製表技術参考資料 17

ISSN 1348－7329

諸外国のデータエディティング

及び

混淆正規分布モデルによる多変量外れ値検出法についての研究

NSTAC

Working Paper No.17

平成 24 年 8 月

独立行政法人　統計センター

製表技術参考資料は、独立行政法人 統計センターの職員がその業務に関連して行った製表技術に関する研究の結果を紹介するためのものである。

ただし、本資料に示された見解は、執筆者の個人的見解である。

目　次

諸外国のデータエディティング
及び
混淆正規分布モデルによる多変量外れ値検出法についての研究*

高橋　将宜**

要 旨

本稿は、海外におけるデータエディティングに関する最新の研究動向を調査・研究したものである。この目的のために、2009 年及び 2011 年に、国連欧州経済委員会(UNECE: United Nations Economic Commission for Europe)[1]の統計データエディティングに関するワークセッションにおいて報告された 88 論文を調査し、とりわけ、その中から選択的エディティング(Selective Editing)に関する論文を精査した。さらに、イタリア国家統計局による多変量外れ値[2]検出に関する論文を詳しく検討し、独立行政法人統計センター（以下、「統計センター」とする）における将来の業務への応用可能性を探求している。本稿では、この調査の結果をもとに、諸外国におけるデータエディティングの概要及び混淆正規分布モデル(Contaminated Normal Model)[3]による多変量外れ値検出法について以下のとおりまとめた。

第 1 節において、データエディティングの概観を示す。ここでは、過去に統計センターにて行った研究を紹介し、様々な用語の定義を再検討し、選択的エディティングの重要性を示している。第 2 節では、データエディティング及び選択的エディティングに関して活発な議論を行っている国際会議として、国連欧州経済委員会(UNECE)の統計データエディティングに関するワークセッションの概要を紹介する。第 3 節では、イタリアにおける選択的エディティングの変遷を紹介する。第 4 節では、混淆正規分布モデルの概論を示し、第 5 節では、混淆正規分布モデルによる選択的エディティング手法を示す。第 6 節では、シミュレーションデータにより R の *SeleMix* パッケージのデモンストレーションを行い、第 7 節では、経済センサスの経理項目を模したデータとして、EDINET データを使用したデモンストレーションを行う。最後に将来の可能性に関する議論で締めくくる。

---

* 本稿は、平成 23 年度第 2 回統計技術研究会（平成 24 年 3 月 13 日）において報告した資料を増補・改訂したものである。統計技術研究会の参加者の皆様、統計技術研究課の研究員の皆様に感謝の意を表したい。また、本稿の内容は執筆者の個人的見解を示すものであり、機関の見解を示すものではない。

** 統計センター 情報技術部 統計技術研究課 上級研究員

---

1 http://www.unece.org/

2 外れ値とは、データの全体的なパターンから大きく逸脱した観測値であり、測定誤差、他の母集団に属すべき観測値、特異な観測値のことである(Weiss, 2005, p.122)。

3 本稿では、Contaminated Model の訳語として「混**淆**（こんこう）モデル」を使用し、Mixture Model の訳語として「混**合**モデル」を使用する。Mixture Model「混**合**モデル」の一部が、Contaminated Model「混**淆**モデル」であると言える。技術的な詳細は、渡辺，山口 (2000)の第 4 章を参照されたい。

諸外国のデータエディティング及び混淆正規分布モデルによる多変量外れ値検出法についての研究

高橋　将宜

序論（研究の目的）

統計センターでは、設立以来、データエディティング及び欠測値補定に関して研究を進めており、国際的な研究動向の把握にも努めてきた。過去には、国連欧州経済委員会(UNECE)の統計データエディティングに関するワークセッションにも参加し情報収集してきたが、近年は参加していない。

一方で、総務省と経済産業省の共管により、2012 年 2 月に経済センサス-活動調査が初めて実施され、調査結果の精度確保のために、売上高などの経理項目におけるデータエディティングが重要になってきている。

このため、統計センターの参加しなかった前回（2011 年）及び前々回（2009 年）に開催された統計データエディティングに関するワークセッションにおける報告について、2011 年の秋よりワーキングペーパーを中心に文献収集及び有用と思われる論文の調査を行った。本稿は、その成果として、今後の我が国統計調査におけるデータエディティング研究に資する材料を取り上げたものである。

1 データエディティング

1.1 データエディティングに関する過去の研究概要

統計センターでは、過去においてもデータエディティングに関する研究を行ってきた。堀内（2006)では、英国（2005)、スウェーデン（2005)、オランダ（2005)、ドイツ（2005）、オーストラリア（2000）、カナダ（1992)における選択的エディティングに関する論文を精査した。選択的エディティングとは、「疑わしい個別データ各々について、集計値レベルの影響度や疑わしさの度合いを所定の算式によって推定（スコア化）し、その大きさが一定値以上のものに絞って、人手による確認を行う」(p.26)ものである。堀内（2006)では、「スコアの算出法は国や調査によって異なっているが、算出法の違いに係わらず、報告されている適用例では、エディティング業務の大幅な効率化が達成されている」(p.32)ことが分かった。

また、畠山（2008)では、国連欧州経済委員会(UNECE)による *Statistical Data Editing Vol.3*（統計データ・エディティング Vol.3）の概要をまとめた。データエディティングとは、「データのエラーを検出するプロセス」であり、エラーとは、「データ測定値とそのデータに対応する真値との差」と定義された(p.4)。その目的は、「データ品質へのデータエディテ

