プライバシーを保護した垂直分割線形回帰システムの実装とDPCデータセットを用いた評価

濱永千佳[1] 菊池浩明[1] 康永秀生[2] 松居宏樹[2] 橋本英樹[2]

概要：プライバシー保護をしつつ情報を活用する方法にはプライバシー保護データパブリッシクングや秘密分散などの様々な手法が提案されている．本研究では，準同型性公開鍵暗号を用いることで，データの価値を失うことなく活用するプライバシー保護データマイニングを取り上げ，最もシンプルな統計計算である線形回帰について実装を試みる．個人情報保護という観点から，多くの項目が個人情報となり得る医療データベースに注目し，実在する患者のデータを用いて同一の患者群についての 2 つのデータセットを生成する．各々異なるデータセットを有するユーザ間において，互いにデータセットを参照させることなく，安全に正しく線形回帰を実行することを本研究の目的とする．2 者間における垂直分割方式での秘匿計算の 3 つの方式を提案し，実装したシステムを用いて，DPC データセットを解析した．実験結果により，どの提案方式でも正確な結果を算出できることと，2 種の重回帰について提案方式のパフォーマンスを報告する．

Development of Privacy-Preserving Linear Regression System of Vertically Partitioned Data and Estimation with DPC dataset

CHIKA HAMANAGA[1] HIROAKI KIKUCHI[1] HIDEO YASUNAGA[2] HIROKI MATSUI[2]
HIDEKI HASHIMOTO[2]

1. はじめに

2015 年に個人情報保護法が改正され，マイナンバー制度も導入された．ベネッセ社で大規模な個人情報流出事件 [1] も発生しており，個人情報保護に関して，社会の関心が高まっている．プライバシー保護をしつつ情報を活用する方法には匿名加工などのプライバシー保護データパブリッシクングや秘密分散などの様々な手法が提案されているが，本研究では，準同型性公開鍵暗号を用いることで，データの価値を失うことなく活用するプライバシー保護データマイニング (Privacy Preserving Data Mining, PPDM) を取り上げる．

本研究では，医療データベースに注目する．医療データベースには，年齢，性別などの基本情報に加え，既往歴や入退院時の状況，診療情報など，多くの情報が登録されており，ほとんどの項目が個人情報と成り得るからである．本稿では，最もシンプルな統計計算である線形回帰を取り上げる．かつて [2] において，医療データベースを用いた PPDM を提案したが，ランダムに合成したデータを用いて，適用可能性を見積もるにとどまっていた．そこで，本研究では，システムを実装し，実在の患者から得られたデータを用いた実行結果と，2 種のアルゴリズムの重回帰の計算について比較を示す．

5000 例規模の脳卒中患者の実在する患者のデータを用いて，同一の患者群についての情報を持つ 2 つのデータセットを生成する．データセット A には年齢，退院時の生死などの基本個人情報を含む．データセット B には既往歴の有無や脳卒中の重症度などの診療情報を含む．両データセットは分散管理されており，平文のままで情報を持ち出すこと，活用することが認められていないという状況を想定し，加法準同型性公開鍵暗号を用いた 2 者間における垂直分割方式での PPDM を行うことを目的とする．個人情

[1] 明治大学総合数理学部
School of Interdisciplinary Mathematical Science, Meiji University
[2] 東京大学大学院医学系研究科
Graduate School of Medicine, The University of Tokyo

表 1　実験 1 心疾患データ

変数		データ数	最小・最大値	平均値	管理者
入院日数 (日)	x_1	655	0 - 101	14.56	B
年齢 (歳)	x_2	655	0 - 95	52.25	B
総入院費用 (円)	y	655	597 - 1291937	77930	A

表 2　実験 2 脳卒中データ

変数		最小・最大値	平均値	管理者
Death	y	0 - 1	0.12	A
Age	x_1	40 - 106	72.03	A
Sex	x_2	1 - 2	1.431	B
JapanComaScale	x_3	0 - 3	0.957	B
modifiedRankinScale	x_4	0 - 5	3.556	B
StrokeType	x_5	1 - 3	1.432	B
LiverDisease	x_6	0 - 1	0.022	B

報を秘匿したままで線形回帰を行う際の正確性，パフォーマンスの 2 点を明らかにすることを試み，漏洩の可能性のあるデータについても検討する．

暗号化処理，秘匿計算の実行，復号処理と傾き，切片の算出の 3 段階では，暗号化処理により多くの時間がかかることが予想される．また，秘匿計算を行うユーザがどの情報を相手に渡すかにより，秘匿計算の実行と復号処理のどちらに負担がかかるかを実験により示す．

暗号化処理に時間を要するが，線形回帰を秘匿計算で行うことにより，個人情報の漏洩の可能性のない，統計的な値のみを明らかにできることを実証する．

2. DPC

2.1 DPC データセット

DPC データセットは，病名や治療行為の表コードによる患者の大規模データベースであり，疾患，治療の組合せのデータからなる [3]．年齢，性別，疫病名，重症度，退院時転帰などの診療情報を含んでいる．DPC データベースは 255 万 3283 件のデータが登録されており，内 7 万 828 件が脳梗塞に関する情報である．

