公的統計への秘密計算適用に向けたミクロデータの統計分析

田中 哲士[1] 阿部 穂日[2] 高橋 慧[1] 菊池 亮[1] 土井 厚志[1] 千田 浩司[1] 白川 清美[3]

概要：公的統計のミクロデータに基づく実証分析により，我が国における社会・経済状況の実態や様々な知見の獲得が期待される．しかし調査対象となった個人または法人その他の団体のプライバシー・機密に係わる情報漏洩の懸念があることから，現在，その利用は限定的である．特に懸念される情報漏洩のリスクは，(1) ミクロデータへの不正なアクセスまたは過失等による情報漏洩，(2) 分析に利用するミクロデータや分析結果からの特定の個人や法人等の識別，が挙げられる．本研究では，前記 (1) の対策として，データを暗号化したまま所望の分析を可能とする秘密計算技術の統計分析への適用，特に，公的統計の分析でよく用いられている多変量解析に対する秘密計算の適用について検討する．具体的には，多変量解析の一手法である線形回帰が，秘密計算で高速に実行できる総和や内積から計算可能であることを利用し，線形回帰を行う効果的な秘密計算手法を提案する．さらに提案手法を実装し，十分実用的な処理時間で実行可能であること，かつ分析結果を正しく得られることを実証した．

Analyzing Microdata of Official Statistics with Secure Computation

SATOSHI TANAKA[1] YUTAKA ABE[2] SATOSHI TAKAHASHI[1] RYO KIKUCHI[1] ATSUSHI DOI[1]
KOJI CHIDA[1] KIYOMI SHIRAKAWA[3]

1. はじめに

1.1 背景

2007 年 5 月に統計法が改正され，公的統計 (国や地方公共団体等の公的機関が作成する統計) は行政に限らず広く国民が利用できるようになった．そして公益を目的とした利用者に，公的統計の作成機関が既存の調査票情報（個々の調査票のデータであり一般にミクロデータという）から新たな集計表を作成・提供したり，匿名化されたミクロデータ (統計法では匿名データという) を提供する等，ミクロデータの二次的な利用が進み，我が国の社会・経済状況の実態に関する多様かつ高度な分析・研究の促進が期待されている [1].

我が国の公的統計のミクロデータ利用として，オーダーメード集計 [2]，オンサイト利用 [3]，匿名データの利用 [4]，一般用ミクロデータの利用 [5] 等がある．これらは何れも，調査対象となった個人または法人その他の団体 (以降，「個人等」と略記する) のプライバシー・機密に配慮し，提供する情報の制限やセキュリティの対策が施されている．例えばオンサイト利用は，ミクロデータの使用を情報セキュリティが確保された施設内に限定している．また匿名データは，特定の個人等の識別（他の情報との照合による識別を含む）ができないようにミクロデータを加工しており，元のミクロデータと比べ提供する情報を制限している．

一方，ミクロデータ利用の利便性を向上させるため，リモートアクセスの活用も検討が行われている [6]．リモートアクセスにより場所を問わずミクロデータの分析を即時に行うことが可能となれば，ミクロデータの二次的な利用の更なる加速が期待できる．しかしインターネット等を介したリモートアクセスの場合，不正アクセスや設定ミス等の過失，更にはシステムの脆弱性による情報漏洩の対策に十分配慮する必要がある (課題 1)．また，分析に利用するミクロデータや分析結果から特定の個人等が識別されないようミクロデータまたは分析結果を加工することも不可欠である (課題 2)．特定の個人等の識別を防ぐ加工は匿名データ等でも行われているが，即時に処理するための自動化や，課題 1 と合わせて課題解決する手法についても検討しなければならない．

[1] 日本電信電話株式会社 NTT セキュアプラットフォーム研究所
[2] 独立行政法人 統計センター
[3] 国立大学法人 一橋大学 経済研究所

1.2 本研究の結果

既存の情報漏洩対策ソリューションは数多くあるが，完全な対策は存在せず，個別の対策を組み合わせてリスクを軽減する方法が一般的である．我々は前記課題 1 の一対策として，データを暗号化したまま所望の分析を可能とする秘密計算技術の統計分析への適用を検討した．特に公的統計の分析でよく用いられる多変量解析に着目し，多変量解析の基本的な手法である線形回帰について，秘密計算によって即時処理が可能なことを実証した．秘密計算の適用により，たとえ分析時であってもミクロデータは常に暗号化されているため，不正アクセスやシステムの脆弱性等により意図せずデータが持ち出されたとしても，ミクロデータは暗号化された状態であるため被害を抑止できる．

