeling for Colorimetric Sensor Array

2) 임의 데이터 생성 후 모델링

performance

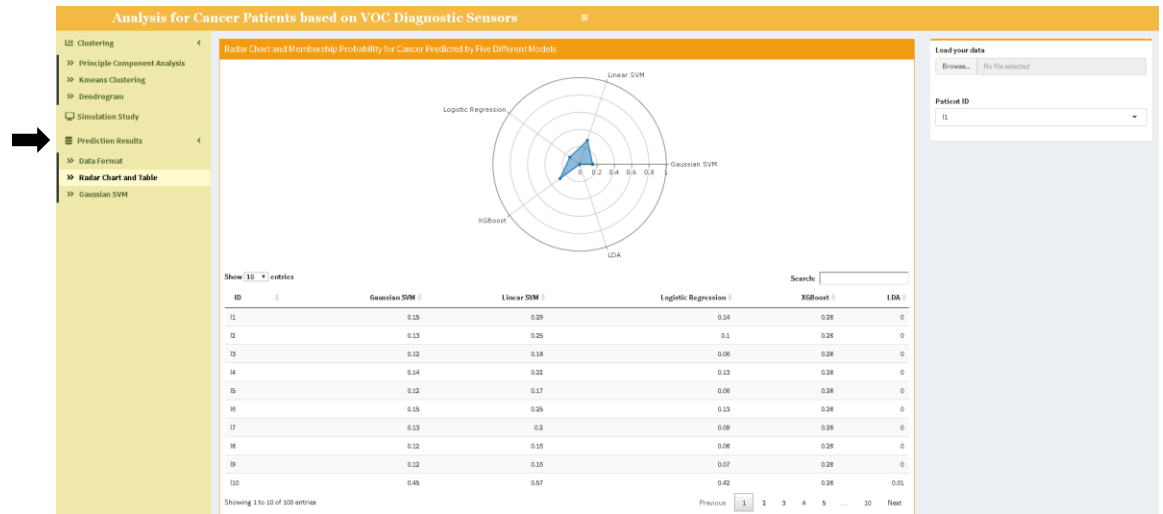
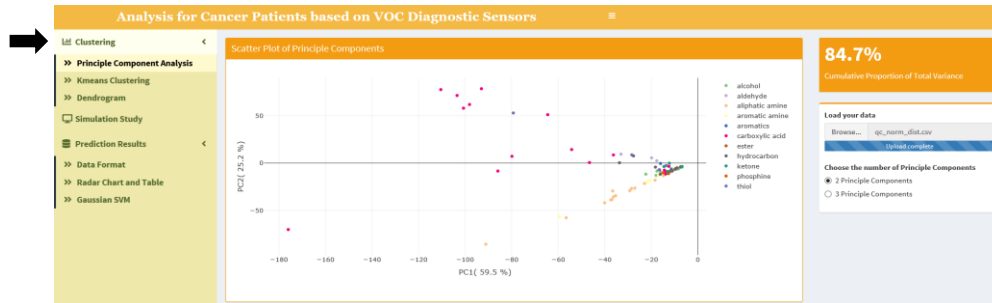


모델링 결과

임의로 생성한 자료를 사용했기 때문에 전반적으로 모든 모형이 좋은 성능을 보였다.

Validity Evaluation and Modeling for Colorimetric Sensor Array

3) 시각화 툴 개발



시각화 툴 화면

연구 결과를 포함하여, 새로운 실험 자료를 업로드하면 해당 자료에 대해주성분 분석, 군집분석을 수행한 결과를 볼 수 있게 구현하였고, 환자의 호기가스를 어레이에 노출시킨 결과를 업로드하면 앞서 구축한 모델로 소화기암 여부를 판단한 결과를 확인할 수 있다.

Comparison of Evaluation Index for improving model performance using Bootstrap Confidence Intervals

Purpose

- NRI, IDI, cNRI, dAUC 지표 타당성 비교
- 각 지표에 대한 구간 추정 결과,
 - 1) 실제로 모형 성능 향상이 없을 때 있다고 판단할 확률(1종 오류)이 통제되는지
 - 2) 모형 성능 향상이 명백해 짐에 따라 얼마나 빠르게 성능 향상이 있다고 판단하는지
 - 3) 위 두가지를 판단하기 위해 추정된 신뢰 구간의 포함 확률이 적절한지 다양한 상황의 모의실험을 통해 비교