2.2 実験使用データ

本稿の実験で用いるデータとその統計量を，表 1，表 2 に示す．表 1 は擬似的に合成したもので，表 2 は実際の患者のデータである．

表 1 は心疾患の擬似患者データである．虚血性心疾患を患ったデータのみを抽出しており，年齢，入院日数が総入院費用にどの程度影を及ぼすのかを実験する．

5 章で後述する実験 2 において，表 2 に示す脳卒中の患者のデータを使用する．脳卒中の分類 1 項目，既往歴 6 項目，入院時の状態 2 項目と，患者の病状についての多くの項目の中から 7 項目を抽出し，実験を行う．年齢，性別，患者の退院時の生死を表す説明変数 Death に加えて，病歴から肝疾患 LiverDisease の有無を表す項目と入院時の病状

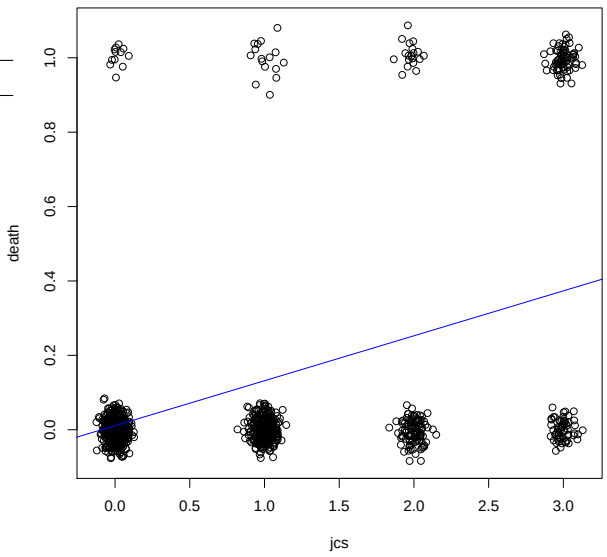

図 1　JapanComaScale

から 3 項目を選んでいる．

JapanComaScale は，入院時の意識レベルを 4 段階に分類したものであり，数字が小さいほど意識が明瞭である．modifiedRankinScale は，身体障害についての項目であり，数値が大きいほど寝たきりなど重度の障害である．StrokeType は脳卒中の病型を示しており，脳梗塞，脳内出血，くも膜下出血の 3 分類である [4]．

JapanComaScale，modifiedRankinScale それぞれの項目における，患者の割合を図 1，図 2 に示す．JapanComaScale は離散値であるが，散布を可視化するため，x，y に小さな (x に標準偏差 0.05 平均 0，y に標準偏差 0.03 平均 0) 正規乱数を加えている．例えば，JapanComaScale= 0，Death= 1 は 12 例である．modifiedRankinScale も同様である (x に標準偏差 0.01 平均 0，y に標準偏差 0.02 平均 0)．死亡 Death の数値が 0 のとき，患者は退院して生存していることを表している．

5 章の秘匿計算を用いた回帰の実験において，死亡 Death を目的変数として使用する．

2.3 データの分割について

市町村などの地方自治体とその地域の病院の間において，地域への健康の促進のため等にデータを活用するユースケースを考える．例えば，自治体などの公的機関は死亡届などから直接の原因を知ることができたとしても，既往歴や血圧等の診療情報を知ることは難しい．また，病院などの医療機関は，医療サービスのために個人情報をデータとして持っているが，個人情報保護の観点から外部への情報提供を厳しく管理しており，簡単には情報を提供できない．

そのような場合でも 2 者間で安全にデータを秘匿計算す

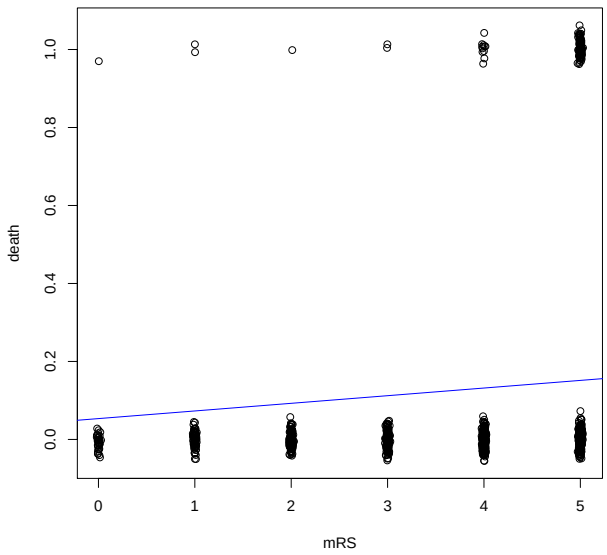

図 2　modifiedRankinScale

ることを目的とする．ユーザ A を官公庁などの公的機関，ユーザ B を病院などの医療機関として想定している．表 1，表 2 それぞれを 2 つに分け，それぞれが所有しているデータを示す．表 1 のデータは，y を持つユーザ A と x_1，x_2 を持つユーザ B が分散管理している．表 2 については，y，x_1 を持つユーザ A と，それ以外の全てのデータを持つユーザ B としており，互いに所有するデータを秘匿したいと考えているとする．