本研究の具体的な結果は以下のとおりである．

- 秘密計算で高速に実行できる総和や内積から線形回帰が計算可能であることを利用し，線形回帰を行う効果的な秘密計算手法を提案した．
- 前記提案手法について，既存の秘密計算システム [9] を拡張して実装し，十分実用的な処理時間で実行可能であること，かつ分析結果も正しく得られることを実証した．

2. 準備

本研究では前節で述べたとおり，公的統計の分析でよく用いられる多変量解析の秘密計算を十分実用的な処理時間で実行可能とすることを目的とし，方式検討および実装評価を行う．実装評価の一例として，独立行政法人統計センターが提供している擬似ミクロデータ [10] を用いて，線形回帰の利用例である Chow 検定 [12] による構造変化テストを実行した．本節では，本研究に関連する線形回帰，構造変化テスト，そして秘密計算について説明する．

2.1 線形回帰

複数の属性を持つデータセットに対して何らかの統計的な処理を行い，分析を行うことを多変量解析と呼ぶ．複数の属性を持つデータに対して統計分析を行う多変量解析の代表的な手法として，回帰分析やクラスタリングが挙げられる．本研究では，多変量解析の代表的な手法の一つである線形回帰に着目した．以下に線形回帰について簡単に説明する．

N 個のレコードを持つ数値属性 $\boldsymbol{y} := \{y_1, \ldots, y_N\}$ を考える．このとき，同様に N 個のレコードを持つ，k 種類の数値属性 $X := \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}(\boldsymbol{x}_i := \{x_{i,1}, \ldots, x_{i,N}\})$ を用いて，

$$\boldsymbol{y} = w_0 + w_1\boldsymbol{x}_1 + \cdots + w_k\boldsymbol{x}_k = w_0 + \sum_{i=1}^{k} w_i\boldsymbol{x}_i, \quad (1)$$

で表現される線形回帰モデルにより，$\boldsymbol{y}$ をもっとも良く説明できるようにパラメータ $\boldsymbol{w} := \{w_0, \ldots, w_k\}$ を選択することを線形回帰と呼ぶ．

実際には，推定値 $\widetilde{\boldsymbol{y}} := \{\widetilde{y_1}, \ldots, \widetilde{y_N}\}$ が最小となるように $\boldsymbol{w}$ を選択する．このとき，$\widetilde{y_j} := w_0 + \sum_{i=1}^{k} w_i x_{i,j}$ である．特に，$\boldsymbol{y}$ と $\widetilde{\boldsymbol{y}}$ 間の二乗誤差の総和である残差平方和 SSE,

$$\mathrm{SSE} := \sum_{j=1}^{N} (y_j - \widetilde{y_j})^2 = \sum_{j=1}^{N} \left(y_j - w_0 - \sum_{i=1}^{k} w_i x_{i,j} \right)^2, \quad (2)$$

を最小にする手法を最小二乗法と呼ぶ．

最小二乗法では，SSE に対する各 w_i の偏微分値 $\frac{\partial \mathrm{SSE}}{\partial w_i}$ が 0 になるような w_i を選択すれば良く，以下の正規方程式，

$$X'^T X' \boldsymbol{w} = X'^T \boldsymbol{y}, \quad (3)$$

$$N \begin{pmatrix} 1 & \overline{x_1} & \ldots & \overline{x_k} \\ \overline{x_1} & \overline{x_1^2} & \ldots & \overline{x_1 x_k} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{x_k} & \overline{x_1 x_k} & \ldots & \overline{x_k^2} \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_k \end{pmatrix} = N \begin{pmatrix} \overline{y} \\ \overline{x_1 y} \\ \vdots \\ \overline{x_k y} \end{pmatrix},$$