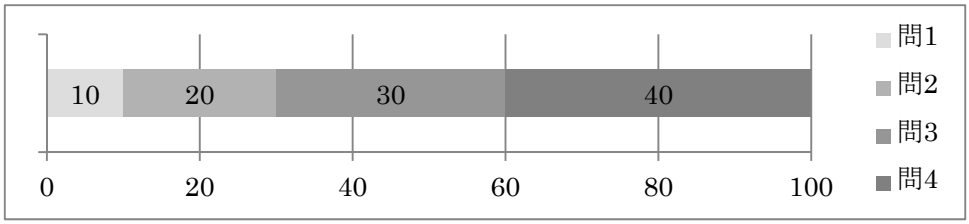
ィングプロセスの影響を評価し、どのようにうまくプロセスを実行するかについて国家統計機関を支援する」ものである(p.1)。

以上のとおり、統計センターにおいては、過去にも諸外国におけるデータエディティングに関して、精力的に研究を行ってきたが、選択的エディティングに関しては 6 年の歳月が、またデータエディティング全般に関しても 4 年の歳月が流れたため、本稿は、これらの過去の研究を踏まえて、最新の動向を改めて注視するものである。また、統計センターにおいて行った多変量外れ値検出法の研究としては、岡本 (2004)が詳しく、M-推定量、MCD 法、射影追跡法、Forward Search 法、感染アルゴリズム法、NNVE 法、クラスター分析法、ロバスト回帰木モデルの調査研究を行っている。しかし、岡本 (2004)では、「多変量外れ値検出法の一種に、外れ値の無い補助変量と外れ値を検出したい変量間でロバスト回帰を行い、回帰モデルの推定値との残差の大きさから外れ値を検出する方法がある。これも重要な方法論ではあるが、本稿では割愛」しており、本研究は、この穴を埋めるものである。近年では、和田 (2010)において、ロバストな多変量外れ値検出法として MSD (Modified Stahel-Donoho)法が紹介されており、データの特性に応じて、本稿で紹介する混淆正規分布モデルに基づく多変量外れ値検出法とあわせて実務に応用されたい。

1.2 動機付け：テストの例

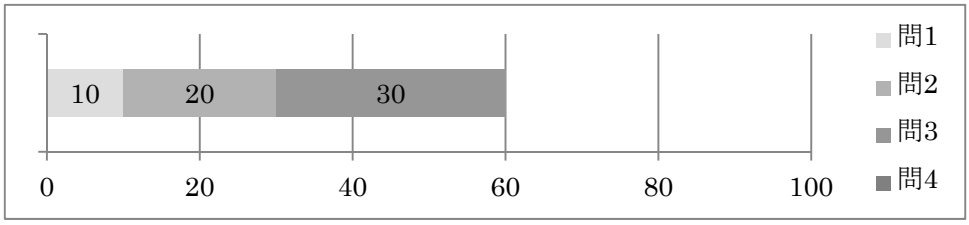
本節では、そもそも、なぜ選択的エディティングが必要なのかについて簡単に触れておきたい。直感的な例として、以下のような 4 問、100 点満点のテストを考えてみる。設問の難易度はすべて同じと仮定し、解答時間は 30 分とする。解答に要する時間は 1 問あたり 10 分かかると仮定する。4 問の配点は以下のとおり。問 1＝10 点、問 2＝20 点、問 3＝30 点、問 4＝40 点。この配点を図示すれば、図 1.1 のようになる。

図 1.1



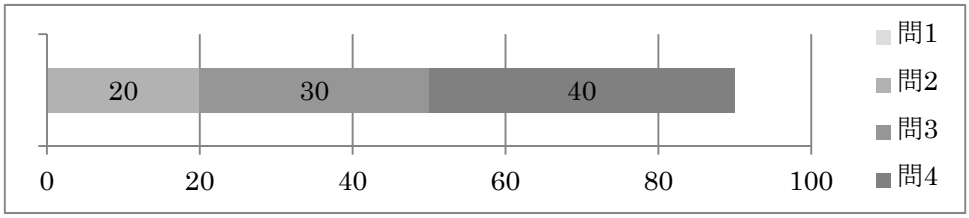
仮に、配点を知らずに 1 問目から順番に解答したとしよう。この場合、問 1＝10 点、問 2＝20 点、問 3＝30 点の合計 60 点しか取れないことになる（図 1.2)。

図 1.2



しかし、配点を知っており、その重要度を優先させて解答するならば、問 4＝40 点、問 3＝30 点、問 2＝20 点の合計 90 点を獲得できることになる（図 1.3）。

図 1.3



テストにおいては、100 点満点を取ることが理想的ではあるが、制限時間など、様々な理由により満点を達成することは困難であることが多い。したがって、配点の低い設問を先に解くよりも、配点の高い設問から解き始める方が効率的で得策である。

今回の例において重要なポイントは、能力は同じであるにも関わらず、点数が飛躍的に伸びたという事実である。つまり、同じ能力、同じ制約であるならば、配点の高さを優先して解答することが有益であることがはっきりと分かる。データエディティングに当てはめた場合、同様の制約の下で、同じだけの人手をかけられるならば、優先スコアを知ることで、より精度の高いエディティングを行うことができると言えるのである。

逆に言えば、もしも同じレベルの精度を保てばよいのであれば、選択的エディティングを用いることにより、短い時間でエディティングを行うことができ、また、少ない人手で同じレベルのエディティングを行うことができると言える。

すなわち、選択的エディティングは、まさに製表の三原則[4]の申し子なのである。

[4] 製表の三原則とは、製表業務において統計センターの目指す指針であり、「正確性：統計精度を確保すること」、「迅速性：早期に結果を提供すること」、「経済性：効率的な手段・方法を用いること」の 3 つのことである。http://www.nstac.go.jp/services/index.html