3. 従来研究

先行研究として，秘匿せずに行う線形回帰をとりあげる．藤森らは，DPC データセットから大腸の悪性腫瘍の根治的手術がなされた 8733 例を用いて，診療プロセスの分析と医療の質評価の活用を行っている（[3]，4 章診療プロセスと臨床評価，pp.57-82）．開腹手術，腹腔鏡手術の平均麻酔時間の関係を回帰で求めているが（[3]，図 4-12），個人情報を秘匿していない平文の状態で分析しているため，特異な値を持つ患者はデータから個人を特定されるリスクがある．

65 才，69 才という年齢をまとめて 60 代と一般化する匿名加工をすることで，個人を識別できないようにすることもできるが，精度を損失する．

4. 提案方式

各々異なるデータセットを有するユーザ A とユーザ B が，互いにデータセットを参照させることなく，安全に正しく線形回帰を実行することを本システムの目的とする．

加法準同型性を満たした公開鍵暗号（本研究では Paillier 暗号 [5]）を用いて，互いのデータを暗号化したままで，線

表 3　ユーザが持つ情報

	ユーザ A	ユーザ B
担当	暗号・復号	回帰計算
所持データ	$y_1, y_2, \ldots, y_n$	$x_{1,1}, x_{2,1}, \ldots, x_{n,1}$
所持暗号情報	秘密鍵 $(p, q), n$	公開鍵 (n, g)

形回帰 $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m$ を実行し，線形式の係数 α, β_1, β_2, ..., β_m を得る．

単回帰

$$y = \alpha + \beta x \tag{1}$$

2 変数の重回帰

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 \tag{2}$$

3 変数以上の多変数に対応した重回帰

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m \tag{3}$$

と表し，3 つに分けて方式を提案する．ここで，レコード数を n，説明変数 x の数を m とする．

4.1 単回帰

目的変数 y と（1）式の右辺の差の二乗の総和 $S = \sum_{i=1}^{n}(y_i - \alpha - \beta x)^2$ を求め，傾き β，切片 α について偏微分した．偏導関数を 0 として，次の方程式を立てる．

$$\begin{cases} \dfrac{\partial S}{\partial \alpha} = -2\displaystyle\sum_{i=1}^{n}(y_i - \alpha - \beta x_i) = 0 \\ \dfrac{\partial S}{\partial \beta} = -2\displaystyle\sum_{i=1}^{n} x_i(y_i - \alpha - \beta x_i) = 0 \end{cases}$$

これを解いて，傾き β は，

$$\beta = \frac{n\displaystyle\sum_{i=0}^{n} x_i y_i - \sum_{i=0}^{n} x_i \sum_{i=0}^{n} y_i}{x_i^2 - (\displaystyle\sum_{i=0}^{n} x_i)^2} = \frac{D}{C} \tag{4}$$

切片 α は，

$$\alpha = \frac{\displaystyle\sum_{i=0}^{n} x_i^2 \sum_{i=0}^{n} y_i - \sum_{i=0}^{n} x_i \sum_{i=0}^{n} x_i y_i}{x_i^2 - (\displaystyle\sum_{i=0}^{n} x_i)^2} = \frac{E}{C} \tag{5}$$

により算出される．ユーザ B は秘匿計算を用いて，A には x_i を見せずに，傾き β，切片 α の 2 値をユーザ A の持つ y の暗号文とユーザ B の持つデータで計算する．ユーザ A は復号することで傾き β，切片 α だけを得る．ユーザ各々が持つ情報を表 3 に示す．

垂直分割での計算にあたり，加算と乗算は秘匿して行ったが，$\frac{D}{C}$，$\frac{E}{C}$ は平文で求める．傾き β，切片 α は，割り切れない数をとることが多く，乗法逆元を求めて掛け算とし

Algorithm 1：scLinear(単回帰)		
	A	暗号鍵を生成
	$A \to B$	公開鍵を共有
1.	A	データ y_1, y_2, ..., y_n を暗号化.
	$A \to B$	$Enc(y_1)$, ..., $Enc(y_n)$　を送る.
2.	B	データ $x_{1,1}$, $x_{2,1}$, $\cdots$,$x_{n,1}$ から,
		$Enc(C) = n\sum x_i^2 - \sum x_i \sum x_i$
		$Enc(y)$，データ x から,
		$Enc(D) =$
		$(\prod Enc(y_i)^{x_i n})((\prod Enc(y_i))^{\sum x_i})^{-1}$
		$Enc(E) =$
		$(\prod Enc(y_i))^{\sum {x_i}^2}((\prod Enc(y_i)^{x_i})^{\sum x_i})^{-1}$
		を出力.
3.	$B \to A$	$Enc(C)$, $Enc(D)$, $Enc(E)$ を送る.
4.	A	復号し，C,D,E を求め,
		$\beta = \dfrac{D}{C}$，$\alpha = \dfrac{E}{C}$ より,
		傾き β，切片 α を求める