Definition

1) 지표 정의

M_1 : 기존 모형

M_2 : 새로운 바이오 마커 추가한 모형

$D = \{0,1\}$: 질병의 여부를 나타내는 이진 변수

$NRI = [P(up|D = 1) - P(down|D = 1)] + [P(down|D = 0) - P(up|D = 0)]$

$c. NRI = E\{sign(P(M_2) - P(M_1))|D = 1\} - E\{sign(P(M_2) - P(M_1))|D = 0\}$

$IDI = E\{P(M_2) - P(M_1)|D = 1\} - E\{P(M_2) - P(M_1)|D = 0\}$

$\Delta AUC = AUC(M_2) - AUC(M_1)$

Methodology

- 특정 질병과 관련된 예후 인자를 식별하기 위한 지표로 NRI, IDI, cNRI, dAUC 존재 (Pencina et al. 2011)
- 정규성 가정 하에서 가설 검정 시, 귀무가설 하에서 정규성 가정이 위배돼 제1종 오류가 너무 크거나 검정력이 매우 작은 상황 발생
→ Bootstrap 신뢰구간 사용하여 해결 (Shao et al. 2015, Olga et al. 2017)
- 포함비율(coverage probability)와 검정력 함수(power function)

$$\text{포함비율} = P(\{CI_{lower} \leq \text{모수값} \leq CI_{upper}\})$$

$$\text{검정력 함수} = 1 - P(\{CI_{lower} \leq 0 \leq CI_{upper}\})$$

2) 신뢰구간

Asymptotic - 점근적 정규성 근사를 적용한 신뢰구간

$$\hat{\theta} - z_{\alpha/2} \widehat{se}(\hat{\theta}) < \theta_0 < \hat{\theta} + z_{\alpha/2} \widehat{se}(\hat{\theta})$$

Boot I - Asymptotic 신뢰구간에서 $\widehat{se}(\hat{\theta})$ 을 붓스트랩 표준오차로 대체한 신뢰구간

$$\hat{\theta} - z_{\alpha/2} \widehat{se}^*(\hat{\theta}) < \theta_0 < \hat{\theta} + z_{\alpha/2} \widehat{se}^*(\hat{\theta})$$

Boot II - $\tau = |\hat{\theta} - \theta_0|$ 의 붓스트랩 버전인 $\hat{\tau} = |\hat{\theta} - \hat{\theta}|$ 을 사용한 신뢰구간

$$\hat{\theta} - Q_{\hat{\tau}}(1 - \alpha) < \theta_0 < \hat{\theta} + Q_{\hat{\tau}}(1 - \alpha)$$

Comparison of Evaluation Index for improving model performance using Bootstrap Confidence Intervals

1) Simulation

Scenario I

- True Model : $\log \frac{p}{1-p} = 1 - 2X_1 + 3kX_2$
 $X_i \sim N(1, 2^2)$, $i = 1, 2$
 $k = \{0, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.6, 0.8, 1\}$
- $M_1: \log \frac{p}{1-p} = \alpha_0 + \alpha_1 X_1$
 $M_2: \log \frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
- 단순한 두 모형 비교

Scenario II

- True Model : $\log \frac{p}{1-p} = -3 + 0.02X_1 + 0.9X_2 + 0.01X_3 + 0.01X_4 - 0.01X_5 + 0.2X_6 + 1.5kX_7$
 $X_1 \sim N(40, 12^2)$, $X_i \sim \text{Ber}(0.5)$, $i = 2, 6$, $X_3 \sim \text{Unif}(120, 320)$, $X_4 \sim N(110, 1^2)$
 $X_5 \sim \text{Unif}(30, 70)$, $X_7 \sim N(1, 2^2)$
 $k = \{0, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.6, 0.8, 1\}$
- $M_1: \log \frac{p}{1-p} = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_5 + \alpha_6 X_6$
 $M_2: \log \frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7$
- 각 설명 변수는 나이, 성별, 총 콜레스테롤, SBP, HDL, 흡연여부와 같은 범위를 가짐.
 → **현실에 대응하는 모형**

Scenario III

- True Model : $\log \frac{p}{1-p} = 1 + 0.5X_1 - 2kX_2 + 1.5kX_3 + 3kX_4 - 2.7kX_5 - kX_6$
 $X_i \sim N(1, 2^2)$, $i = 1, 2, 3, 4, 5, 6$
 $k = \{0, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.6, 0.8, 1\}$
- $M_1: \log \frac{p}{1-p} = \alpha_0 + \alpha_1 X_1$
 $M_2: \log \frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6$
- $k = 0$ 일 때, M_2 가 매우 **과적합**되는 양상을 보일 것이며,
 $k \neq 0$ 에서는, M_1 이 M_2 에 비해 훨씬 안좋은 모형이 될 것임

