統計研究彙報　第76号　2019年３月　(101-114)

ロバスト回帰推定へのウェイト関数や残差尺度の影響について

和田　かず美†
野呂　竜夫‡

Consideration on the Influence of Weight Functions and the Scale for Robust Regression Estimator

WADA Kazumi
NORO Tatsuo

本研究では、欠測値の補完への利用を目的として、外れ値の影響を緩和することのできる回帰 M-推定量の繰返し加重最小二乗（IRLS: Iteratively Reweighted Least Squares）アルゴリズムを取り上げる。外れ値の影響の緩和の程度を決めるウェイト関数やその調整定数、また尺度基準には複数の選択肢があるが、その選択や調整定数の設定は文献により様々である。そこで、挙動の異なる代表的なウェイト関数である Tukey の biweight 関数と Huber のウェイト関数、よく使用される残差尺度の基準である平均絶対偏差（AAD）と中央絶対偏差（MAD）を選択し、その組み合わせで構成される四種類の推定量を算出する関数を作成した。比較可能性を確保するためにまず調整定数の基準を統一し、モンテカルロシミュレーションで計算効率と推定効率を比較することにより、各推定量の性質の違いと調整定数や収束条件の効果を確認し、調査集計への適用時に最適な設定の選択方法を明らかにした。またこの比較により、一般的に使用されている MAD 尺度の Huber のウェイト関数ではなく、AAD 尺度の Huber のウェイト関数を用いた推定量の有用性について確認できた。

キーワード：M-推定量、Tukey の biweight、Huber ウェイト、平均絶対偏差、中央絶対偏差

In this research, we aim to use Iterative Reweighted Least Squares (IRLS) algorithm of M-estimator that can alleviate the influence of outliers for regression imputation.　There are choices about weight functions that determine the influence of outliers, their tuning constants, and scale parameters; however, their selections and settings varies depending on literature.　Therefore, we select Tukey's biweight and Huber weight for the weight function, the average absolute deviation (AAD) and the median absolute deviation (MAD) for scale parameter, and prepared R functions of four estimators made by their combinations.　A Monte Carlo simulation is conducted to examine their differences together with the effect of the tuning constant and convergence condition.　We standardize the tuning constants among scales and weight functions to ensure comparability.

We confirmed the effect of those settings and clarified how to select them when applied to the statistical survey data processing.　In addition, we propose a useful estimator of Huber weight with the AAD scale, which outperforms Huber weight with MAD scale regarding computational efficiency and compares favorably regarding estimation efficiency.

Key words:　M-estimators, IRLS, Tukey's biweight, Huber weight, Average absolute deviation (AAD), Median absolute deviation (MAD)

† 独立行政法人統計センター情報技術センター技術研究開発課
‡ 総務省統計研究研修所

1.　はじめに

統計調査の集計業務では、しばしば欠測の発生は避けられず、欠測のまま集計すると結果が偏る恐れがある場合に、欠測値を何らかの値で補う補完（imputation）という処理を行う。補完には様々な方法があるが、ここでは回帰モデルにより補完値を推定する、回帰補完を前提とする。

一般に、回帰パラメータの推定には、最小二乗法（OLS: Ordinary Least Squares）が使用されるが、推定に使用するデータに外れ値が存在し、特にそれが梃子比（回帰分析の観測点ごとに説明変数のデータを変えずに、目的変数の値を１だけ変えた場合の予測値の変化量）の高くなる位置にある場合に、結果として得られる回帰パラメータに大きく影響し、問題のデータを除外したときの推定値と大きく乖離することがある。

国連欧州地域委員会（UNECE: United Nations Economic Commission for Europe）は、各国統計部局における方法論や概念の調和を促進し、データ収集やデータの品質管理に関連した経験の共有を目的として、統計的データエディティングに関するワークセッションを定期開催しており、このワークセッションでは、各国統計部局の知識共有のために、「統計的データエディティング（Statistical data editing）」と題する一連の刊行物を作成してきた。「方法論と技術（Methods and techniques）」という副題で 1997 年に刊行された第二分冊は、上述の回帰補完時の外れ値の対策として、Bienias et al. (1997) が、J. W. Tukey が 1960 年代に提唱した探索的データ解析 (Exploratory Data Analysis: EDA) に基づく抵抗回帰 (resistant regression) 法を適用した米センサス局の事例を紹介している。

抵抗回帰とは、繰返し加重最小二乗（IRLS: Iteratively Reweighted Least Squares）アルゴリズムにより得られる古典的な M-推定量で、モデルからの乖離を示す残差が大きいほど小さくなるウェイトを各観測値に付与し、回帰パラメータ推定を加重最小二乗法（WLS: Weighted Least Squares）により行う操作を何度か繰り返すことにより、外れ値の影響を緩和した回帰パラメータの推定を実現する方法である。Bienias et al. (1997) は、ウェイト関数として Tukey の biweight 関数（Beaton and Tukey 1974）、残差の尺度として平均絶対偏差（AAD: Average Absolute Deviation）を採用しているが、その理由について説明はされていない。また、理論上優れた性質を持つ代表的なウェイト関数の一つである Huber (1964) のウェイト関数は、主に中央絶対偏差（MAD: Median Absolute Deviation）との組み合わせで使用される場合が多い。MAD の使用は、位置パラメータ推定において、理論的に偏りに関してミニマックスで多くの外れ値に耐えうることが裏づけられているという点で最善であることが知られているが、回帰の場合にも MAD が良いとされる根拠は十分ではない。