1.3 エディティングの種類と定義

データエディティングは、様々な国の統計局において独立して発展してきた経緯もあり、その用語には統一性のない部分があり、議論の混乱の原因となることがある。そこで、本節では、重要な用語を簡単に定義しておきたい。ここで用いる定義は、主に、Farwell (2005)、Scarrott (2007, p.5)、Ton De Waal *et al.* (2011, 第 6 章)に準拠している。

ミクロエディティングとは、「他のユニットの回答には基づくことなく、現在のユニットの回答及び補助的な情報にのみ基づいて、個別ユニットの回答の妥当性及び一貫性に関してエディティングを行う」ものである。また、インプットエディティングとは、「データ収集段階、データ入力段階、データ処理段階において行われるエディティング」である。ミクロエディティングとインプットエディティングとは、同義のものとして扱われることも多い。

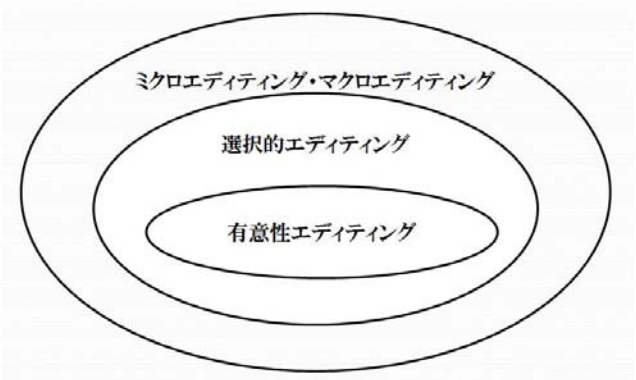
ミクロエディティングでエディットされたデータは、マクロエディティングへと移され、ここでは、集計レベルで生産された統計値に影響を与えるエラーに焦点をあてる。マクロエディティングとは、「多くのユニットの回答に基づいて、個別ユニットの回答の妥当性及び一貫性に関してエディティングを行う」ものであり、「全データベースのレコードや異なった領域の調査変数に関してチェック」を行うものである。また、アウトプットエディティングとは、「多くのユニットの回答を分析した結果に基づいて、データ処理段階の後に行われるエディティング」のことである。マクロエディティングとアウトプットエディティングとは、同義のものとして扱われることも多い。

選択的エディティングは、ミクロエディティング及びマクロエディティングのギャップの架け橋となるものであり、ユニット回答をエディティングの対象としている点がミクロ的であり、調査結果に影響を与えるエラーを優先化している点がマクロ的である。選択的エディティングとは、「エラーである可能性があり調査結果に影響を及ぼし得る回答の修正及び補定を優先化する手法」のことである。

選択的エディティングの亜種として、オーストラリア統計局による有意性エディティング(Significance Editing)が知られている。有意性エディティングとは、スコア関数を用いて、調査変数に与えるエラーの影響を直接的に数値化する手法であり、選択的エディティングの一種と言える。先行文献においては、選択的エディティングと有意性エディティングとの違いに関して、若干の混乱が見られる。また、オランダ統計局による妥当性指標(Plausibility Indicator)など、多数の亜種が存在するが、ほとんどの手法に共通しているのは、何らかのスコア関数を用いてエディティングの優先化を行うということである。

以上の定義を簡易的に図示すれば、図 1.4 のような内包関係となる。つまり、すべての有意性エディティングは選択的エディティングであるが、すべての選択的エディティングは必ずしも有意性エディティングであるとは限らないということが分かる。

図 1.4



1.4 「選択的エディティング」概念の起源と未解決の問題

選択的エディティングは、様々な国において開発されてきた手法の集大成と言え、いつ、どこで、誰が始めたかということに関してピンポイントで示すことは難しい。しかし、先行文献においては、Latouche and Berthelot (1990, 1992)に起源を求めるものが多い(Farwell, 2005; Scarrott, 2007; Ton De Waal *et al.*, 2011)。当時、カナダ統計局方法論部の上級研究員であった Michel Latouche と Jean-Marie Berthelot は、スコア関数を用いてエディティングに優先付けをした最初期の有力な研究を行った。スコア関数を用いない例やマクロエディティング手法の例を含めれば、1980 年代までさかのぼることが可能だが(Breiman *et al.*, 1984; Hidiroglou and Berthelot, 1986; Granquist, 1990)、現在の選択的エディティング手法は、スコア関数を用いてマクロ的及びミクロ的エディティングを行う方法が最も一般的であり、そういった意味で選択的エディティングの起源である。

Latouche and Berthelot (1992)は、スコア関数作成に際して、以下の 4 つの指針を示し、現在の選択的エディティング手法は、多かれ少なかれ、この 4 つの指針を継承していると言える。(1) 回答ユニットの重要性 (2) 疑わしい回答の与える影響の度合い (3) 疑わしい回答の数 (4) 項目または変数の相対的な重要性。しかし、Latouche and Berthelot (1992)による指針は漠然としたものであり、スコア関数を具体的にどのようにして作成するかということは、今日においても未解決の問題である。むしろ、各国の統計機関が、各々の調査ごとに、各々のデータに即して、多数の手法を提供しており、汎用的な方法は存在しない[5]。また、選択的エディティングは、量的連続変数のエディティングに関して非常に有効な手段であるが、名目変数や順序変数など、自然なメトリックが存在しない場合には、選択的エディティングの応用は難しく、未解決の問題である(Scarrott, 2007, p.25)[6]。

[5] 唯一の例外として、Arbués, Revilla, and Saldaña (2011)による確率論的な統合理論構築の試みが挙げられるが、これに至っても、国際的なフレームワークとなっている訳ではない。