Scenario IV

- True Model : $\log \frac{p}{1-p} = 1 + 2X_1 - 0.9kX^{2k+1}$
 $X_i \sim \text{Unif}(0, 2)$, $k = \{0, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.6, 0.8, 1\}$
- $M_1: \log \frac{p}{1-p} = \alpha_0 + \alpha_1 X_1$
 $M_2: \log \frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \beta_2 X_2^{2k+1}$
- 두 모형이 **동일한 설명변수를 사용하지만,**
모형에 차이가 있는 경우

Comparison of Evaluation Index for improving model performance using Bootstrap Confidence Intervals

2) Results

k	n	dAUC						NRI						cNRI						IDI									
		TRUE	Asymptotic	CP	P	CP	P	TRUE	Asymptotic	CP	P	CP	P	TRUE	Asymptotic	CP	P	CP	P	TRUE	Asymptotic	CP	P	CP	P				
0	500	0.0000	1	0	1	0	0.0000	0.972	0.128	0.984	0.016	0.996	0.004	0.0000	0.946	0.354	0.96	0.04	0.963	0.037	0.0000	0.968	0.032	1	0	1	0		
0.01	500	0.0001	0.993	0.001	1	0	0.0007	0.787	0.138	0.995	0.009	0.999	0.002	0.0644	0.683	0.429	0.972	0.054	0.975	0.058	0.0002	0.913	0.038	1	0	1	0		
0.05	500	0.0015	0.885	0.022	0.951	0.015	0.98	0.018	0.0050	0.865	0.232	0.96	0.094	0.969	0.063	0.3119	0.669	0.77	0.936	0.371	0.966	0.367	0.0056	0.75	0.347	0.922	0.02	0.975	0.024
0.1	500	0.0060	0.929	0.482	0.941	0.44	0.948	0.448	0.0177	0.889	0.429	0.938	0.31	0.947	0.26	0.5811	0.699	0.996	0.954	0.911	0.958	0.908	0.0217	0.751	0.924	0.931	0.49	0.943	0.489
0.2	500	0.0227	0.945	1	0.945	1	0.944	1	0.0619	0.9	0.899	0.938	0.833	0.945	0.733	0.9513	0.786	1	0.943	1	0.946	1	0.781	0.765	1	0.946	1	0.946	1
0.3	500	0.0473	0.946	1	0.948	1	0.947	1	0.1203	0.922	0.998	0.948	0.984	0.96	0.94	1.1885	0.872	1	0.959	1	0.962	1	0.1536	0.777	1	0.952	1	0.948	1
0.4	500	0.0762	0.941	1	0.942	1	0.938	1	0.1836	0.905	1	0.936	1	0.945	0.995	1.3054	0.893	1	0.943	1	0.946	1	0.2346	0.758	1	0.951	1	0.951	1
0.6	500	0.1356	0.957	1	0.958	1	0.957	1	0.3045	0.905	1	0.956	1	0.965	1	1.4792	0.92	1	0.959	1	0.962	1	0.3862	0.747	1	0.946	1	0.948	1
0.8	500	0.1884	0.952	1	0.95	1	0.949	1	0.4071	0.842	1	0.946	1	0.955	1	1.5861	0.923	1	0.956	1	0.959	1	0.5076	0.742	1	0.952	1	0.953	1
1	500	0.2313	0.958	1	0.958	1	0.959	1	0.4887	0.815	1	0.953	1	0.964	1	1.6593	0.923	1	0.944	1	0.946	1	0.5986	0.767	1	0.951	1	0.948	1