本研究では、この抵抗回帰法の補完の実務への適用を目的に、ウェイト関数として Tukey の biweight 関数と Huber のウェイト関数を、残差の尺度には MAD と AAD を取り上げ、これらの組み合わせから構成される四種類の推定量について、その性質の違いと、調整定数や収束条件の影響を比較するため、乱数データを用いたモンテカルロシミュレーションにより検討を行い、実務適用に最適な設定方法を明らかにする。

第２節では、まず回帰の M-推定量やその計算のための IRLS アルゴリズム、検討対象となるウェイト関数、残差の尺度パラメータ及び調整定数の役割について説明し、そして計算アルゴリズムと併せてその収束条件について説明する。第３節では、シミュレーションに使用するデータの設計や、比較条件等について述べ、第４節においてその結果を計算効率と推定効率の観点から整理する。最後に、第５節で実用化に向けた推定量の選択と設定の方法についてまとめる。

2.　回帰 M-推定量とその計算アルゴリズム

2.1　回帰 M-推定量

Huber（1973）は、Huber (1964) で考案した位置パラメータの M-推定量の考え方をもとに、次のような回帰 M-推定量を提案した。

まず、次のような線形回帰モデルを考える。

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i = \boldsymbol{x_i}\boldsymbol{\beta} + \varepsilon_i \tag{1}$$

ここで、目的変数を $\boldsymbol{y}$、説明変数を $\boldsymbol{X}$ 、回帰係数を $\boldsymbol{\beta}$、誤差を $\boldsymbol{\varepsilon}$ とおくと、

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{x_1} \\ \vdots \\ \boldsymbol{x_n} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

で、データサイズは n、説明変数の数は p として、式（1）は次のように表現される。

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

このとき、y_i の推定値は、

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip} = \boldsymbol{x_i}\widehat{\boldsymbol{\beta}}$$

ここで誤差の尺度パラメータを $\hat{\sigma}$、回帰残差を r_i とすると、標準化残差は $e_i = r_i/\hat{\sigma} = (y_i - \boldsymbol{x_i^\top}\widehat{\boldsymbol{\beta}})/\hat{\sigma}$ により得られ、尺度共変な回帰 M-推定量は、損失関数を ρ として次式により定義される。

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\beta} \sum_i \rho\left(\frac{y_i - \boldsymbol{x_i}\boldsymbol{\beta}}{\hat{\sigma}}\right) \tag{2}$$

式 (2) を満足する $\widehat{\boldsymbol{\beta}}$ は、ρ の導関数を $\rho' = \psi$ として、次の式を解くことにより得られる。

$$\sum_i \psi\left(\frac{y_i - \boldsymbol{x_i}\widehat{\boldsymbol{\beta}}}{\hat{\sigma}}\right)\boldsymbol{x_i^\top} = 0$$

ここで、ウェイト関数を $w(e) = \psi(e)/e$, $w_i = w(e_i)$と定義して、この推定方程式は次のように表現することができる。

$$\sum_i w_i e_i \boldsymbol{x_i^\top} = 0 \tag{3}$$

2.2　IRLS

式 (3) を解いて $\boldsymbol{\beta}$ を推定するためには繰り返し計算が必要で、ここでは Holland and Welsch (1977) が推奨する繰返し加重最小二乗法（IRLS: Iteratively weighted linear least squares）を使用する。

この方法は、適当な初期値 $\widehat{\boldsymbol{\beta}}^{(0)}$ を用いて、より良い次の $\widehat{\boldsymbol{\beta}}^{(1)}$ を次式に基づき算出し、収束条件を満たすまでこれを繰り返して推定値を改善する。このとき、尺度パラメータ $\hat{\sigma}$ も併せて推定することにより、尺度共変の M-推定量を得ることができる。

$$\widehat{\boldsymbol{\beta}}^{(j)} = \widehat{\boldsymbol{\beta}}^{(j-1)} + \left\{ \boldsymbol{X}^{\top} \left[\boldsymbol{W} \left(\frac{y_i - \boldsymbol{x_i}\widehat{\boldsymbol{\beta}}^{(j-1)}}{\widehat{\sigma}} \right) \boldsymbol{X} \right]^{-1} \right\} \boldsymbol{X}^{\top} \left[\boldsymbol{W} \left(\frac{y_i - \boldsymbol{x_i}\widehat{\boldsymbol{\beta}}^{(j-1)}}{\widehat{\sigma}} \right) \left(y_i - \boldsymbol{x_i}\widehat{\boldsymbol{\beta}}^{(j-1)} \right) \right]$$

ここで、$\boldsymbol{W}$ は対角成分が $w(e_i)$、それ以外の成分が 0 となる $n \times n$ の正方行列とする。

IRLS は、Beaton and Tukey（1974）が “biweight regression fitting” と名づけ、Tukey の biweight 関数とともに考案した。尺度パラメータには H-spread と呼ばれる下側ヒンジ（中央値以下のデータの中央値）と上側ヒンジ（中央値以上のデータの中央値）の差が使用されている。この H-spread は、外れ値に強い尺度の一つで、統計ソフト R の箱ひげ図等で現在も使用されるが、データサイズが小さい場合にそれが奇数か偶数かで基準が若干変動することが知られている。Holland and Welsch (1977) は、M-推定量の解法として、ニュートン法や Huber の方法ではなく IRLS を推奨している。