ても，元の数値には戻らない．従って，プログラムとして実装した秘匿計算は，和，差，積である．

この問題を解決するため，Algorithm 1 に示す単回帰の提案方式では，傾き β，切片 α についての計算を，ユーザ B が暗号文 $Enc(C)$，$Enc(D)$，$Enc(E)$ をユーザ A に渡し，ユーザ A が復号した後で，式（4），（5）より傾き β，切片 α を算出する．

4.2　2 変数の重回帰

2 変数に限定した重回帰 $y = \alpha + \beta_1 x_1 + \beta_2 x_2$ については，単回帰と同じく，偏微分を用いて解く．目的変数 y と（2）式の右辺の差の二乗の総和 $S = \sum_{i=1}^{n}(y_i - \alpha - \beta_1 x_{1i} - \beta_2 x_{2i})^2$ を求め，傾き β_1，β_2，切片 α について偏微分した．偏導関数を 0 として，方程式

$$\begin{cases} \dfrac{\partial S}{\partial \beta_1} = -2\displaystyle\sum_{i=1}^{n} x_{1i}(y_i - \beta_1 x_{1i} - \beta_2 x_{2i} - \alpha) = 0 \\ \dfrac{\partial S}{\partial \beta_2} = -2\displaystyle\sum_{i=1}^{n} x_{2i}(y_i - \beta_1 x_{1i} - \beta_2 x_{2i} - \alpha) = 0 \\ \dfrac{\partial S}{\partial \alpha} = -2\displaystyle\sum_{i=1}^{n}(y_i - \beta_1 x_{1i} - \beta_2 x_{2i} - \alpha) = 0 \end{cases}$$

を解いて，傾き β_1，β_2，切片 α を

$$\beta_1 = \frac{\sum x_{1i} y_i \sum x_{2i}^2 - \sum y_i x_{2i} \sum x_{1i} x_{2i}}{\sum x_{1i}^2 \sum x_{2i}^2 - \sum x_{1i} {x_{2i}}^2} = \frac{D_2}{C_2}$$

$$\beta_2 = \frac{\sum x_{2i} y_i \sum x_{1i}^2 - \sum y_i x_{1i} \sum x_{1i} x_{2i}}{\sum x_{1i}^2 \sum x_{2i}^2 - \sum x_{1i} {x_{2i}}^2} = \frac{E_2}{C_2}$$

$$\alpha = n\sum y_i - \beta_1 \sum x_{1i} - \beta_2 \sum x_{2i}$$

で算出する（Algorithm 2）．

4.3　多変数に対応した重回帰

我々は，過去に [2] において，水平分割方式と垂直分割方式での多変数の重回帰について提案した．本稿では，垂直分割方式について再検討し，実在する患者データでの実験と，2 変数の重回帰との比較を行った．計算方法を以下に示す．

Algorithm 2：scLinear(2 変数の重回帰)		
1.		Algorithm 1 の 1 までは同じ.
2.	B	データ x_1，x_2，$Enc(y)$ から,
		$\sum x_1$，$\sum x_2$，$\sum y$ を求め,
		$Enc(C_2) = \sum x_1^2 \sum x_2^2 - \sum x_1 {x_2}^2$
		$Enc(D_2) =$
		$(\prod Enc(y_i)^{x_{1i}})^{\sum {x_{2i}}^2}((\prod Enc(y_i)^{x_{2i}})^{\sum x_{1i}x_{2i}})^{-1}$
		$Enc(E_2) =$
		$(\prod Enc(y_i)^{x_{2i}})^{\sum {x_{1i}}^2}((\prod Enc(y_i)^{x_{1i}})^{\sum x_{1i}x_{2i}})^{-1}$
		を出力.
3.	$B \to A$	$Enc(C_2)$，$Enc(D_2)$，$Enc(E_2)$,
		$Enc(\sum x_1)$，$Enc(\sum x_2)$ を送る.
4.	A	復号し，C_2,D_2,E_2,$\sum x_1$,$\sum x_2$ を求め,
		傾き β_1，β_2，切片 α を求める.

$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m$ についても，単回帰，2 変数の重回帰と同じく，偏微分を用いる．行列式での計算を用いて，線形式の係数 α, β_1, β_2, ..., β_m を算出する（Algorithm 3）．

$S = \sum(y_i - \alpha - \beta_1 x_{1i} - \cdots - \beta_m x_{mi})^2$ を求め，それらを最小化する各係数 β を，総和の微分を 0 とおいて次の方程式を立てる．

$$\begin{cases} \dfrac{\partial S}{\partial \alpha} = -2\displaystyle\sum_{i}^{n}(y_i - \alpha - \beta_1 x_{i,1} - \cdots - \beta_m x_{i,m}) = 0 \\ \dfrac{\partial S}{\partial \beta_1} = -2\displaystyle\sum_{i}^{n} x_{i,1}(y_i - \alpha - \beta_1 x_{i,1} - \cdots - \beta_m x_{i,m}) = 0 \\ \cdots \end{cases}$$