k	n	dAUC						NRI						cNRI						IDI									
		TRUE	Asymptotic	CP	P	CP	P	TRUE	Asymptotic	CP	P	CP	P	TRUE	Asymptotic	CP	P	CP	P	TRUE	Asymptotic	CP	P	CP	P				
0	500	0.0000	0.999	0.001	1	0	1	0	0.0000	0.82	0.18	0.972	0.028	0.993	0.007	0.0000	0.934	0.066	0.967	0.033	0.969	0.031	0.0000	0.961	0.039	0.999	0.001	0.999	0.001
0.01	500	0.0001	0.994	0.003	1	0	0.998	0.002	0.0011	0.818	0.177	0.984	0.018	0.994	0.007	0.0237	0.969	0.059	0.981	0.027	0.982	0.025	0.0002	0.937	0.046	1	0	1	0
0.05	500	0.0035	0.859	0.006	0.998	0.006	0.997	0.01	0.0066	0.866	0.22	0.972	0.045	0.986	0.028	0.1202	0.971	0.226	0.986	0.148	0.99	0.137	0.0038	0.697	0.201	0.997	0.005	0.998	0.005
0.1	500	0.0137	0.899	0.138	0.912	0.126	0.936	0.158	0.0220	0.847	0.387	0.949	0.203	0.969	0.182	0.2376	0.921	0.825	0.935	0.546	0.957	0.531	0.0149	0.709	0.698	0.925	0.114	0.943	0.121
0.2	500	0.0486	0.922	0.841	0.918	0.867	0.918	0.884	0.0753	0.9	0.731	0.945	0.588	0.953	0.498	0.4600	0.946	0.992	0.967	0.985	0.969	0.983	0.0544	0.735	0.998	0.94	0.866	0.943	0.866
0.3	500	0.0936	0.926	0.997	0.914	0.998	0.912	0.998	0.1458	0.988	0.954	0.948	0.9	0.968	0.806	0.6561	0.932	1	0.946	1	0.957	1	0.1097	0.731	1	0.94	0.999	0.94	0.999
0.4	500	0.1398	0.919	1	0.916	1	0.914	1	0.2206	0.909	0.997	0.958	0.986	0.982	0.935	0.8248	0.94	1	0.959	1	0.963	1	0.1728	0.754	1	0.942	1	0.94	1
0.6	500	0.2200	0.93	1	0.922	1	0.926	1	0.3573	0.885	1	0.964	1	0.983	0.997	1.0824	0.937	1	0.955	1	0.957	1	0.2993	0.789	1	0.947	1	0.946	1
0.8	500	0.2796	0.92	1	0.902	1	0.923	1	0.4656	0.855	1	0.952	1	0.982	1	1.2599	0.924	1	0.944	1	0.952	1	0.4081	0.777	1	0.941	1	0.939	1
1	500	0.3213	0.917	1	0.906	1	0.926	1	0.5478	0.837	1	0.959	1	0.985	1	1.3841	0.928	1	0.954	1	0.961	1	0.4949	0.821	1	0.955	1	0.952	1

k	n	dAUC						NRI						cNRI						IDI									
		TRUE	Asymptotic	CP	P	CP	P	TRUE	Asymptotic	CP	P	CP	P	TRUE	Asymptotic	CP	P	CP	P	TRUE	Asymptotic	CP	P	CP	P				
0	500	0.0000	0.964	0.036	0.999	0.001	1	0	0.0000	0.731	0.269	0.94	0.06	0.948	0.052	0.0000	0.921	0.479	0.938	0.364	0.716	0.284	0.0000	0.981	0.419	0.995	0.005	0.999	0.001
0.01	500	0.0012	0.972	0.044	0.993	0.013	0.999	0.003	0.0031	0.736	0.138	0.931	0.077	0.943	0.071	0.0796	0.762	0.524	0.846	0.403	0.869	0.328	0.0015	0.634	0.509	0.995	0.009	0.998	0.002
0.05	500	0.0268	0.952	0.698	0.964	0.626	0.983	0.439	0.0436	0.818	0.621	0.9	0.425	0.921	0.321	0.3797	0.914	0.984	0.953	0.971	0.96	0.904	0.0349	0.746	0.995	0.967	0.674	0.981	0.419
0.1	500	0.0879	0.963	0.999	0.967	0.999	0.969	0.999	0.1425	0.948	0.984	0.94	0.945	0.939	0.828	0.6967	0.935	1	0.962	1	0.969	1	0.1240	0.766	1	0.951	1	0.966	1
0.2	500	0.2114	0.944	1	0.947	1	0.943	1	0.3508	0.815	1	0.902	1	0.925	1	1.1091	0.934	1	0.957	1	0.963	1	0.3551	0.744	1	0.94	1	0.946	1
0.3	500	0.2938	0.962	1	0.967	1	0.97	1	0.4996	0.82	1	0.906	1	0.915	1	1.3373	0.923	1	0.949	1	0.957	1	0.4946	0.766	1	0.932	1	0.938	1
0.4	500	0.3449	0.949	1	0.959	1	0.963	1	0.5996	0.674	1	0.853	1	0.878	1	1.4764	0.92	1	0.945	1	0.95	1	0.6007	0.758	1	0.934	1	0.938	1
0.6	500	0.4001	0.951	1	0.945	1	0.945	1	0.7192	0.569	1	0.76	1	0.794	1	1.6351	0.917	1	0.94	1	0.948	1	0.7244	0.749	1	0.929	1	0.933	1
0.8	500	0.4279	0.945	1	0.922	1	0.939	1	0.7864	0.496	1	0.693	1	0.749	1	1.7214	0.916	1	0.947	1	0.952	1	0.7915	0.788	1	0.935	1	0.941	1
1	500	0.4442	0.931	1	0.884	1	0.967	1	0.8289	0.43	1	0.696	1	0.749	1	1.7752	0.924	1	0.954	1	0.954	1	0.8327	0.797	1	0.942	1	0.946	1