[6] こういった穴を埋めるための研究として、現在、統計センターでは、新井, 伊藤, 阿部, 佐々木, 袈岩, 土井 (2010)のプロジェクトにおいて、名目変数の選択的エディティングに関する研究を継続中であることを指摘しておく。

2 UNECE 統計データエディティングに関するワークセッション

国連欧州経済委員会(UNECE)主催による統計データエディティングに関するワークセッションは、データエディティング及び選択的エディティングに関して活発な議論を行っている国際会議である。この会議は、1 年半周期で開催され、欧州を中心に米国、カナダ、オーストラリア、アジアなどの各国統計機関が参集、討議を行うものである。その内容は、データエディティングの革新的な手法や技術開発、統計の加工処理におけるデータエディティングの工程など多岐に渡り、この会議において対象としている聴衆は、センサスや行政情報源などから得られたデータのエディティングや補定に関わる統計家であり、社会経済的な様々な分野を対象とする。

直近では、2009 年と 2011 年に開催された。2009 年 10 月には、スイスのヌーシャテルにおいて開催され、8 つの事項（うち 1 つはキャンセル）[7]が討議された。また、最新の会議は、2011 年 5 月にスロベニアの首都リュブリャナにおいて開催され、8 つの事項[8]が討議された。次回は、2012 年 9 月 24 日から 26 日までの日程で、 ノルウェーの首都オスロにおいて開催予定であり、7 つの事項[9]が討議される予定である。

3 イタリアにおける選択的エディティングの変遷

2009 年及び 2011 年の統計データエディティングに関するワークセッションにおいて報告された全 88 論文を調査したところ、イタリア国家統計局による混淆正規分布モデルに基づいた多変量外れ値検出法の研究が最新的で実用的な内容であることが分かり、現在、統計センターにおける将来の業務への適用可能性を検討しているところである。

そこで、本節では、このワークセッションにおいてイタリア国家統計局が報告した諸論文を簡単に紹介し、イタリアにおける選択的エディティングの変遷を見ていきたい。イタリア国家統計局では、Marco Di Zio 氏を中心とするデータエディティングチームにより、2002 年頃からデータエディティングに関する研究を盛んに行っている。

---

[7] (1) エディティングと補定の自動化及びソフトウェアへの応用; (2) 情報源の近くでのエディティング; (3) 行政及びセンサスデータのエディティングと補定; (4) 企業/母集団統計の成功事例（キャンセル）; (5) エディティングと補定の新手法の実装戦略の成功例; (6) 新たな手法; (7) データエディティングと補定の品質への影響を測る指標; (8) 選択的マクロエディティング。報告された全論文は、以下のウェブサイトにて閲覧及びダウンロード可能である。http://www.unece.org/stats/documents/2009.10.sde.html

[8] (1) 行政及びセンサスデータのエディティング; (2) 電子収集物のエディティング; (3) マクロエディティング手法; (4) ミクロエディティング–手法とソフトウェア; (5) 変容する組織文化; (6) 国際協力; (7) 新たな手法; (8) 将来の展望と K ベースウィキ。報告された全論文は、以下のウェブサイトにて閲覧及びダウンロード可能である。http://www.unece.org/stats/documents/2011.05.sde.html

[9] (1) 選択的及びマクロエディティング; (2) エディティングのグローバルな解決策; (3) 複数情報源及び混合モードからのデータ統合の文脈におけるエディティング及び補定; (4) エディティングプロセスの効率性を分析するためのメタデータ及びパラデータの使用法; (5) データエディティング及び補定のためのソフトウェアとツール; (6) 新たな手法; (7) センサスデータのエディティング及び補定。2012 年に報告される論文は、以下のウェブサイトに順次公開され、閲覧及びダウンロード可能となる予定である。http://www.unece.org/stats/documents/2012.09.sde.html

Di Zio, Luzi, and Manzari (2002)では、エディティング及び補定(E&I: Editing and Imputation)プロセスの影響評価を行っている。この論文では、以下の評価目的に基づき、イタリアの経験談を報告している：(1)任意の E&I 問題に関して、任意の E&I 手法の統計的性質を検証すること；(2)任意の調査目的のために、最良の手法を選ぶこと；(3)任意の E&I 手法のパフォーマンスを監視し最適化すること；(4)非標本誤差に関する情報を入手すること；(5)E&I の元データに与える影響を測ること。評価目的 1 から 4 は、プロセスの信頼性に関する情報に関連しており、評価目的 5 は最終データの信頼性に関する情報をユーザーに提供することを目的としている。これらの評価目的のために、シミュレーションを使い、真のデータセットに汚染データ(Corrupted Value)を人工的に加え、エディティング及び補定を行っている。

Di Zio, Guarnera, and Luzi (2003)では、混合モデルに基づき、エラーを確率的に形式化するモデルベースの手法を提唱している。調査変数 $\boldsymbol{X}$ を確率変数とし、その期待値は $\boldsymbol{E(X)=\mu}$、分散は$\boldsymbol{Var(X)=\sigma^2}$とする。エラーには体系的エラーとランダムエラーの 2 種類がある。体系的エラーは期待値に影響を与える：$\boldsymbol{E(X)=g(m)}$。一般的には、加法エラーメカニズムが想定されているので、$\boldsymbol{g(m)=m+C}$（ここで$\boldsymbol{C}$は定数）である。一方、ランダムエラーは $\boldsymbol{X}$の分散構造に影響を与えるので、$\boldsymbol{Var(X)=\sigma_\varepsilon^2}$であり、一般的には$\boldsymbol{\sigma_\varepsilon^2 \gg \sigma^2}$である。この論文では、体系的エラーに着目しており、混合モデルを用いることで、クラスター分析と同様に、特定のエラーパターンにしたがってデータを異なるグループに分類することを目的としている。Di Zio, Guarnera, Luzi, and Manzari (2005)では、2003 年のモデルを拡張し、体系的な測定単位エラー検出のための有限混合モデルを構築している。