これらを整理して，$m+1$ 個の連立方程式を

$$FX = G \tag{6}$$

と行列で表し，これを満たす X を求めればよい．ここで，

$$F = \begin{pmatrix} \sum x_{i,1}^2 & \sum x_{i,1}x_{i,2} & \cdots & \sum x_{i,1} \\ \sum x_{i,1}x_{i,2} & \sum x_{i,2}^2 & \cdots & \sum x_{i,2} \\ \vdots & & & \vdots \\ \sum x_{i,1} & \sum x_{i,2} & \cdots & \sum 1 \end{pmatrix},$$

$$X = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \\ \alpha \end{pmatrix}, \qquad G = \begin{pmatrix} \sum x_{i,1}y_i \\ \vdots \\ \sum x_{i,m}y_i \\ \sum y_i \end{pmatrix}$$

とする．F の逆行列を左からかけて，係数を得る．

4.4　評価

表 4 で，提案方式の比較を示す．多変数の重回帰につい

Algorithm 3：scLinear(多変数の重回帰)		
1.		Algorithm 1 の 1 までは同じ．
2.	B	行列 F，G を求める．
		（ここで，A のみで求められるものは計算しない．）
3.	$B \to A$	$Enc(F)$，$Enc(G)$ を送る．
4.	A	復号し，F，G を求め，
		$FX = G$ となる X を求める．

ては，2 変数の場合と比べやすくするため $m = 2$ の場合で記載している．m の場合においては，F からは $\frac{(m+m^2)}{2} - 1$ 個，G からは m 個の情報を送信している．しかし，計算の面では，合計値のみを求める多変数の重回帰に対し，係数を求める際に必要となる最小限の数値のみをユーザ A に渡す 2 変数の重回帰の方が，秘匿計算において煩雑な計算を要求する．

$m = 2$ の場合において，暗号化した状態でユーザ B がユーザ A に公開する情報を比較する．Algorithm 2 で送るデータは，C_2，D_2，E_2，$\sum x_1$，$\sum x_2$ の 5 つである．C_2，D_2，E_2 は計算途中の数値であり，$\sum x_1$，$\sum x_2$ は統計値である．一方，Algorithm 3 で送るデータは，

$$F_2 = \begin{pmatrix} \sum x_1^2 & \sum x_1 x_2 & \sum x_1 \\ \sum x_1 x_2 & \sum x_2^2 & \sum x_2 \\ \sum x_1 & \sum x_2 & \sum 1 \end{pmatrix},$$

$$G_2 = \begin{pmatrix} \sum x_1 y_i \\ \sum x_2 y_i \\ \sum y_i \end{pmatrix}$$

であり，7 つの情報を送信する．（$\sum x_1 x_2$ などの重複する情報を除いてカウントする．また，$\sum y_i$，$\sum 1$ のユーザ A のみで計算できるものは除いている．）この Algorithm 3 で公開する情報は，全て統計値である．

Algorithm 2 と 3 の違いは，計算途中の結果を公開して情報量を少なくするか，統計値を公開することで情報量を多くするかである．

5. 実験

5.1 実験目的

表 1，表 2 において，y の値をユーザ A が，それ以外の値をユーザ B が持つとしたとき，提案方法を表 5 の環境上に実装した．

システム scLinear を用いて，以下の 2 項目を評価する．

1. 本システムの計算結果の正確性．
2. 本システムのパフォーマンス．

5.2 実験方法

用意した DPC データについて，線形回帰を実施する．本研究では，2 つの実験から，提案方式について評価する．

ユーザ間でのデータのやりとりはネットワーク通信で行われることが多いが，本システムでは通信の過程を省略し，通信内容を一次ファイルに出力する手法をとっている．

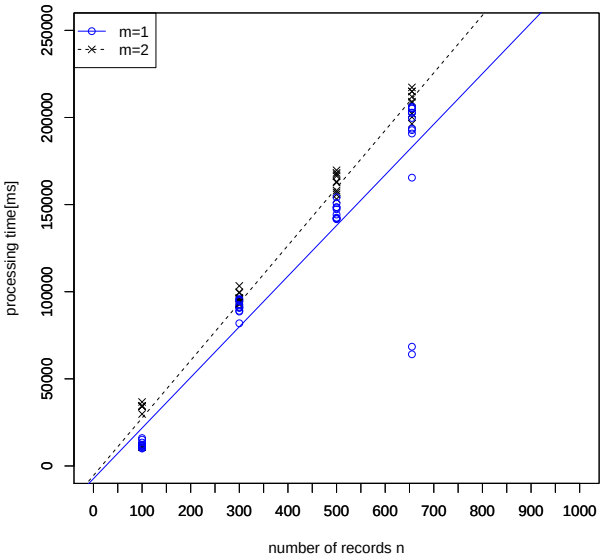

図 3 実験 1：システム実行時間

5.2.1 実験 1

用意した擬似 DPC データについて単回帰，2 変数の重回帰をそれぞれ 10 回ずつ実施する．$n = 100$ 行，300 行，500 行，655 行の 4 つの異なるデータセットについて測定する．