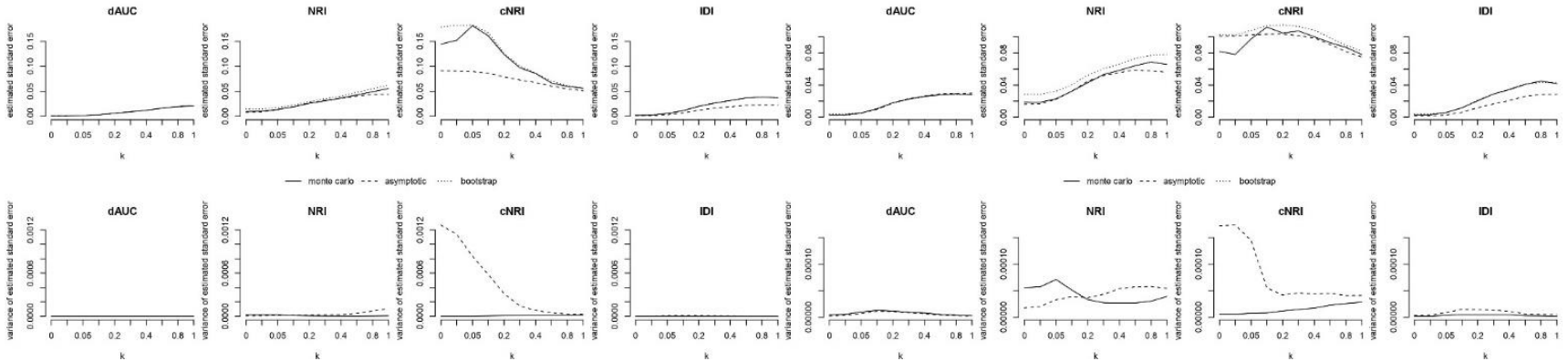
k	n	dAUC						NRI						cNRI						IDI									
		TRUE	Asymptotic	CP	P	CP	P	TRUE	Asymptotic	CP	P	CP	P	TRUE	Asymptotic	CP	P	CP	P	TRUE	Asymptotic	CP	P	CP	P				
0	500	0.0000	0.986	0.014	1	0	1	0	0.0000	0.951	0.049	0.998	0.002	0.998	0.002	0.0000	0.908	0.092	0.966	0.034	0.971	0.029	0.0000	0.934	0.066	1	0	1	0
0.01	500	0.0000	0.979	0.016	0.998	0.002	0.999	0.002	0.0001	0.27	0.48	0.94	0.006	0.999	0.001	0.0119	0.937	0.081	0.972	0.028	0.974	0.031	0.0000	0.857	0.058	1	0	1	0
0.05	500	0.0001	0.954	0.01	0.999	0.001	0.999	0.001	0.0008	0.259	0.251	0.995	0.005	0.999	0.001	0.0298	0.936	0.088	0.976	0.035	0.979	0.03	0.0002	0.248	0.065	1	0	1	0
0.1	500	0.0003	0.971	0.01	0.999	0.001	1	0.001	0.0016	0.212	0.297	0.996	0.004	0.999	0.001	0.0624	0.945	0.116	0.977	0.047	0.984	0.045	0.0006	0.276	0.1	1	0	1	0
0.2	500	0.0012	0.955	0.02	1	0	1	0	0.0019	0.44	0.116	0.993	0.008	0.999	0.002	0.1215	0.935	0.188	0.989	0.085	0.996	0.081	0.0027	0.368	0.21	1	0	1	0
0.3	500	0.0031	0.979	0.059	0.999	0.006	0.999	0.005	0.0064	0.629	0.229	0.994	0.028	0.994	0.017	0.1773	0.9	0.354	0.974	0.173	0.987	0.168	0.0069	0.477	0.465	0.994	0.005	0.996	0.005
0.4	500	0.0058	0.919	0.139	0.969	0.039	0.987	0.034	0.0086	0.734	0.329	0.976	0.064	0.98	0.039	0.2348	0.861	0.629	0.961	0.337	0.982	0.311	0.0141	0.587	0.783	0.939	0.046	0.984	0.037
0.6	500	0.0113	0.919	0.527	0.96	0.325	0.971	0.297	0.0124	0.783	0.422	0.944	0.159	0.957	0.156	0.3477	0.784	0.911	0.952	0.713	0.959	0.688	0.0674	0.695	0.946	0.622	0.61	0.951	0.54
0.8	500	0.0133	0.929	0.687	0.947	0.499	0.959	0.516	0.0197	0.811	0.517	0.949	0.215	0.967	0.225	0.4596	0.74	0.984	0.952	0.884	0.963	0.872	0.0530	0.687	1	0.941	0.954	0.946	0.931
1	500	0.0136	0.935	0.689	0.946	0.634	0.958	0.596	0.0239	0.81	0.558	0.959	0.243	0.986	0.266	0.5986	0.725	0.997	0.951	0.955	0.956	0.956	0.0604	0.666	1	0.944	0.989	0.946	0.948