を解けば良い．ここで，$X' := \{\boldsymbol{1}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ であり，$\overline{x_i}$ は，$\boldsymbol{x}_i$ の平均値である．これにより，$\boldsymbol{w}$ は，

$$\boldsymbol{w} = (X'^T X')^{-1} X'^T \boldsymbol{y}, \quad (4)$$

によって計算される．

求めた $\boldsymbol{w}$ が $\boldsymbol{y}$ を良く説明できているかを評価する手法として，赤池情報量規準（AIC） [13] による評価が知られている．線形回帰において AIC は，

$$\mathrm{AIC} = N\left(\log\left(2\pi \frac{\mathrm{SSE}}{N}\right) + 1\right) + 2(k+2), \quad (5)$$

で算出できる．このとき，AIC が小さい程，$\boldsymbol{w}$ は $\boldsymbol{y}$ を良く説明できていると評価する．

2.2 構造変化テスト

構造変化テストはある 2 つのデータセットの構造，即ち，回帰モデルが同じものであるか，異なるものであるかを判定する検定である．構造変化テストの代表的な手法である Chow 検定では，$k+1$ 種類の属性 $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k, \boldsymbol{y}\}$ を持つ 2 つの集合 D_1 と D_2 および 2 つの集合をまとめた長さ N の集合 D_{1+2} における $\boldsymbol{y}$ と線形回帰による推定量 $\widetilde{\boldsymbol{y}}$ の残差平方和 SSE_1, SSE_2, SSE_{1+2} を用いる次の式，

$$F = \frac{\mathrm{SSE}_{1+2} - (\mathrm{SSE}_1 + \mathrm{SSE}_2)}{\mathrm{SSE}_1 + \mathrm{SSE}_2} \cdot \frac{N - 2k}{k} \quad (6)$$

が統計量 F となることを利用する．即ち，帰無仮説「D_1 と D_2 が同じ集合である」に対して，F を用いて F 検定を行う．

2.3 秘密計算

M 個のデータを持つ平文 $\boldsymbol{a} := \{a_1, \ldots, a_M\}$ を，各 a_i に対して暗号化した暗号文 $[\![\boldsymbol{a}]\!] := \{[a_1], \ldots, [a_M]\}$ を考える

($[a_i]$ は a_i の暗号文)．このとき，$\boldsymbol{a}$ を入力とする関数 $f(\boldsymbol{a})$ と $[\![\boldsymbol{a}]\!]$ を入力とする処理の手順 $F([\![\boldsymbol{a}]\!])$（プロトコルと呼ぶ）において，$F([\![\boldsymbol{a}]\!])$ の結果が $f(\boldsymbol{a})$ の暗号文となっている（即ち，$F([\![\boldsymbol{a}]\!]) = [f(\boldsymbol{a})]$），かつ，$F([\![\boldsymbol{a}]\!])$ から結果以外の一切の情報が得られない場合，f の秘密計算方式と呼ぶ．

任意の関数 f に対して，秘密計算方式となるプロトコル F を実現するような暗号化手法の一つとして，秘密分散に基づく方式（秘密分散ベースの秘密計算と呼ぶ）が知られている [7], [8]．秘密分散は，あるデータ a を n 個のシェア $[a]_1, \ldots, [a]_n$ と呼ばれるデータの断片に分割し，各 $[a]_i$ を別々のサーバに分散して配置する方式である．このとき，個々の $[a]_i$ からは，元の a について何も判らず，k 個以上のシェアが集まった場合にのみ a に復元できるようにシェアを生成する．このようシェアの生成を行う場合，秘密分散は k 個以上のサーバが結託し，a に対する復元を試みない限り情報理論的に安全である．

秘密分散ベースの秘密計算において，ボトルネックとなるのは通信の回数である．秘密分散ベースの秘密計算では，シェアが正しい計算結果を保つためにサーバ間で適度に通信を行う必要がある．更に，この通信はサーバ間で同期して行わなければならない．従って，サーバ間の通信は，サーバ内でシェアに対して行う処理よりも重く，通信回数が実行時間に対して支配的となる．