2.3　ウェイト関数

Tukey の biweight 関数は式 (4)、Huber のウェイト関数は式 (5) で表され、これらの関数の形は、図 1 のようになる。

$$w(e) = \begin{cases} \left[1 - \left(\frac{e}{c} \right)^2 \right]^2 & |e| \le c \\ 0 & |e| > c \end{cases} \tag{4}$$

$$w(e) = \begin{cases} 1 & |e| \le k \\ \frac{k}{|e|} & |e| > k \end{cases} \tag{5}$$

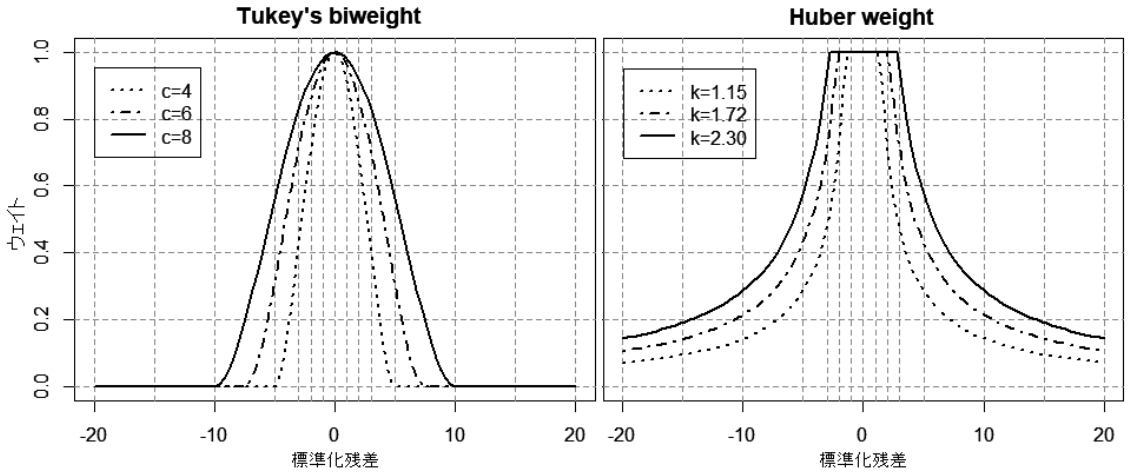

図 1. ウェイト関数の形状（和田 2012, p.28）

式 (4) の c 及び式 (5) の k は、調整定数あるいはチューニングコンスタントと呼ばれ、推定量のロバスト性をユーザーが調整するために使用する。図 1 に示すとおり、いずれのウェイト関数も、残差が 0 あるいは非常に小さい観測値のウェイトに 1 を付与するが、残差が大きくなるに従い、ウェイトは 0 に近づいていく。このとき、大きい調整定数を設定するほど、ウェイトの減少が緩やかになり、そのために推定のロバスト性は低減する。

Tukey の biweight 関数は、Beaton and Tukey (1974) が考案し、極端な外れ値のウェイトに0を付与することにより、その影響を完全に排除することができるという性質を持ち、Huber 関数と並び広く利用されている。

一方、Huber の関数を用いた M-推定量は、Huber 推定量と呼ばれる。Huber (1964) が考案し、位置パラメータの M-推定量について、蓑谷 (1992) が最小好適分布（Least favorable distribution, l.f.d）と訳した最も不利な分布を前提に、位置パラメータの M-推定量の漸近分散を最小化するというミニマックス問題を解くことにより得られた。極端な外れ値であってもウェイトが 0 にはならないために、必ず大局解に収束するという計算処理上も良い性質を持つことが知られている（e.g. Antoch and Ekblom 1995）。

2.4　尺度パラメータと調整定数

中央値を得る関数を median、平均値を得る関数を mean として、MAD と AAD による残差 r_i を標準化するための尺度パラメータは、それぞれ式 (6) 及び (7) で表現される。

$$\hat{\sigma}_{\mathrm{MAD}} = \mathrm{median}(|r_i - \mathrm{median}(r_i)|) \tag{6}$$

$$\hat{\sigma}_{\mathrm{AAD}} = \mathrm{mean}(|r_i - \mathrm{mean}(r_i)|) \tag{7}$$

回帰ではなく位置パラメータのロバスト推定について大規模なモンテカルロ実験を行った Andrews et al. (1972) は、M-推定量の尺度の基準として AAD やヒンジあるいは四分位差ではなく MAD が良いことを示した。これは理論的にも偏りに関してミニマックスであり、破局点（breakdown point）が 50% となることが裏づけられているが、回帰の場合に MAD を最善とする根拠は十分ではないとされている（Huber and Ronchetti 2009, pp.172-173）。破局点とは、推定量のロバスト性を測る指標の一つで、ある標本にどの程度の外れ値が混入したときに、推定量が無意味な値になるのかを示している（e.g. Rousseeuw and Leroy 1987, p.9）。

Huber のウェイト関数はその理論上の優れた性質から広く利用され、Holland and Welsch (1977) は、MAD を尺度パラメータとして、様々なウェイト関数を比較するモンテカルロシミュレーションを行い、Huber のウェイト関数が biweight 関数よりも推定効率が高いことを示した。

一方、Mosteller and Tukey (1977, p.358) や Holland and Welsch (1977) は、Tukey の biweight 関数の尺度パラメータとして MAD を採用しているが、Bienias et al. (1997) は AAD を採用し、収束が早く計算が簡便で容易に第三世代言語で実装可能という実用の観点から優れた性質を持つと述べている。また、Hampel (2001) は、biweight 関数について、Huber の関数を用いた推定量が optimal な状況ではわずかに劣るが、通常 Huber 推定量よりも同等か優れると述べている。