Di Zio and Guarnera (2006)では、混合正規モデルのフレキシブルさを利用してデータ分布の近似を求め、セミパラメトリックな補定法である予測平均値マッチング(Predictive Mean Matching)を行っている。ここでは、説明変数の値を直接的に用いるのではなく、観測された説明変数を条件として欠測変数の期待値を基に距離を算出する最近隣ドナー手法により欠測値を補定する。通常の予測平均値マッチングでは、期待値は線形回帰モデルを用いて算出されているが、この論文では、距離算出のために使用する期待値を混合正規モデルによって推定している。この手法により、変数間の非線形的な関係にも対処することができる。

以上のように、イタリア国家統計局では混合モデルの研究を積極的に行い、エラーの検出及び補定に適用してきた。Di Zio, Guarnera, and Luzi (2008)では、Ghosh-Dastidar and Schafer (2006)を参考に、外れ値検出法として混淆正規分布モデルを採用した。2003 年/2005 年には体系的エラーに着目したが、2008 年にはランダムエラーに着目し、 潜在的に影響力のある外れ値の検出法の研究を行っている。エラーのないデータの分散を増大させることにより、エラーデータの分布が得られるという仮定に基づいて、エラー確率及びエラーの影響度の両方を推定できる多変量エラーモデルを提唱している。

さらに、Bellisai, Di Zio, Guarnera, and Luzi (2009)では、混淆正規分布モデルによる

多変量外れ値検出法の理論化を発展させた。混淆正規分布モデルを使うことで、エラーと残差の変動部分との区別をつけることが可能となり、観測値のスコアを該当ユニットのエラーの期待値と直接的に関連付けることができるようになった。そして、期待値条件付最大化法(ECM: Expectation Conditional Maximization)アルゴリズムを使用して、最尤推定値(MLE: Maximum Likelihood Estimate)を求める方法も確立した。

2008 年及び 2009 年の論文を集大成させるものとして、Buglielli, Di Zio, Guarnera, and Pogelli (2011)では、混淆正規分布モデルによる外れ値検出法を、R の *SeleMix* パッケージとしてソフトウェア化した。 このパッケージは、ウェブサイト[10]より無料でダウンロードし、R に実装することにより誰でも使用可能となっている[11]。

4 混淆正規分布モデル概論[12]

前節で見たとおり、イタリア国家統計局では、混淆正規分布モデルを用いたデータエディティングの研究を盛んに行っている。本節では、混淆正規分布モデルを簡単に紹介する。

そもそも一般的な統計的分析において使用される確率分布は、一様分布を除くと、単峰の分布であることが多い。しかし、現実の世界では複数の峰を持つ分布が存在する。このような場合、最も単純なモデルとして式(1)のように、単峰の分布($\boldsymbol{f_i}$)を集め、それぞれに重み($\boldsymbol{w_i}$)を付けて足し合わせるという方法が考えられる。ここで、$\boldsymbol{w_i}$は正の値($\boldsymbol{0 < w_i \leq 1}$)であり、その総和は 1 である($\boldsymbol{\sum_{i=1}^{n} w_i = 1}$)。

$$\boldsymbol{f = \sum_{i=1}^{n} (w_i f_i)} \qquad \textbf{(1)}$$

このとき、単峰の分布$\boldsymbol{f_i}$の個数 n が分かっている場合、各々の$\boldsymbol{f_i}$に含まれるパラメータの推定を行うだけでよいが、n が分かっていない場合、n をどのようにして推定するかという混合分布問題を考えなければならない(金田, 新居, 2009, p.1)[13]。しかし、多変量外れ値検出法としての混淆正規分布モデルでは、観測データは、真のデータとエラーのあるデータの 2 つの分布が混在したものから得られたと考える。つまり、n は自ずと 2 以下と定まる。

一般的に、$\boldsymbol{n \leq 2}$の混淆正規分布モデルは、以下の式(2)で表すことができる。すなわち、

---

10 http://cran.r-project.org/web/packages/SeleMix/index.html

11 以下の手順で実装することにより使用可能となっている：(1) Windows binary: SeleMix_0.8.1.zip を自分の PC に保存; (2) ダウンロードした Zip ファイルをいったん解凍; (3) 解凍した Zip ファイルを再度、圧縮; (4) R を起動し、「パッケージ→ローカルにある zip ファイルからのパッケージのインストール」をクリックし、(3)で圧縮したファイルを選択。ファイル名などは、2012 年 5 月 28 日現在の情報である。また、以上の手順は、Windows の場合である。Mac ユーザーの場合は、MacOS X binary: SeleMix_0.8.1.tgz を自分のパソコンに保存。本来であれば、上記手順の(1)と(4)のみでインストールできるはずであるが、筆者の PC においては、上記の手順を踏むことでインストールをすることができた。

12 外れ値生成モデルとしての混淆正規分布モデルについては、Barnett and Lewis (1994, pp.43-52)を参照されたい。

13 ベイズ尤度比を応用した n 数に関する斬新な検定については、藤原(2009)を参照されたい。

変数$\boldsymbol{x}$が混淆正規分布しているとは、$\boldsymbol{1-p}$の確率により平均$\boldsymbol{\mu}$、分散$\boldsymbol{\sigma^2}$の正規分布から生成される部分と確率$\boldsymbol{p}$により何らかの確率密度関数$\boldsymbol{g(x)}$により生成される部分から構成されることを意味する。