通常，秘密分散ベースの秘密計算では，加算の秘密計算に 0 回, シェアと平文の乗算（定数倍）の秘密計算に 0 回，シェア同士の乗算の秘密計算に 1 回の通信が必要となる．また，ベクトル同士の内積についても，1 回の通信が必要となる事が知られている．これは，同じタイミングに独立に計算できる乗算の通信をまとめて行うことで，1 回の通信で処理できるようになるためである．

3. 関連研究

秘密計算に関する既存の研究では，関数 f を限定せず汎用的に実現可能な方式と，f を限定した方式がある．秘密計算は一般に，f を通常計算する場合と比べ，処理時間が増加する．そのため，f を限定した高速化の研究も数多く行われている．本節では，線形回帰を実現する既存の秘密計算の実装例を紹介する．

秘密計算を用いた回帰分析の実装においては，最も基本的な方式である線形回帰の実装例が知られており，Bogdanov らによる線形回帰の実装 [11] や Lu らの線形回帰の実装 [17] が存在する．

Bogdanov らの手法は，バイナリ演算やソーティング等の秘密計算方式を組み合わせて，行列式を用いた逆行列計算，Gauss の消去法，共益勾配法といった線型方程式の解法で用いられるアルゴリズムの秘密計算方式を実現している．実装では，秘密計算システムの一つである Sharemind [14] を用いており，10,000 レコードのデータセットに対する 1 属性の線形回帰（線形単回帰）を 3.8 秒で処理可能である．

Lu らの手法は，反復法により $({X'}^T X')$ の逆行列を計算し，求めた逆行列を用いて線型方程式を解き，パラメータを算出している．この方式は，完全準同型暗号を用いて実装され，adult と呼ばれる 6 属性× 32,561 レコードのデータセットに対して 870 秒で線形回帰の処理が可能である．

4. 提案方式

4.1 線形回帰の秘密計算

平均値 $\overline{x_i}$ から，各 $x_{i,\ell}$ に関する情報は漏れないと仮定する．同様に，内積値 $N\overline{x_i x_j}$ に対しても，各 $x_{i,\ell}$, $x_{j,\ell}$ に関するに関する情報が漏れないことを仮定する．

式 3 から，正規方程式は各属性の総和および内積値から構成できる．従って，線形回帰は実際には次の 2 つのステップに分けることができる．

(1) $N\overline{x_i}$ $N\overline{y}$ および $N\overline{x_i x_j}$, $N\overline{y x_i}$ を算出し，$({X'}^T X)\boldsymbol{w} = {X'}^T \boldsymbol{y}$ を構成する．

(2) 線形代数により，$\boldsymbol{w}$ を求める．

本提案方式では，両方のステップを秘密計算で実装するのではなく，ステップ 1 についてのみ秘密計算方式で求め，求めた総和と内積値を平文に復号し，ステップ 2 は平文上で高速に処理し，$\boldsymbol{w}$ を求める．

この方式では，データセットの総和と内積値が公開されるため，総和と内積値については情報が漏れることになる．しかし，前記の安全性に関する仮定により，X の各要素 $x_{i,\ell}$ を漏らさずに，最小二乗法による線形回帰を実現できる．

具体的なプロトコルを以下に示す．

プロトコル 1 線形回帰の秘密計算

パーティ: クライアント P，n 台の秘密計算サーバ $\{S_1, \ldots, S_n\}$. 各 S_t は k 属性× N レコードのデータセットのシェア $[\![X]\!]_t$ と，N レコードの属性のシェア $[\![\boldsymbol{y}]\!]_t$ を持つ．

出力: P は各属性の総和値 $N\overline{x_i}$, $N\overline{y}$, 属性間の内積値 $N\overline{x_i x_j}$, $N\overline{y x_i}$, 線形回帰のパラメータ $\boldsymbol{w}$ を得る．