Comparison of Evaluation Index for improving model performance using Bootstrap Confidence Intervals

2) Results

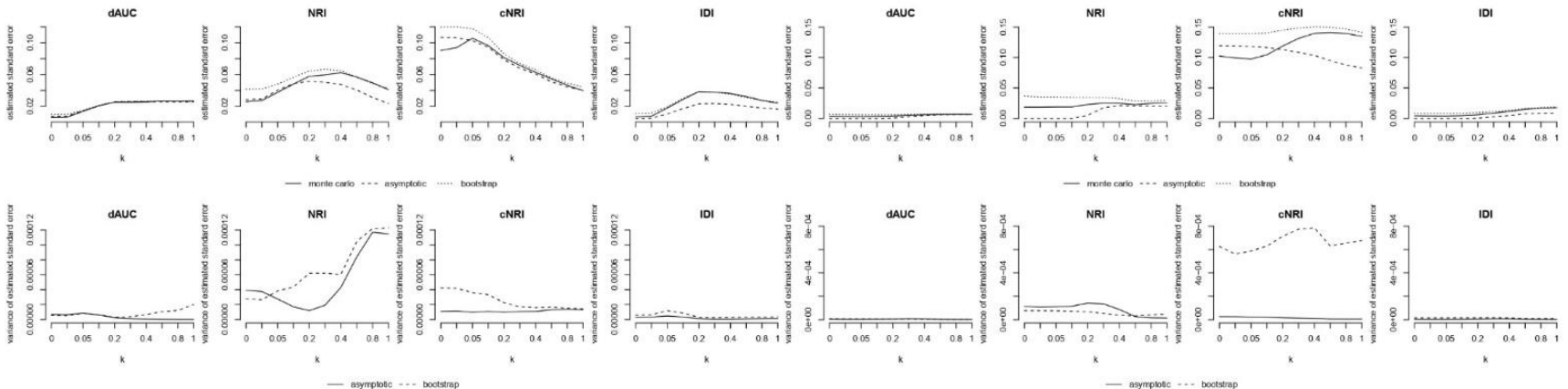
Estimated standard errors of each statistics according to k (scenario-I)

Estimated standard errors of each statistics according to k (scenario-II)



Estimated standard errors of each statistics according to k (scenario-III)

Estimated standard errors of each statistics according to k (scenario-IV)



각 지표의 monte carlo simulation을 통한 표준오차, 공식 기반의 표준오차, Bootstrap 표준오차
일반적으로 *NRI*, *cNRI*는 세 표준오차 간의 차이가 크기 때문에 앞선 모의실험에서 문제가 생긴 것으로 생각할 수 있다.