Holland and Welsh (1977) は、Tukey と Huber のウェイト関数を MAD 基準で比較しているが、尺度は SD 基準に補正しており、正規分布を前提として漸近有効性が 95%になる SD 基準の調整定数となる $c = 4.685$ と $k = 1.345$ を採用している。一方、Bienias et al. (2007) は、尺度が AAD で biweight 関数を用いた M-推定量の調整定数を、経験則から 4 から 8 の間で設定することを推奨している。これらの情報と SD、AAD 及び MAD についての以下のような関係式から、その対応を表１に示す。ここで、Φ は標準正規分布の累積分布関数とする。

$$\frac{\hat{\sigma}_{\mathrm{AAD}}}{\hat{\sigma}_{\mathrm{SD}}} = \frac{E|e|}{\sqrt{E(e^2)}} = \sqrt{\frac{2}{\pi}} \approx 0.80$$

$$\hat{\sigma}_{\mathrm{SD}} = 1/\Phi^{-1}\left(\frac{3}{4}\right) \cdot \hat{\sigma}_{\mathrm{MAD}} \approx 1.4826 \cdot \hat{\sigma}_{\mathrm{MAD}}$$

これをもとに、表 2 に示す、AAD 基準の Tukey の調整定数が $c = 4, 6, 8$ の場合に対応する各尺度基準とウェイト関数の組み合わせの調整定数の値を得た。本研究では、これを調整定数の試算条件として採用する。

表 1．調整定数の基準値

	漸近有効性 95%値		
	SD	AAD	MAD
Tukey の c	4.685	3.738	3.160
Huber の k	1.345	1.073	0.907

表 2. 調整定数の対応表

ウェイト関数	Tukey の biweight 関数			Huber のウェイト関数		
尺度	SD	AAD	MAD	SD	AAD	MAD
調整定数	5.01	4	3.38	1.44	1.15	0.97
	7.52	6	5.07	2.16	1.72	1.46
	10.03	8	6.76	2.88	2.30	1.94

2.5　計算アルゴリズム

使用している計算アルゴリズムは、和田 (2012) と同様に Bienias et al. (1997) に準拠する。処理の流れは以下のとおり。

1)　初期値算出

回帰モデル (1) に基づき下式に示す OLS で算出した回帰パラメータを、初期値 $\widehat{\boldsymbol{\beta}}^{(0)}$ として使用する。このときの残差 $r_i^{(0)}$ から、尺度パラメータ $\hat{\sigma}^{(0)}$ を算出、ウェイト関数に基づき 0 から 1 までの値をとる IRLS ウェイト $w_i^{(0)}$ を得る。w_i の値はウェイト関数に依存するが、各観測値が推定された回帰モデルから乖離するほど小さい値になるため、外れ値が回帰推定に及ぼす影響が緩和される。

$$\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} = (\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}\boldsymbol{y}$$

2)　繰返し 1 回目

式 (8) に示すように、$w_i^{(0)}$ を用いた加重最小二乗法 (WLS) により $\widehat{\boldsymbol{\beta}}^{(1)}$ を算出する。このとき、W は対角成分が w_i で、それ以外は 0 となる $n \times n$ の正方行列である。新たな回帰パラメータに基づく残差 $r_i^{(1)}$ から尺度パラメータ $\hat{\sigma}^{(1)}$ を算出し、新たな IRLS ウェイト$w_i^{(1)}$を得る。

$$\widehat{\boldsymbol{\beta}}_{\mathrm{WLS}} = (\boldsymbol{X}^{\top}\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}\boldsymbol{W}\boldsymbol{y} \tag{8}$$

3)　繰返し j 回目

j 回目の繰返しの際、$j-1$ 回目の残差 $r_i^{(j-1)}$ とその尺度パラメータ $\hat{\sigma}^{(\mathrm{j}-1)}$ により算出した IRLS ウェイト $w_i^{(j-1)}$ を用いて、式 (8) に基づき回帰パラメータ $\widehat{\boldsymbol{\beta}}^{(j)}$ を求め、同様に残差、尺度パラメータ及びウェイトを更新する。

4)　収束条件

$1-\hat{\sigma}^{(j)}/\hat{\sigma}^{(j-1)}$ の値が 0 に収束するまで 3)を繰り返す。Bienias et al. (1997) の選択した収束条件は、0.01 未満である。

3.　モンテカルロシミュレーション

ここでは、前節で概説した、IRLS アルゴリズムによる回帰 M-推定量について、二種類のウェイト関数と二種類の残差尺度の組み合わせを用いて四種類の推定量を作成し、計算効率及び推定効率について比較を行い、調整定数や収束条件の影響についても検討を行う。

3.1　シミュレーションデータの設計

次のような単回帰モデルを考える。

$$y = \beta_0 + \beta_1 \boldsymbol{x} + \boldsymbol{\varepsilon} \tag{9}$$

説明変数 $\boldsymbol{x} = (x_1, \ldots x_n)$ は、(0, 10) の値をとる独立な一様乱数、誤差項 $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$ は独立に自由度 $\nu = (2, 3, 5, 10, \infty)$ の t 分布に従う乱数を使用する。説明変数も、本来は正規分布とすべきであるが、梃子比の高い外れ値が発生しやすいように、あえて一様乱数を選択した。そして、切片 $\beta_0 = 5$ 、傾き $\beta_1 = 2$ として、目的変数 $\boldsymbol{y} = (y_1, \ldots, y_n)$ の値はモデル (9) に従い作成する。各データセットの大きさを 100、同一データセット内の誤差項の自由度を同一として、誤差項の自由度別にそれぞれ 100,000 データセットを作成した。なお、自由度が∞ (無限大) の t 分布は、標準正規分布と同等である。