1: 各 S_t は他の秘密計算サーバと総和の秘密計算を用いて，総和のシェア $[N\overline{x_i}]_t$, $[N\overline{y}]_t$ を求める．
2: 各 S_t は他の秘密計算サーバと内積値の秘密計算を用いて，内積値のシェア $[N\overline{x_i x_j}]_t$, $[N\overline{y x_i}]_t$ を求める．
3: 各 S_t は P に対して，シェア $[N\overline{x_i}]_t$, $[N\overline{y}]_t$, $[N\overline{x_i x_j}]_t$, $[N\overline{y x_i}]_t$ を送信する．
4: P は受け取ったシェアから，総和 $N\overline{x_i}$, $N\overline{y}$ および内積値 $N\overline{x_i x_j}$, $N\overline{y x_i}$ を復元する．
5: P は復元した総和と内積値を用いて，式 3 の正規方程式 $({X'}^T X')\boldsymbol{w} = {X'}^T \boldsymbol{y}$ を構築する．
6: P は Gauss の消去法等を用いて正規方程式を解き，パラメータ $\boldsymbol{w}$ を求める．

4.2 Chow 検定の秘密計算

Chow 検定では残差平方和が必要となるが，パラメータと残差列から元のデータに関する情報が漏れる恐れがあ

るため，残差列を直接出力することはできない．そこで，式 2 の変形をすると，

$$
\begin{aligned}
SSE &= \sum_{j=1}^{N} (y_j - \widetilde{y_j})^2 \\
&= \sum_{j=1}^{N} (y_j - w_0 - \sum_{i=1}^{k} w_j x_{i,j})^2 \\
&= N\overline{y^2} - 2\sum_{i=1}^{k} w_j N\overline{yx_i} \\
&\quad + \sum_{i=1}^{k}\sum_{\ell=1}^{k} w_i w_\ell N\overline{x_i x_\ell}, \qquad (7)
\end{aligned}
$$

となる．即ち，残差平方和は線形回帰の秘密計算で得られる，総和，内積値とパラメータから算出できる．従って，以下に示すプロトコルを用いる事で，$x_{i,\ell}$ を漏らさずに残差平方和を計算できる．

プロトコル 2 残差平方和の秘密計算

パーティ: クライアント P，n 台の秘密計算サーバ $\{S_1, \ldots, S_n\}$. 各 S_t は k 属性× N レコードのデータセットのシェア $[\![X]\!]_t$ と，N レコードの属性のシェア $[\![\boldsymbol{y}]\!]_t$ を持つ．

出力: P は残差平方和 SSE を得る．

1: P はプロトコル 1 を用いて，各 S_t と通信しながら線形回帰の秘密計算を行い，総和 $N\overline{x_i}$, $N\overline{y}$, 内積値 $N\overline{x_i x_j}$, $N\overline{yx_i}$ および パラメータ $\boldsymbol{w}$ を得る．

2: P は総和・内積値およびパラメータを用いて，式 7 から，残差平方和 SSE を求める．

これにより，Chow 検定の秘密計算は次のプロトコルで実現できる．

プロトコル 3 Chow 検定の秘密計算

パーティ: クライアント P，n 台の秘密計算サーバ $\{S_1, \ldots, S_n\}$. 各 S_t は k 属性× N_1 レコードのデータセットのシェア $[\![X_1]\!]_t$, N_1 レコードの属性のシェア $[\![\boldsymbol{y}_1]\!]_t$ と，k 属性× N_2 レコードのデータセットのシェア $[\![X_2]\!]_t$, N_2 レコードの属性のシェア $[\![\boldsymbol{y}_2]\!]_t$ を持つ．

出力: P は帰無仮説 「$\{X_1, \boldsymbol{y}_1\}$ と $\{X_2, \boldsymbol{y}_2\}$ は等しい」に対する検定結果を得る．

1: P はプロトコル 2 を用いて，各 S_t と通信しながら，残差平方和 SSE_1, SSE_2, SSE_{1+2} を得る．

2: P は残差平方和と $N := N_1 + N_2$, k を用いて，式 6 から統計量 F を求める．

3: P は F が棄却域に入る場合は帰無仮説を棄却する．そうでない場合は，帰無仮説を受理する．

5. 実験

本研究では，独立行政法人統計センターが提供している教育用擬似ミクロデータ [10]（擬似ミクロデータ）に対して，秘密計算による構造変化テストを実行し，結果の正しさの確認および処理時間の計測を行った．擬似ミクロデータは，統計局の平成 16 年全国消費実態調査 [16] を基に作成された擬似データセットであり，32,027 世帯の支出と収入に関する 197 の属性データを擬似乱数を用いて生成しまとめたものである．