Numerically Stable Algorithms estimating Linear Regression Coefficients in RHadoop

Purpose

- RHadoop에서의 QR분해를 이용한 선형회귀계수 추정 알고리즘 개발
- 기존 병렬 Normal Equation 알고리즘과 비교

Environment

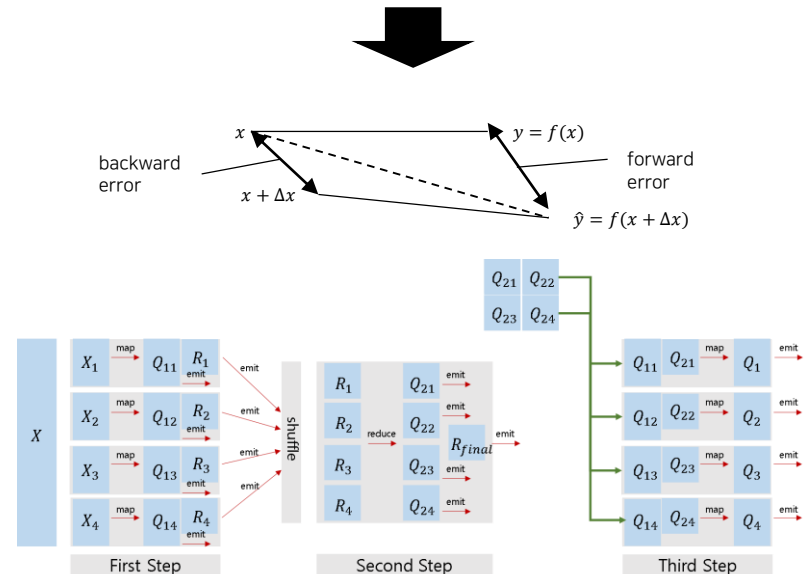
노드	master	1
	slave	4
서버 사양	CPU	Intel i7-8700K
	RAM(master/slave)	64GB / 32GB
	HDD	1TB
소프트웨어 버전	OS	Ubuntu 16.04 LTS
	Java	1.8.0
	Hadoop	2.6.0
	R	3.4.4
	rhdfs	1.0.8
	rmr2	3.3.1

Issues

기존의 Direct TSQR 알고리즘은 총 3단계의 맵리듀스 과정을 거치기 때문에 비효율적

Methodology

- Direct TSQR 알고리즘을 수정, 보완하여 알고리즘 작성
- 수치적 안정성(numerical stability)과 시스템 처리시간을 기준으로 비교

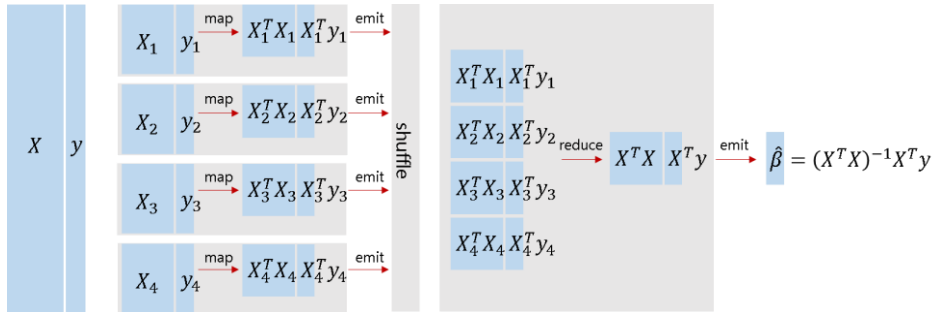


Solutions

두 단계에 걸쳐 계산되는 Q_1, Q_2 를 마지막에 결합하지 않고, **계산될 때마다 블록 별로 필요한 연산을 수행** (다음 페이지 참고)

Numerically Stable Algorithms estimating Linear Regression Coefficients in RHadoop

1) Algorithms

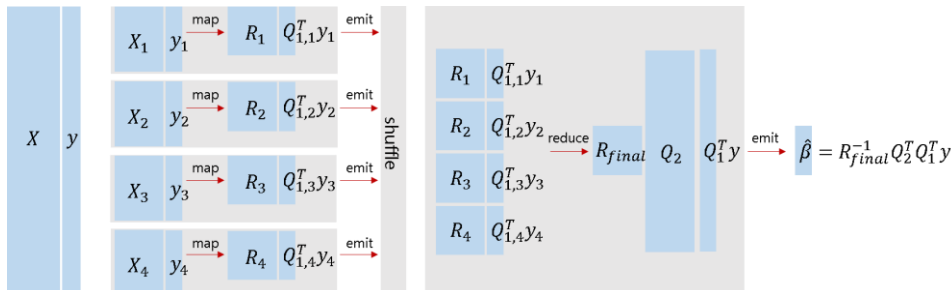
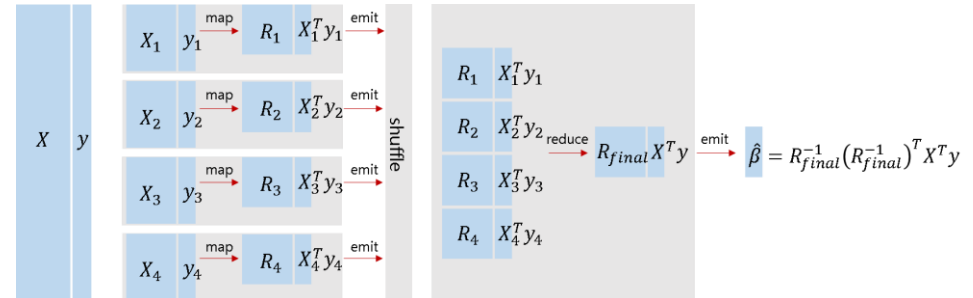