3.2　比較条件

乱数シミュレーションにより比較を行う設定条件を、表 3 に示す。ウェイト関数は Tukey の biweight 関数と Huber のウェイト関数、残差の尺度について AAD と MAD を取り上げ、これらを組み合わせた四種類の推定量について、表 2 に示した AAD 基準の Tukey の調整定数が $c = 4, 6, 8$ と同等の調整定数の設定で、計算効率と推定効率の比較を行う。さらに、これらの各試算条件について、2.4 節のアルゴリズム末尾に示した繰返し計算の収束条件を変えた場合に、これが計算効率と推定効率に与える影響についても検討する。

計算効率は IRLS の繰返し計算回数により比較するが、繰返し計算は初期値計算も 1 回とカウントし、上限を 150 回として、それまでに収束しない場合は計算を打ち切るものとする。推定効率は、式 (10) に示すように、データセット毎の $\hat{y}$ の平均値とモデル (9) に基づく理論値との平均二乗誤差 (MSE: Mean Squared Error) を比較する。ここで、M はデータセット数、n は各データセットの大きさ、$\hat{\boldsymbol{\beta}}^{(m)}$ は m 番目のデータセットに基づく回帰パラメータの推定値、$\boldsymbol{\beta}^*$は乱数発生に用いた真値、$\bar{x}^*$ は x の平均の理論値である。

$$\frac{1}{M}\sum_{m=1}^{M}\left\{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\beta}_0^{(m)} + \hat{\beta}_1^{(m)} x_i\right) - \left(\beta_0^* + \beta_1^* \bar{x}^*\right)\right\}^2 \tag{10}$$

単回帰モデル (9) のパラメータを推定するシミュレーションを、統計ソフト R で行うが、その際に MAD を計算する R の MAD 関数は、結果数値の尺度が SD に準拠するよう補定されるため、実際に設定する調整定数の値は、表 2 の SD 基準の数値を使用する。

シミュレーションに使用した調整定数や収束条件を変えることのできる四種類の推定量の R 関数は、和田（2012）収録の Tirls 関数及び Hirls を元に作成し、github レポジトリ［https://github.com/kazwd2008/IRLS］内のファイル Tirls.r 及び Hirls.r に収録している。ファイル Tirls.r には Tukey の biweight 関数を使用している推定量を収録しており、その中の Tirls.ave が尺度に AAD、Tirls.med が MAD を使用する関数である。同様に、Hirls.r には Huber のウェイト関数を使用した関数を収録し、Hirls.ave が AAD、Hirls.med が MAD を使用している。

表 3. シミュレーションの比較条件

A. ウェイト関数：	① Tukey の biweight 関数	② Huber のウエイト関数	
B. 尺度基準：	① 平均絶対偏差(AAD)		
	② 中央絶対偏差(MAD)		
C. 調整定数：			
Tukey , AAD	① 4	② 6	③ 8
Tukey, MAD	① 5.01	② 7.52	③ 10.03
Huber, AAD	① 1.15	② 1.72	③ 2.30
Huber, MAD	① 1.44	② 2.16	③ 2.88
D. 収束条件			
	① 0.01	② 0.001	③ 0.0001
E. シミュレーションデータの誤差項の自由度（t 分布）			
	① 2 ② 3 ③ 5 ④ 10 ⑤ ∞（無限大）		

4. 結果

ウェイト関数と尺度基準の組み合わせにより得られる四種類の推定量について、繰返し計算の回数による計算効率を 4.1 節、データセット毎の推定値平均の MSE による推定効率を 4.2 節で検討する。収束条件の影響についても検討する。

なお、調整定数の違いは、Tukey の biweight 関数の AAD 基準での調整定数との対応で表記する。具体的には、本節の数表で MAD 基準の場合の「TK6」という表記は、その調整定数が AAD 基準で 6 の場合に対応する SD の調整定数 7.52 の試算を指し、「HB6」とあれば、AAD 基準ならば Huber のウェイト関数の調整定数 1.72、MAD 基準ならば SD の調整定数 2.16 での試算であることを指している。

4.1 計算効率について

まず、乱数データの自由度別の IRLS の繰返し計算の最大回数を、試算条件毎に表 4 に示す。尺度基準が AAD の場合、ウェイト関数によらず MAD よりも最大繰返し回数が少ない。一方で、Tukey と MAD の組み合わせは、繰返し回数がかなり増えるかあるいはループの危険性が高い。少なくとも今回使用したデータセットについて、収束条件が 0.01 の場合に Tukey と AAD の組み合わせは最大 14 回で収束しているため、最大繰返し回数の比較について、MAD と比較した AAD の優位性は明らかである。