表 1 実験環境（管理サーバ，秘密計算サーバ）

	管理／秘密計算サーバ
OS	CentOS 6.4(仮想マシン)
CPU	Intel Xeon E3-1220 v5 (3.00GHz, 4 コア)
メモリ	6GB
ホスト OS	CentOS 7
ホストメモリ	8GB

表 2 実験環境（クライアント端末）

	クライアント
OS	Windows 7 (64bit)
CPU	Intel Corei7-6600U （2.60 GHz, 2 コア）
メモリ	16GB
使用言語	R-3.1.0

5.1 実装環境

本研究では，秘密計算を用いた統計分析処理の実装に桐淵らが開発した秘密計算システム [9] を用いた．本システムは，データの暗号化に秘密分散を用いている秘密計算のシステムで，3 台の秘密計算サーバ，1 台の管理サーバ，そして，複数台のクライアント端末で構成される．本システムの全体図を図 1 に示す．

この秘密計算システムの分析では，はじめに分析者が数値計算ソフトである R を利用し，クライアント端末を通じて管理サーバに分析コマンドを送信する．その後，管理サーバが秘密計算サーバに対して，秘密計算サーバに秘密計算処理の命令を送り，全秘密計算サーバは通信しながら秘密計算の処理を行う．最終的に，各サーバが持つ計算結果のシェアを，管理サーバを通じてクライアント端末に送信し，分析者は分析の結果を R 上で取得する．

また，分析者は取得した結果を R を用いて更に深く分析することができる．本研究では，秘密計算システムが出力する平均値および内積値を利用して，正規方程式を構成し効率的な解決を図っている．

本実験における，秘密計算システムの管理サーバおよび秘密計算サーバの構成を表 1 に，クライアント端末の構成を表 2 に示す．各サーバは，それぞれ別個の CentOS7 サーバ上に，CentOS 6.4 の仮想マシンとして構築されている．また，各機器は 1Gbps (Mbp) の LAN で接続されている．

5.2 教育用擬似ミクロデータの分析

5.2.1 線形回帰

擬似ミクロデータに対して，集合全体および 3 人以下の有業人員毎の世帯人員 1 人当たりの実支出額の対数値に対

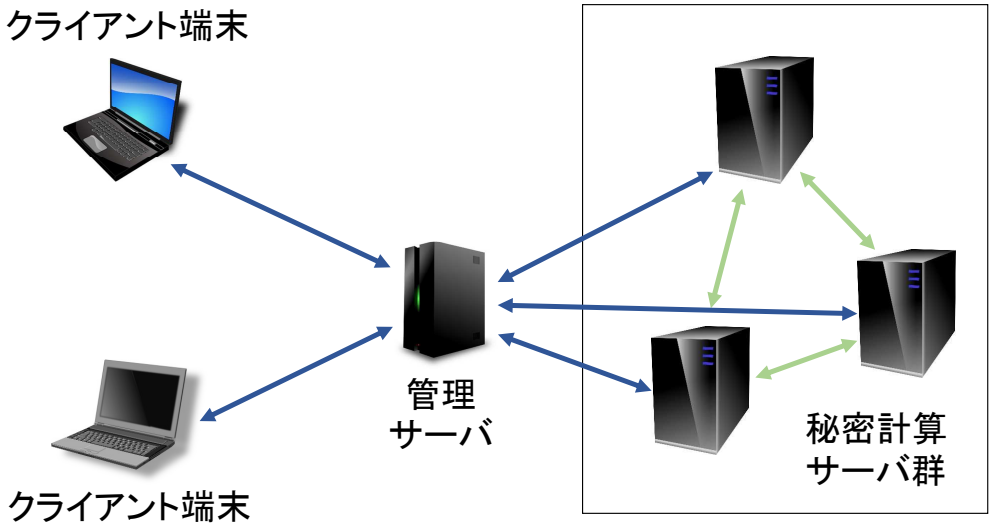

図 1 秘密計算システム [9] の全体図

表 3 実支出額に対する収入総額の線形回帰

有業人員	世帯数	パラメータ		残差平方和	AIC	実行時間
		収入総額	切片			
全体	32,027	0.852	0.981	2313.3	6731.2	1.02s
1 人	13,913	0.851	1.012	992.9	2762.1	1.07s
2 人	13,459	0.864	0.831	947.0	2480.7	1.05s
3 人	2,950	0.854	0.953	224.9	786.1	0.46s