기존 병렬 Normal Equation 알고리즘

시스템 처리시간 빠르지만, 단일 컴퓨팅에서의 연산과 동일한 연산을 수행하기 때문에 수치적으로 불안정할 것으로 예상

Indirect TSQR 수정 보완한 선형회귀계수 추정 알고리즘

기존 Indirect TSQR 알고리즘이 QR의 Q를 구하는데 있어 수치적으로 불안정하기 때문에 수치적으로 불안정할 것으로 예상

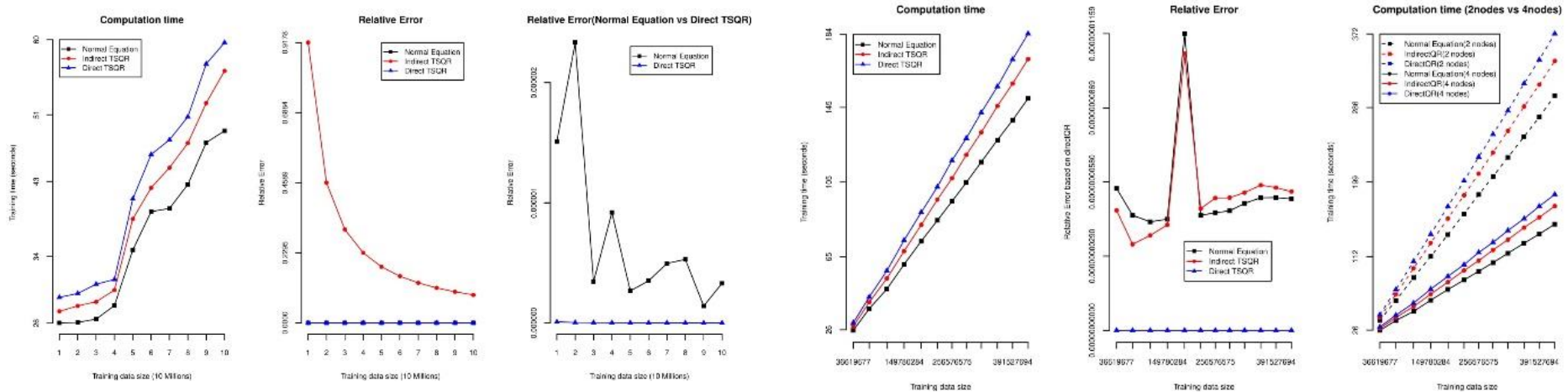


Direct TSQR 수정 보완한 선형회귀계수 추정 알고리즘

수치적으로 안정적이고 시스템 처리시간은 다른 알고리즘보다 조금 느릴 것으로 예상

Numerically Stable Algorithms estimating Linear Regression Coefficients in RHadoop

2) Results



sparse한 자료에 대한 각 알고리즘의 시스템 처리시간(좌),
상대 오차(가운데, 우)

뉴욕 옐로우 택시 자료에 대한 각 알고리즘의 시스템 처리시간(좌), 상대 오차(가운데)
노드 수 2개일 때와 4개일 때의 시스템 처리시간 비교(우)



결론

수치적 안정성 : Normal Equation \approx Indirect TSQR \ll Direct TSQR
시스템 처리시간 : Normal Equation \leq Indirect TSQR \leq Direct TSQR

Thank You

:)

Master of Science
Dept. Statistics and Actuarial Science
Soongsil University

Phone : 010-3103-7944
E-Mail : kms950216@gmail.com
Github : <https://github.com/mins0216>