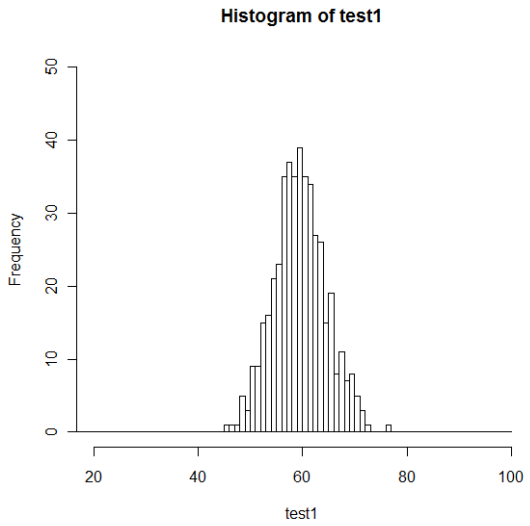
$$\boldsymbol{f(x) = (1-p)(2\pi\sigma^2)^{-1/2} exp\left(-\frac{1}{2\sigma^2}[x-\mu]^2\right) + pg(x)} \qquad \textbf{(2)}$$

もし確率密度関数$\boldsymbol{g(x)}$で汚染(contaminate)している側の分布の分散が大きい場合、あるいは、平均値が$\boldsymbol{\mu}$とは大幅に異なる場合、汚染している側の分布から得られた観測値は、他の観測値から大きく外れている可能性が高い(DeGroot and Schervish, 2002, p.577)。また、$\boldsymbol{g(x)}$の平均値が$\boldsymbol{\mu}$であれば峰は 1 つとなり、$\boldsymbol{g(x)}$の平均値が$\boldsymbol{\mu}$でなければ峰は 2 つとなる。

直感的な例[14]として、ある中学校で行われた 100 点満点の中間テストを考えてみる。テスト 1 は平均 60 点、標準偏差 5、受験者数 450 人であり、テスト 2 は平均 60 点、標準偏差 15、受験者数 50 人であると想定する。ここでは仮に、テスト 1 を真のデータと想定し、テスト 2 をエラーデータと想定してみる。具体的なイメージとしては、 テスト 1 は 3 年生全体の英語のテスト、テスト 2 は 3 年 1 組の理科のテスト、と考えてみる。すなわち、英語の全テスト結果の中に、ある特定のクラスの理科のテスト結果が混ざり、英語のテストの点数を汚染しているということである。

図 4.1 はテスト 1（真のデータ）を模した正規乱数のヒストグラムである。テスト 1 の最大値は 77、最小値は 46、中央値は 60、平均値は 59.99、標準偏差は 5.01 である[15]。

図 4.1



[14] 本節は混淆正規分布モデルの直感的なイメージを提示するもので、技術的な詳細は第 5 節に譲る。
[15] 本稿で用いたシミュレーションデータでは、小数点第二位が正確でない場合があるが、これは四捨五入によるものである。

図 4.2 はテスト 2（エラーデータ）を模した正規乱数のヒストグラムである。テスト 2 の最大値は 87、最小値は 13、中央値は 60、平均値は 59.96、標準偏差は 14.92 である。

図 4.2

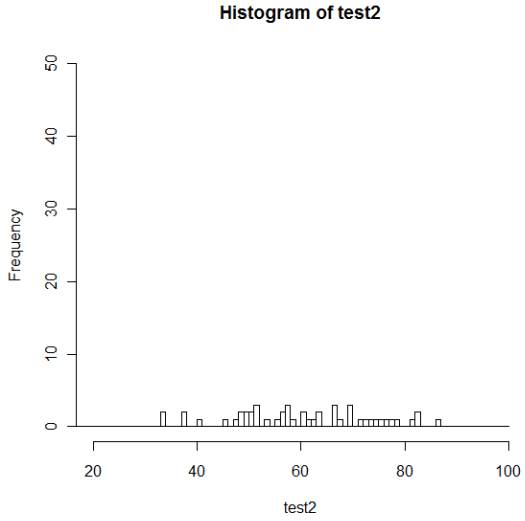


図 4.3 はテスト 1 とテスト 2 を模した正規乱数の結果が混在しているヒストグラムである。具体的には、図 4.1 と図 4.2 のデータを単純に結合したものである。図 4.3 のデータの最大値は 87、最小値は 13、中央値は 60、平均値は 59.99、標準偏差は 6.67 である。図 4.3 は、実際に観測者の手元にある観測データを表していると考えることができる。

図 4.3

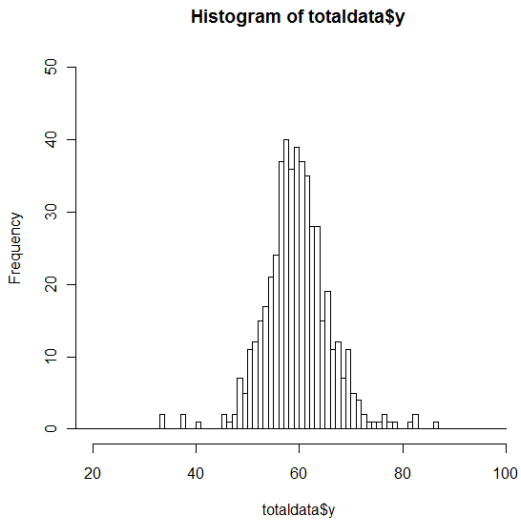


図 4.1 と図 4.3 を見比べると、45 以上 75 未満の分布はほとんど同じであることが視覚的に分かるであろう。一方、45 未満及び 75 以上の部分に影響力のある外れ値が存在することも視覚的に分かるであろう。図 4.2 から分かるように、エラーそのものは図 4.3 の分布全体に満遍なく存在している訳だが、潜在的に影響力の強い外れ値というものは、その中でも一部分であることが分かる。