する，世帯人員 1 人当たりの収入総額の対数値の線形回帰（線形単回帰）を行った．

この実験においては，線形回帰を行う前に除算により世帯人員 1 人当たりの実支出額を計算し，更に対数変換を行う必要がある．しかし，秘密計算を用いた除算や対数計算は計算コストが高いことが知られており，秘密計算システムでは提供されていない．そのため，そのままでは線形回帰は実現できなかった．そこで，本実験では事前計算として擬似ミクロデータを秘密計算システム上に登録する前に，世帯人員 1 人当たりの収入総額と実支出額の対数値を算出し追加のデータとして登録を行った．線形回帰の実行時間，導出されたパラメータと，そのときの残差平方和および AIC の値を表 3 に示す．

実行時間においては，いずれの場合においても 1 秒で実行可能であり，実用的な時間で処理ができているといえる．また，各パラメータに対して，秘密計算上で算出した結果 w^* と平文による計算で算出した回帰結果 w の相対誤差の絶対値 $|\frac{w^*-w}{w}|$ を計算したところ，最大で 1.67×10^{-8} であった．従って，秘密計算上でも十分に精度の良い回帰計算ができているといえる．

5.2.2 構造変化テスト

有業人員 1 人, 2 人, 3 人の世帯集合間に対して，有意水準を 5%として Chow 検定を用いた構造変化テストを行い，それぞれの集合の線形モデルが同じ構造を持つか判定した．判定の結果を表 4 に示す．表 4 から，有業人員が 2 人と 3 人の世帯集合において p 値が有意水準 5%を上回っており，この 2 つは同じ構造を持つ．

表 4 有業人員ごとの構造変化テスト

有業人員		統計量 F	p 値	実行時間
1 人	2 人	22.059	2.677×10^{-10}	5.15s
1 人	3 人	6.729	0.0012	3.65s
2 人	3 人	0.353	0.7025	3.58s

実行時間は最大でも 5 秒で実行可能であり，実用的な時間で処理可能である．また，統計量 F に対して秘密計算で算出した値と平文による計算で求めた値の相対誤差の絶対値は，最大で 1.78×10^{-7} であり，十分な精度で検定を行えているといえる．

6. 考察

6.1 比較

プロトコル 1 より，本提案手法は，総和と内積値を計算する際にのみ秘密計算を行う．このとき，総和の通信回数は 0 回，内積値の計算における通信回数は 1 回である．ここで，内積値 $N\overline{x_i x_j}$, $N\overline{y x_i}$ を並列に計算することで，従って，線形回帰の秘密計算における通信は内積による 1 回と，シェアの送信による 1 回の計 2 回となる．即ち，本提案手法は $O(1)$ 回の通信を必要とする．

一方で，正規方程式を秘密計算で解決する場合，例えば，Gauss の消去法を用いるでは，行列 $(X'^T X')$ のサイズ，即ち $k+1$ ステップの乗算を含む処理が必要となる．そのため，Pivot の選択を行わない場合でも，$O(k)$ 回の通信が必要となる．従って，本提案手法は線形回帰の処理を全て秘

密計算上で実装する手法よりも通信回数が少なく，高速である．

また，本提案手法は，3 万レコードのデータに対して 1 秒で線形単回帰の秘密計算を処理できる．これは，Bogdanov らの方式よりも多いデータをより速い時間で処理できており，Bogdanov らの方式よりも高速である．

6.2 課題 1 および 2 の対策の展望

本稿ではこれまで，1 節で挙げた課題 1 の対策について検討してきた．しかし実際には，同様に 1 節で挙げた課題 2 の対策も不可欠である．課題 2 の対策として，線形回帰等の演算結果の開示基準を与える試みもみられる [18]．この例 [18] では，原則ベースとして