5.2.2 実験 2

DPC データセットについて $m = 3$，4，5，6 の重回帰を実施する．$n = 1000$ 行，2000 行，5000 行のデータセットについて測定する．

5.3 実験 1 の結果

5.3.1 正確性

表 6 に単回帰の計算結果を，表 7 に 2 変数の重回帰の計算結果を示す．

scLinear の実行結果は，単回帰においては R の計算結果と差がなかったが，2 変数の重回帰については，小数第 3 位以降に R の結果との違いが見られた．$n = 655$ の β_1，α と，$n = 300$ における α である．特に α での誤差が目立つが，本システムが小数点第 5 位まで算出して各変数を求めており，β_1，β_2 を求めた後に $\alpha = n\sum y - \beta_1 \sum x_1 - \beta_2 \sum x_2$ の算出を行っているため，β_1，β_2 でのわずかな誤差が累積していることが原因であると考えている．scLinear は，単回帰においては非常に高い精度（小数第 3 位）で，2 変数の重回帰においては小数点第 2 位までの精度を持つと分かった．

5.3.2 パフォーマンス

図 3 と図 4 に scLinear の処理時間を示す．ここで，点線は 2 変数の重回帰を，実線は単回帰の結果を示している．

表 4　提案手法と従来手法の比較

	(1) 単回帰	(2) 2 変数の重回帰	(3) 多変数の重回帰 ($m=2$)
モデル	$y=\alpha+\beta x$	$y=\alpha+\beta_1 x_1+\beta_2 x_2$	$y=\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_m x_m$
B から A へ送るデータ	C，D，E	C_2，D_2，E_2，$\sum x_1$，$\sum x_2$	F_2，G_2
暗号文数	3	5	7

表 5　実験環境

	実験 1	実験 2
OS	Windows 7	Windows 7
メモリ	4 GB	11.7 GB
CPU	Intel(R) Core(TM) i5-3337U	Intel Xeon X5460
クロック	1.8 GHz	3.16 GHz
使用言語	Java(1.8.0_91-b14) R(3.1.0)	Java(1.8.0_45-b15) R(3.1.2)
鍵長	2048[bit]	

表 6　実行結果（単回帰）

	傾き β		切片 α	
n	scLinear	R	scLinear	R
655	5099.358	5099.358	5002.521	5002.521
500	67751.810	67751.810	1021.751	1021.751
300	72369.082	72369.082	322.378	322.378
100	89508.076	89508.076	786.790	786.790

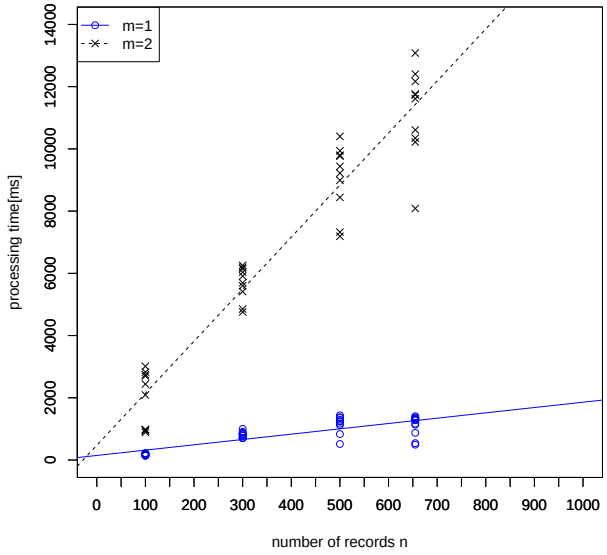

図 4　実験 1：Linear クラス実行時間

Algorithm1 の step2 を行なう Linear クラスにおいて，単回帰分析は 1 データあたりの平均処理時間が 1.9 ミリ秒であった．一方，Algorithm2 では，1 データに対する計算が 1 つ多く，秘匿計算において大きなレコードの数値（例えば，日数 50 日）での 50 乗を計算するなど，複雑な計算を要求される．1 データあたり平均 16.7 ミリ秒を要した．暗号化の処理時間には平均 320 ミリ秒と単回帰，重回帰に差が見られなかった．復号化については，単回帰は平均 1.0 ミリ秒，重回帰は平均 0.6 ミリ秒であり，大きな差はなかった（表 8）．

5.4　実験 2 の結果

5.4.1　正確性

表 9 に 6 変数の重回帰での計算結果を示す．秘匿計算を行った結果を scLinear に，R での実行結果を coefficient に示している．

scLinear の実行結果は，R の結果と差が見られなかった．また，$n=1000$，2000 においても差がなかったため，scLinear は正確に計算できていると言える．

5.4.2　パフォーマンス

図 5，図 6 に異なる条件下での scLinear の処理時間を示す．図 5 は，サイズの異なる 3 つのデータセットにおけるシステムの実行時間について，図 6 は同じデータにおける変数 m を変えた場合の，暗号化処理を除いたシステム実行時間である．ここで，一点鎖線はユーザ A の処理時間を，点線はユーザ B の処理時間を，破線はデータの暗号化にかかった時間を，実線は合計時間を示している．