表４．最大繰返し回数

尺度	AAD						MAD					
ウェイト&調整定数	TK4	TK6	TK8	HB4	HB6	HB8	TK4	TK6	TK8	HB4	HB6	HB8
収束条件	0.01						0.01					
df 2	6	5	5	5	5	5	36	22	150	18	19	13
df 3	6	5	5	6	5	4	23	17	150	11	11	11
df 5	7	5	5	5	5	4	25	16	13	14	12	14
df 10	6	5	4	5	5	4	15	10	8	11	9	8
df ∞	6	5	4	6	5	4	12	9	5	10	8	6
収束条件	0.001						0.001					
df 2	9	7	7	7	7	6	39	146	150	29	37	19
df 3	10	7	6	8	6	6	37	25	150	17	14	20
df 5	10	7	6	8	6	6	115	27	19	17	19	19
df 10	10	7	6	8	6	5	24	16	11	15	13	12
df ∞	10	7	5	8	6	5	19	14	7	15	11	8
収束条件	0.0001						0.0001					
df 2	13	10	9	10	8	7	150	150	150	41	54	26
df 3	13	9	8	11	8	7	46	32	150	23	20	30
df 5	13	9	7	11	8	7	150	37	26	22	25	25
df 10	15	9	7	11	8	7	33	21	14	21	17	16
df ∞	14	8	7	11	8	7	33	19	8	20	15	11

続いて、表 5 に平均繰返し回数をまとめた。ここでは、繰返しの最大値である 150 回となったデータセットの数値を平均値の計算から除外している。尺度基準が MAD の場合、常に Tukey よりも Huber の繰り返し計算効率が高い。AAD の場合は大きな差異が見えにくいが、収束条件が厳しい場合ほど Huber の計算効率が高いといえる。

次に AAD と MAD を比較すると、MAD 部分の薄い網掛けのセルを除き、ウェイト関数にかかわらず尺度基準が AAD の場合の計算効率が、Tukey で最大 1.43 倍、Huber は最大 1.36 倍である。

4.2　推定効率について

表 6 に、試算条件別に各推定量のデータセット別の y の推定値平均と、乱数モデルから得られる理論値との MSE を示した。 x は、$(0, 10)$ の値をとる独立な一様乱数なので、各データセットの x の平均の理論値は 5 となる。ここで、切片 $\beta_0 = 5$ 、傾き $\beta_1 = 2$ から、y の平均の理論値は式 (9) に基づき 15 であることがわかる。

まず、収束条件に着目する。ウェイト関数や尺度基準にかかわらず、収束条件による差異が非常に小さい。また、むしろ収束条件を厳しくしても標準誤差が改善しないか、ごくわずかに悪化がみられる場合もある。

表５．平均繰返し回数

尺度	AAD						MAD					
ウェイト&調整定数	TK4	TK6	TK8	HB4	HB6	HB8	TK4	TK6	TK8	HB4	HB6	HB8
収束条件	0.01						0.01					
df 2	3.43	3.26	3.15	3.30	3.14	3.04	4.90	4.33	4.04	4.47	4.01	3.78
df 3	3.24	3.03	2.89	3.08	2.91	2.81	4.47	3.86	3.53	4.09	3.61	3.35
df 5	3.07	2.82	2.65	2.90	2.73	2.62	4.10	3.45	3.12	3.80	3.29	2.96
df 10	2.96	2.65	2.48	2.80	2.61	2.47	3.79	3.15	2.86	3.61	3.04	2.60
df ∞	2.86	2.51	2.34	2.72	2.51	2.31	3.50	2.91	2.69	3.42	2.77	2.23
収束条件	0.001						0.001					
df 2	4.95	4.39	4.10	4.49	4.07	3.85	6.51	5.54	5.09	5.83	5.04	4.70
df 3	4.93	4.22	3.85	4.40	3.90	3.61	6.01	4.97	4.46	5.39	4.56	4.16
df 5	4.86	4.00	3.59	4.31	3.75	3.39	5.55	4.45	3.95	5.06	4.16	3.63
df 10	4.78	3.81	3.37	4.25	3.64	3.22	5.16	4.08	3.63	4.82	3.83	3.09
df ∞	4.69	3.62	3.18	4.23	3.54	3.05	4.78	3.79	3.39	4.58	3.43	2.47
収束条件	0.0001						0.0001					
df 2	6.59	5.59	5.11	5.84	5.10	4.71	8.12	6.77	6.15	7.18	6.09	5.63
df 3	6.80	5.49	4.89	5.93	5.00	4.48	7.56	6.09	5.40	6.71	5.54	4.98
df 5	6.88	5.31	4.61	5.96	4.89	4.26	7.01	5.48	4.78	6.34	5.05	4.30
df 10	6.90	5.12	4.37	5.98	4.81	4.07	6.55	5.04	4.37	6.07	4.64	3.58
df ∞	6.87	4.91	4.16	6.03	4.72	3.88	6.09	4.69	4.09	5.79	4.12	2.71

次に尺度 AAD と MAD の場合を比較すると、ウェイト関数にかかわらず、誤差項が正規分布の場合は MAD の MSE が小さく、また調整定数が小さいほどより裾の重い分布についても MAD の MSE がより小さくなる。分布の裾が重くなるほど AAD の MSE の方が小さいが、自由度 2 では再び、特に Huber のウェイト関数について MAD の MSE の値が小さい傾向がみられた。通常、調査データの分布は正規分布よりも裾が長いことが多いため、そのような場合には尺度に AAD を使用し、調整定数は大きめの数値を選択するのが良いが、裾がかなり重い分布が想定される場合は Huber のウェイト関数に MAD を尺度とし、小さめの調整定数が良いということになる。