- 少なくとも自由度 10 を有している，
- カテゴリ変数のみに基づかない，
- 単一のユニットに基づいているもの（例えば企業一社の時系列データ）ではない，

ことが挙げられている．また，分析の対象となるある特定の集団が 10 ユニット以上ということも述べられている．しかしこのような開示基準は，データが暗号化されていることで識別できない場合がある．そこで課題 1 および 2 の対策として，各演算における開示基準を定めた上で，その開示基準を満たすかどうかデータが暗号化されていることで識別できない場合，秘密計算によって識別する手法を検討することを今後の課題としたい．

7. まとめ

本研究では，公的統計への統計分析を目的として，秘密分散ベースの秘密計算で高速に処理可能な線形回帰の手法を提案した．また，桐淵と五十嵐の秘密計算システム [9] 上で実装し，教育用擬似ミクロデータに対して実証実験を行った．いずれの実験においても，秘密計算による分析は通常の計算による統計分析と差が無く，実用的な時間で処理できることがわかった．今後の課題は，秘密計算を用いて分析結果から特定の個人・法人等が識別されるリスクへの対策を実現することである．

謝辞

本研究は，一橋大学経済研究所共同利用・共同拠点プロジェクト研究からの助成を受けたものである．

参考文献

[1] （公財）統計情報研究開発センター, 公的統計のミクロデータ利用ガイド -社会生活基本調査の匿名データを用いた分析を例として- , Sinfonica 研究叢書 No.23, 2015.
[2] 独立行政法人 統計センター — オーダーメード集計の利用 https://www.nstac.go.jp/services/order.html
[3] 独立行政法人 統計センター | オンサイト利用 http://www.nstac.go.jp/services/on-site.html
[4] 独立行政法人 統計センター — 匿名データの利用 https://www.nstac.go.jp/services/anonymity.html
[5] 独立行政法人 統計センター — 一般用ミクロデータの利用 https://www.nstac.go.jp/services/ippan-microdata.html
[6] 公的統計ミクロデータ研究コンソーシアム — オンサイトネットワークの形成, http://www.rois.ac.jp/tric/micro/moc/onsite.html
[7] Ben-Or, M., Goldwasser, S., and Wigderson, A.: Completeness theorems for non-cryptographic fault-tolerant distributed computation, Proc. of STOC '88, 1-10, 1988.
[8] Chaum, D., Crepeau, C., and Damgård, I.: Multiparty unconditionally secure protocols, Proc. of STOC '88, 11-19, 1988.
[9] 桐淵直人，五十嵐大，"秘匿した状態で処理可能なデータベースの設計", コンピュータセキュリティシンポジウム 2015, pp.419–426, 長崎，10 月，2015 年.
[10] 教育用擬似ミクロデータの開発とその利用〜平成 16 年全国消費実態調査を例として〜http://www.nstac.go.jp/services/pdf/sankousiryou2407.pdf
[11] Dan Bogdanov, Liina Kamm, Swen Laur, Ville Sokk, Rmind: a Tool for Cryptographically Secure Statistical Analysis, IEEE Transactions on Dependable and Secure Computing, IEEE, 2016.
[12] Gregory C. Chow, Tests of equality between sets of coefficients in two linear regressions. Econometrica: Journal of the Econometric Society, 1960, 591-605.
[13] Hirotogu Akaike, Information theory and an extension of the maximum likelihood principle. In: Selected Papers of Hirotugu Akaike. Springer New York, 1998. p. 199-213.
[14] Dan Bogdanov, Sharemind: programmable secure computations with practical applications, Ph.D. dissertation, University of Tartu, 2013.
[15] Max A. Little, Patrick E. McCharry, Stepehn J. Roberts, Declan AE. Costello, Irene M. Moroz, Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection, BioMedical Engineering OnLine, pp.6–23, 2007.
[16] 統計局ホームページ平成 16 年全国消費実態調査 http://www.stat.go.jp/data/zensho/2004/
[17] Wen-jie Lu, Shohei Kawasaki, Jun Sakuma, Using Fully Homomorphic Encryption for Statistical Analysis of Categorical, Ordinal and Numerical Data, the 24th annual Network and Distributed System Security Symposium (NDSS17), 2017.
[18] ESS Net SDC, Guidelines for the checking of output based on microdata research, http://neon.vb.cbs.nl/casc/ESSnet/guidelines_on_outputchecking.pdf