図 5 より，レコード数 n に線形に，システムの処理時間が増加していることが分かる．また，図 6 より，変数 m に対して線形であり，計算に使用する変数が増加すると，比例して全体の処理時間が増加している．（図 5 におけるシステム全体の実行時間 Total は $y=224.576n+9023.692$ で，暗号化処理にかかる時間は $y=223.374n-1649.538$，n はレコード数である）．$n=100$ 万件のデータセットを実行した場合，システム全体の実行は

$$225\times10^6+9024=2.24\times10^8[\text{ms}]=2.60[\text{day}]$$

かかる．そのほとんどは暗号化処理にかかる

$$223\times10^6-1650=2.23\times10^8[\text{ms}]=2.59[\text{day}]$$

時間であり，その後の計算のみであれば約 20 分で実行できる．

5.5　実験考察

多変数に対応させた場合，ユーザ各々でデータを処理する時間が増加しているが（図 6），2 変数の重回帰では，ユーザ A の処理時間は最大でも 2000[ms] と，n に依存しない．

実験 1 において Algorithm 1 の step2 を行なう Linear クラス以外の処理時間では大きな差が見られなかったことから，Linear クラスの処理時間がシステム実行時間に大きな

表 7　実行結果（2 変数の重回帰）

	傾き β_1		傾き β_2		切片 α	
n	scLinear	R	scLinear	R	scLinear	R
655	4995.554	4995.555	41.304	41.304	3042.752	3042.759
500	998.417	998.417	245.842	245.842	54332.010	54332.010
300	302.882	302.882	128.136	128.136	65339.083	65339.084
100	4730.629	4730.629	273.939	273.939	1744.523	1744.523

表 8　実験 1：1 データあたりの平均実行時間 [ms]

モデル	暗号化処理 ユーザA	秘匿計算実行 ユーザB	復号化と算出 ユーザA
単回帰	320	1.9	1.0
2 変数の重回帰	320	16.7	0.6

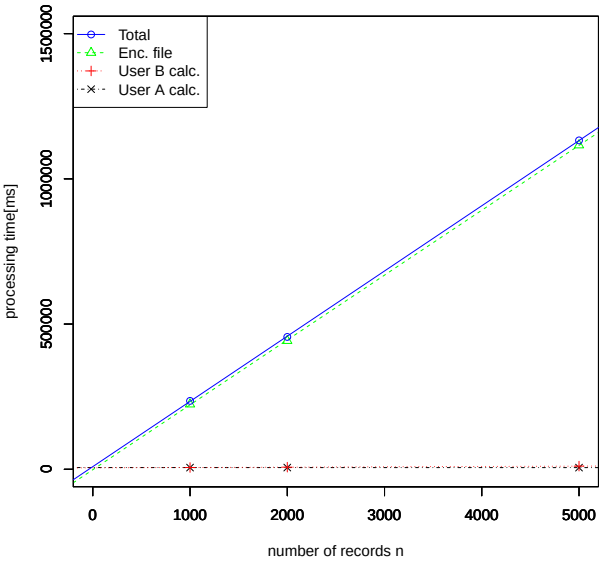

図 5　実験 2：レコード数 n におけるシステムの実行時間

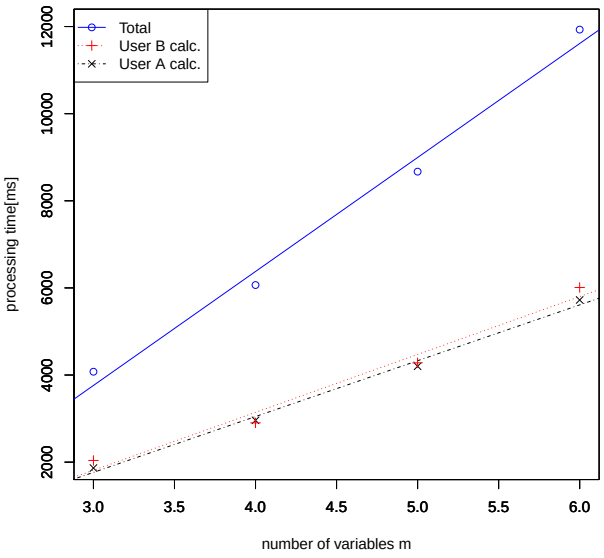

図 6　実験 2：変数 m におけるシステム実行時間（$n = 1000$，暗号化時間は除く）

影響を与えていると考えられる．2 変数の場合は，計算途中のデータをどの程度秘匿したいかによってアルゴリズムを選択して使用できると考える．

3 つのアルゴリズムを提案したが，どの計算方法でも小数点第 5 位程度までは R と非常に近い数値を求めることができていた．重回帰の場合では，ユーザ B の持つデータの内属性ごとの和をユーザ A に公開しているが，性別等のダミー変数（1 若しくは 0）の場合でも，レコード数を多くすることで，統計量となり，中間ファイルから個人を特定できない．