具体的には、(1)ある観測値がエラーである可能性と(2)あるエラーの推定値に与える影響度合の 2 つを考慮することが重要である。今回の例は、シミュレーションであり、平均値と分散が分かっているので、(1)に関して z スコアを用いることができる。算出した z スコアが 0 に近ければ近いほど、観測値がエラーである可能性は低いと言える。

例えば、観測値の最大値である 87 がエラーである可能性を考えてみる。$Z = \frac{87-59.99}{5.01} = 5.391$であり、図 4.1 の分布から 87 という値が得られる可能性は極めてゼロに近いことが分かる。一方、分布の中央付近にある値を考えてみよう。例えば、観測値 63 の場合、$Z = \frac{63-59.99}{5.01} = 0.601$であり、図 4.1 の分布から 63 という値が得られる可能性は極めて高いことが分かる。すなわち、観測値 87 がエラーである可能性は非常に高く、観測値 63 がエラーである可能性は非常に低いと言える。

また、(2)に関して、今回の例では標準偏差の推定値について考えてみる。正しい標準偏差は 5.00 であるが、全データを用いた場合の標準偏差の推定値は 6.67 となっている。エラーによって標準偏差が影響を受け、真値の 1.33 倍になっていることが分かる。もし観測値 87 を削除したならば、標準偏差の推定値は 6.56 まで下がり、真の値に近づくこととなる。一方、観測値 63 を削除したならば、標準偏差の推定値は 6.67 となり、ほとんど変化がないことが分かる。つまり、観測値 87 は潜在的に影響力があり、観測値 63 は潜在的に影響力がないと言える。

ゆえに、今回の例では、観測値 87 はエラーである可能性が高く、かつ、潜在的に影響力の高い観測値であることが分かる。一方、観測値 63 はエラーである可能性が低く、かつ、潜在的に影響力の低い観測値であるということも分かる。

今回の例のように単変量の場合は、視覚的に解決することができるが、多変量の外れ値を検出する際には、視覚による検出は不可能となる。また、今回の例はシミュレーションなので、母集団における平均値と分散が分かっているが、実際のデータ分析においては、このような情報は利用可能ではなく、上述した手法で検出することはできない。したがって、多変量の文脈における影響力の強い外れ値検出法として、混淆正規分布モデルによる選択的エディティング手法に期待が寄せられる。

5 混淆正規分布モデルによる選択的エディティング手法

本節では、イタリア国家統計局による二変量の混淆正規分布モデル及びそれに基づく選択的エディティングの理論を詳細に解説する[16]。

5.1 二変量混淆(対数)正規分布モデル

5.1.1 使用する変数

今回のモデルでは、$\boldsymbol{W}$と$\boldsymbol{Z}$の 2 つの変数を考える。$\boldsymbol{W}$は、すでに補定とエディティングによりエラーの取り除かれた変数であり、$\boldsymbol{Z}$は測定誤差の影響を受けている変数である。また、$\boldsymbol{Z}^*$はエラーのない理論上の観測されない真の変数である。

もし$\boldsymbol{W}$と$\boldsymbol{Z}$が正規分布しているならば、そのまま使用すればよいが、正規分布していない場合には、何らかの変換を行って正規分布に近似させる必要がある。イタリア国家統計局による今回のモデルでは、対数正規分布を念頭に置いており、$\boldsymbol{Y}^*$は変数$\boldsymbol{Z}^*$の対数変換後の変数、$\boldsymbol{Y}$は変数$\boldsymbol{Z}$の対数変換後の変数、$\boldsymbol{X}$は変数$\boldsymbol{W}$の対数変換後の変数である。

$\boldsymbol{W}$：エラーが存在しない説明変数
$\boldsymbol{Z}$：測定誤差による影響を受ける被説明変数
$\boldsymbol{Z}^*$：観測されない真の被説明変数
$\boldsymbol{X}$：変数$\boldsymbol{W}$の対数変換後の変数、$\boldsymbol{X} = \log(\boldsymbol{W})$
$\boldsymbol{Y}$：変数$\boldsymbol{Z}$の対数変換後の変数、$\boldsymbol{Y} = \log(\boldsymbol{Z})$
$\boldsymbol{Y}^*$：変数$\boldsymbol{Z}^*$の対数変換後の変数、$\boldsymbol{Y}^* = \log(\boldsymbol{Z}^*)$

5.1.2 前提条件

$\{\boldsymbol{X} = \boldsymbol{x}\}$を条件として、$\boldsymbol{Y}^*$は平均値$\boldsymbol{\alpha} + \boldsymbol{\beta x}$と分散$\boldsymbol{\sigma^2}$で正規分布していると想定する。また、$\boldsymbol{\alpha}$、$\boldsymbol{\beta}$、$\boldsymbol{\sigma^2}$は推定するべきパラメータである。さらに、エラーが正規分布しているという仮定を追加し、加法エラーメカニズムは以下のように記述できる：$\boldsymbol{Y}^* \to \boldsymbol{Y}^* + \boldsymbol{\epsilon}$であり、$\boldsymbol{\epsilon} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\sigma_\epsilon^2})$である。

[16] 本節の内容は、主に Di Zio, Guarnera, and Luzi (2008)； Bellisai, Di Zio, Guarnera, and Luzi (2009)； Buglielli, Di Zio, and Guarnera (2010)に基づいている。多変量の文脈での理論的展開については、Buglielli, Di Zio, and Guarnera (2011)を参照されたい。