そして、ウェイト関数を比較すると、差異はわずかではあるが、尺度によらず調整定数が小さい場合は Huber の誤差が小さく、調整定数が大きくなるほど Tukey の誤差が小さくなる。また、誤差分布の自由度が大きいほど Huber の誤差の方が小さくなる傾向がみられる。

ところで、Tukey の biweight 関数を使用する場合、ループを起こす可能性があり、今回のシミュレーションでは尺度に MAD を使用した場合にその現象が観察された。この場合、二種類のウェイトのセットが繰り返し計算の中で交互に付与され計算が収束しなくなる。この原因は、ある程度以上極端な外れ値のウェイトは 0 という同一の値になるこの関数の性質に起因する。Huber のウェイト関数については、極端な外れ値でも微小なウェイトが付与され、繰り返し計算の中で推定結果に応じてそのウェイトは変化するため、稀に収束に時間がかかることは起こるが無限ループは発生しない。ただし、biweight 関数において無限ループが起こる場合の

対処は容易で、調整定数の値をわずかに変えるだけで計算を収束させることができる。また、Tukey と MAD の組み合わせは、計算不能に陥ることがある。今回の試算条件ではそのケースは起きていないが、t 分布の自由度を 1 に設定した場合（コーシー分布）に稀に発生する。これは、程度は異なるが非常に極端な外れ値が二つ、梃子比の小さくなる説明変数の中心部ある場合に発生する。計算不能の原因は、この二つの外れ値を正常値、それ以外の全ての観測値を外れ値と誤認識してウェイトにゼロが付与され、回帰計算に必要なデータ数を実質的に満たせなくなるためである。

表６．推定値平均の MSE

尺度		AAD					MAD				
自由度		df 2	df 3	df 5	df 10	df ∞	df 2	df 3	df 5	df 10	df ∞
OLS		0.7534	0.3636	0.3503	0.3468	0.3437	0.7534	0.3636	0.3503	0.3468	0.3437
収束条件(0.01)	TK4	0.3538	0.3497	0.3474	0.3464	0.3452	0.3536	0.3499	0.3473	0.3460	0.3442
	TK6	0.3561	0.3504	0.3473	0.3459	0.3441	0.3560	0.3511	0.3478	0.3461	0.3438
	TK8	0.3588	0.3515	0.3477	0.3460	0.3438	0.3589	0.3525	0.3483	0.3463	0.3437
	HB4	0.3556	0.3500	0.3472	0.3461	0.3446	0.3547	0.3502	0.3472	0.3460	0.3441
	HB6	0.3588	0.3510	0.3474	0.3459	0.3441	0.3581	0.3519	0.3480	0.3461	0.3438
	HB8	0.3624	0.3523	0.3479	0.3461	0.3438	0.3614	0.3535	0.3487	0.3464	0.3437
収束条件(0.001)	TK4	0.3535	0.3498	0.3476	0.3469	0.3458	0.3535	0.3499	0.3473	0.3460	0.3442
	TK6	0.3558	0.3503	0.3473	0.3459	0.3442	0.3560	0.3511	0.3477	0.3461	0.3438
	TK8	0.3586	0.3514	0.3477	0.3460	0.3439	0.3588	0.3525	0.3483	0.3463	0.3437
	HB4	0.3550	0.3497	0.3471	0.3462	0.3449	0.3546	0.3502	0.3472	0.3460	0.3442
	HB6	0.3585	0.3508	0.3473	0.3459	0.3442	0.3581	0.3519	0.3480	0.3461	0.3438
	HB8	0.3622	0.3522	0.3479	0.3461	0.3439	0.3614	0.3534	0.3487	0.3464	0.3437
収束条件(0.0001)	TK4	0.3534	0.3498	0.3478	0.3470	0.3461	0.3535	0.3499	0.3473	0.3460	0.3442
	TK6	0.3558	0.3503	0.3473	0.3460	0.3442	0.3560	0.3511	0.3477	0.3461	0.3438
	TK8	0.3586	0.3514	0.3477	0.3460	0.3439	0.3588	0.3525	0.3483	0.3463	0.3437
	HB4	0.3548	0.3496	0.3471	0.3463	0.3451	0.3546	0.3502	0.3472	0.3460	0.3442
	HB6	0.3584	0.3508	0.3473	0.3459	0.3442	0.3581	0.3519	0.3480	0.3461	0.3438
	HB8	0.3621	0.3522	0.3479	0.3460	0.3439	0.3614	0.3534	0.3487	0.3464	0.3437

5.　まとめと考察

従来、MAD 基準の Huber 推定量は理論上良い性質を持つとされてきたが、本稿の試算条件で、AAD 基準の Tukey の biweight 関数を使用した推定量と比較すると、誤差が正規分布かそれに近い場合に Huber のウェイト関数、より裾の長い誤差項については Tukey のウェイト関数を使用する場合推定効率がわずかに高くなるが、実用上大きな差異はないと言える。一方で、最適な収束条件を 0.01 とすれば、計算効率については、AAD 基準の Tukey の biweight 関数を用いた方が高い。

また、Huber のウェイト関数に AAD を組み合わせた推定量は、これまで利用事例がないが、AAD 基準の Tukey のウェイト関数の推定量と同様に計算効率が高いことがわかった。また、誤差が正規分布かそれに近い場合を除き、MAD 基準の Huber 推定量と比較して、わずかだが推定効率も高い。