実験 2 より，多変数の重回帰において，死亡という目的変数に対して，正負はあるが，JapanComaScale，modifiedRankinScale，StrokeType だけでなく，年齢，性別が生死に影響を与えていると分析できた．JapanComaScale の係数は 0.1283596 であり，入院時の意識レベルが 1 大きく（より悪く）なると，死に近づきやすくなることを示している．特に，性別の係数が −0.0217865 と負の値を示しているが，1 が男性，2 が女性であることから，女性の方が死ににくいと言える．これは，男女の割合を示した図 7 からも，女性の方が長生きであり，平均年齢も高くなっていることを裏付けることができると考える．また，肝臓病 LiverDisease の傾き β が 0.0095 と小さいことから，肝臓病の既往歴が脳卒中での死亡に対してはあまり影響がないと確認できる．

図 1，図 2 に示す実線は，それぞれ JapanComaScale，modifiedRankinScale についての実験 2 で求めた傾き β，切片 α の予測式を示す．なお，他の属性値は平均値を代入しでいる．図 1 の傾きは図 2 と比べて大きい．（JapanComaScale の傾き 0.1283596），（modifiedRankinScale の傾き 0.0121227）．これは，死亡 Death に対して，JapanComaScale が modifiedRankinScale よりも大きな影響を与えていることを示している．

scLinear を用いて，脳卒中の患者の DPC データについて考察し，係数の絶対値が最大である，入院時の意識の状態 JapanComaScale が，死亡 Death という結果に対して最も支配的であると明らかにした．

表 9　線形回帰モデルの係数と提案方式の比較 ($n = 5000$)

	提案方式	R				
variables	scLinear	coefficient	Std. Error	t value	$Pr(>\|t\|)$	
α	−0.1731982	−0.1731982	0.0290099	−5.970	$2.53e-09$	***
Age	0.0015410	0.0015410	0.0003576	4.310	$1.67e-05$	***
Sex	−0.0217865	−0.0217865	0.0083993	−2.594	0.009519	**
JapanComaScale	0.1283596	0.1283596	0.0049296	26.039	$< 2e-16$	***
modifiedRankinScale	0.0121227	0.0121227	0.0034845	3.479	0.000507	***
StrokeType	0.0292522	0.0292522	0.0073582	3.975	$7.12e-05$	***
LiverDisease	0.0095770	0.0095770	0.0324591	0.295	0.767970	

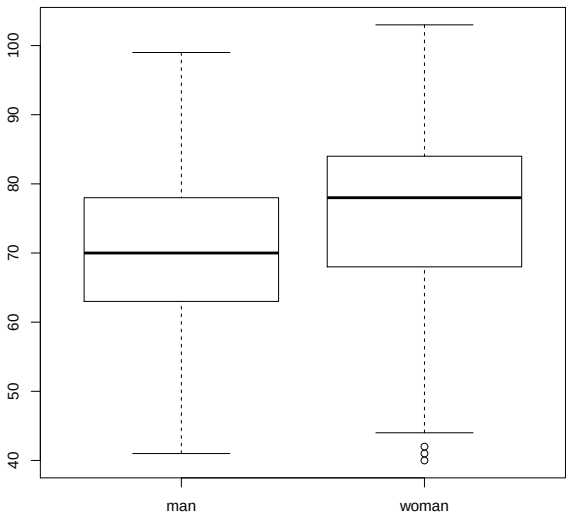

図 7　実験 2：男女の性別割合（$n = 1000$）

6. おわりに

医療機関と公的機関を例とした 2 組織間で，各々が情報を暗号化して秘匿したまま線形回帰を求めるプロトコルを実装し，その性能を評価した．3 つの提案方法を示し，実際の DPC データを用いて実験を行い，どの方法でも正確性のある結果を算出することができることを示し，パフォーマンスを評価した．特に，多変数の重回帰はデータの暗号化処理に時間を要するが，その後の計算に時間を要さない．$n = 100$ 万件とした場合，暗号化を含めると 2.6 日かかるが，その後の計算のみならば，20 分で計算できる．

しかし，統計値のみの公開であっても，相手ユーザに提示する情報があり，暗号化したままではあるが公開している．データをより安全に活用するために，相手に漏れてしまう情報の安全性の評価を今後の課題とする．

謝辞

本研究は JSPS 科研費 15K00194 の助成を受けたものです．

参考文献

[1] ベネッセホールディングス，“事故の概要”（http://www.benesse.co.jp/customer/bcinfo/01.html，2015 年 6 月参照）

[2] 菊池，橋本，康永，“DPC データベースからのプライバシーを保護した線形回帰による入院日数モデルの学習”，DICOMO2014 シンポジウム，pp．219-223，2014.

[3] 松田，伏見，“診療情報による医療評価 : DPC データから見る医療の質”，東京大学出版会．2012.

[4] 日本脳卒中学会，“脳卒中ガイドライン 2009”（http://www.jsts.gr.jp/jss08.html，2016 年 5 月参照）

[5] P. Paillier, Public-Key Cryptosystems Based on Composite Degree Residuosity Classes, EUROCRYPT 1999, pp.223-238，1999.