実用上、ウェイト関数の選択にあたり、推定効率や計算効率以外にも考慮すべきなのは、両者のウェイト関数の性質の違いである。Tukey の biweight 関数の場合は、極端な外れ値のウェイトをゼロとして、その影響を排除できるが、Huber のウェイト関数の場合は極端な外れ値であっても推定から完全に排除することはないという性質の差異から、和田（2012）が述べるように、外れ値の扱いに関する利用者の方針により選択するのが望ましい。

本稿の目的としている補完の場合を考えれば、対象となるデータはクリーニング済みで誤りは含まれていないことが前提である。これを用いて、他の欠測のある観測値を補完するための推定に、影響が大きく他の大多数の観測値と傾向が違う外れ値は、推定から完全に排除できる方が望ましい場合が多く、その場合は Tukey の biweight 関数という選択になる。補完推定から排除されたとしても、当該観測値は誤りではないので、通常それを理由に集計自体から除外されることはない。一方で、補完ではなく直接的に母集団推定を行う場合、たとえ外れ値であっても誤りではない全ての観測値を使用することが望ましい、という方針になれば、Huber のウェイト関数が選択されるであろう。

尺度に AAD が優れる理由を考察すると、AAD よりも MAD のロバスト性が高いことは、式（5）及び（6）から明らかである。AAD がロバストではない平均値からの偏差であるのに対して MAD ではロバストな中央値からの偏差を採用しているため、IRLS で推定が繰り返されるとき、明らかに MAD よりも AAD の方が数値の動きは大きくなる。そして、IRLS は繰返し計算の収束を尺度の推定量の変化率で判定するために、AAD が特に計算効率に大きく寄与していると思われる。

収束条件は、数値を小さくしても推定効率への貢献がほとんどみられない一方で、計算効率は大きく悪化する。Bienias et al (1997) の 0.01 という収束条件は、AAD 基準の Tukey だけでなく、尺度が MAD の場合も含め、Tukey と Huber のウェイト関数両方について良い選択であるといえる。

最後に、データセットに、あまり極端な値ではないがある程度の外れ値が含まれており、その程度が時点や補完ドメイン毎に変化する、という一般的な状況を想定した統計調査の補完プロセスの場合、IRLS の推奨される設定条件は、以下のようになる。

まず統計を作成する担当者は、外れ値に対する方針を決める必要がある。もし、全ての観測値を推定に用いるという方針であれば、Huber のウェイト関数を、自由度 3 から 5 の t 分布に相当するデータであれば AAD 尺度、それ以上に裾が重いことが想定される場合は MAD 尺度とともに使用することが望ましい。このときの調整定数は、適用データセットが正規分布に近ければ大きく、極端な外れ値が見込まれれば小さくすると良い。一方で、影響の大きすぎる極端な外れ値を、推定からは自動的に除外したいと考える場合には、Tukey の biweight 関数を AAD 尺度とともにウェイト関数に使用し、調整定数の選び方は Huber の関数の場合と同じである。一般に、補完の場合は Tukey のウェイト関数を選ぶ場合が多い。

本研究では、IRLS に用いられる性質の異なる代表的なウェイト関数と、尺度基準の組み合わせにより得られる四つの推定量について、モンテカルロシミュレーションにより比較検証を行った。その結果 Bienias et al. (1997) の提案するウェイト関数と尺度の選択を裏付ける結果が得られ、統計調査データの補完において、IRLS の適切な設定方法を明らかにすることができた。また、実用性の高い推定量として、これまでに使用事例がみられない Huber のウェイト関数の尺度に AAD を使用する推定量を提案する。

参考文献

[1] 蓑谷千凰彦 (1992), *計量経済学における頑健推定*, 多賀出版.

[2] 和田かず美 (2012) , 多変量外れ値の検出～繰返し加重最小二乗(IRLS)法による欠測値の補定方法～, *統計研究彙報*, 総務省統計研修所, **69**, 23-52.

[3] Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H. and Tukey, J. W. (1972), *Robust estimates of location – Survey and advances*, Princeton University Press.

[4] Antoch, J., & Ekblom, H. (1995), Recursive robust regression computational aspects and comparison, *Computational statistics & data analysis*, **19**(2), 115-128.

[5] Beaton, A. E. and Tukey, J. W. (1974), The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data, *Technometrics*, **16**(2), 147-185.

[6] Bienias, J. L., Lassman, D. M., Scheleur, S. A., & Hogan, H. (1997), Improving outlier detection in two establishment surveys, *Statistical Data Editing*, 2. UNECE (United Nations Economic Commission for Europe).

[7] Hampel, F. (2001), Robust Statistics: A Brief Introduction and Overview, Research Report No.94, Seminar für Statistik, Eidgenössische Technische Hochschule (ETH), Switzerland.

[8] Holland, P. W. & Welsch, R. E. (1977), Robust Regression Using Iteratively Reweighted Least-Squares, *Communications in Statistics – Theory and Methods*, **6**(9), 813-827.

[9] Huber, P. J. (1964), Robust estimation of a location parameter, *The annals of mathematical statistics*, **35**(1), 73-101.

[10] Huber, P. J. (1973), Robust Regression: Asymptotics, Conjectures and Monte Carlo, *Annals of Statistics*, **1**(5), 799-821.

[11] Huber, P. J. and Ronchetti, E. M. (2009), *Robust statistics*, 2nd ed., NJ: Wiley.

[12] Mosteller, F. and Tukey, J. W. (1977). *Data analysis and regression: a second course in statistics*. Addison-Wesley Series in Behavioral Science: Quantitative Methods.

[13] Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*, John Wiley & Sons, Inc.