5.1.3 Y^*を条件とした場合のYの確率密度関数：混合密度

このモデルの重要な特徴として、エラーが断続的であるという点が挙げられる。エラーが断続的であるとは、エラーはすべてのデータに影響を与えるのではなく、一部のデータにのみ影響を与えているということである。つまり、観測データの分布は、エラーのない真のデータを条件として、2 つの確率分布の混合として表すことができるということである。すなわち、式(3)のように定式化できる。

$$f_{Y|Y^*}(y|y^*) = (1-p)\delta_{y^*} + pN(y; y^*, \sigma_\epsilon^2) \quad (3)$$

Y^*を条件としたYの分布は、$1-p$の確率でデルタ関数δに属し、pの確率で平均y^*、分散σ_ϵ^2の正規分布に属している。ここで、pは混合重み(Mixing Weight)であり、エラーの「事前」確率を表し、δはディラックのデルタ関数であり、$y \neq y^*$の場合にゼロとなり、$y = y^*$の場合に無限大となる関数である。

5.1.4 Xを条件としたYの測定値の混淆正規分布

5.1.2 節において示した前提条件より、Xを条件としたYの測定値は式(4)の混淆正規分布となる。

$$f_{Y|X}(y|x) = (1-p)N(y; \alpha + \beta x, \sigma^2) + pN(y; \alpha + \beta x, \sigma_c^2) \quad (4)$$

すなわち、説明変数xを条件とした場合の被説明変数yの分布は、確率$1-p$により平均$\alpha + \beta x$、分散σ^2の正規分布により生成されており（真のデータ）、確率pにより平均$\alpha + \beta x$、分散σ_c^2の正規分布により生成されている（エラーデータ）。この式は、同じ切片と同じ傾きを持つが異なる残差分散を持つ 2 つの回帰モデルを表している。ここで、σ^2は正しいデータの分散を表し、$\sigma_c^2 = \sigma^2 + \sigma_\epsilon^2$は汚染されたデータ、つまり、エラーデータの分散を表す。

5.1.5 パラメータ推定

観測値$Y = \{y_i : i = 1 \dots n\}$に対し、パラメータ$\alpha, \beta, \sigma^2, \sigma_c^2$は、$p$が与えられていれば$y_i$の同時分布の確率密度(式 5)を最大化することにより、最尤法(MLE)で求められる。

$$\prod_{i=1}^{n} \{(1-p)N(y; \alpha + \beta x, \sigma^2) + p\, N(y; \alpha + \beta x, \sigma_c^2)\} \quad (5)$$

一方、pについては、y_iが汚染データに属することの事後確率がベイズの定理により式(6)となるため、これの期待値が pに一致するという制約条件がある。

$$\frac{p\,N(y;\alpha+\beta x,\sigma_c^2)}{(1-p)N(y;\alpha+\beta x,\sigma^2)+p\,N(y;\alpha+\beta x,\sigma_c^2)} \tag{6}$$

すなわち、式(7)となる。

$$\begin{aligned} p &= \mathrm{E}\left[\frac{p\,N(y;\alpha+\beta x,\sigma_c^2)}{(1-p)N(y;\alpha+\beta x,\sigma^2)+p\;N(y;\alpha+\beta x,\sigma_c^2)}\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{p\;N(y;\alpha+\beta x,\sigma_c^2)}{(1-p)N(y;\alpha+\beta x,\sigma^2)+p\;N(y;\alpha+\beta x,\sigma_c^2)} \end{aligned} \tag{7}$$

この制約条件がついた最適解は解析的には求められない。上式より、パラメータの仮の推計値から期待値 p を計算する過程と、計算された p の下で最尤法(MLE)によりパラメータを求める過程を収束するまで繰り返すことにより求めるアルゴリズムを EM アルゴリズムと言う。

5.2 ECM アルゴリズム

4 節で示したテストの例のように、真のデータとエラーのデータの区別が分からないような問題を不完全データ問題と呼ぶ。4 節では、平均と分散の値があらかじめ分かっていることを想定していたが、現実には、不完全なデータから母集団における平均や分散を推定しなければならない。不完全なデータを完全なものにするためには、平均や分散といった分布に関する情報が必要となるが、その平均や分散を推定するのに不完全データを使うというイタチごっこになってしまう。これを解決するために、何らかの手段によって初期値を決め、そこから繰り返し法を用いて推定を行う方法が提唱されてきた。その代表例が EM (Expectation Maximization：期待値最大化)アルゴリズムである。

EM アルゴリズムとは、期待値計算を行う Expectation ステップ（E ステップ）と尤度最大化計算を行う Maximization ステップ（M ステップ）から構成され、不完全データから最尤推定値(MLE)を導く一般的なアルゴリズムである。EM アルゴリズムでは、まず仮の平均値や分散を決め、分布を仮定して初期値を設定する。この初期値からモデル尤度の期待値を計算し、尤度の最大化計算を行い、得られた期待値を最大化するパラメータを推定し、分布を更新する。期待値計算及び最大化計算を繰り返し、最終的に収束した値が最尤推定値(MLE)である（渡辺, 山口, 2000, p.29; Gill, 2008, p.309）。

ECM アルゴリズムとは、Expectation Conditional Maximization（期待値条件付最大化

法）の略であり、文字通り、EM アルゴリズムの M ステップを条件付最大化ステップ（CM ステップ）に置き換えたものである。ECM アルゴリズムは、複数のパラメータが存在する場合に、その一部のパラメータが与えられた条件のもとで尤度の最大化を行うことで、M ステップを単純化することができる（渡辺, 山口, 2000, p.120）。

以下では、n 個の二変量測定値$(\boldsymbol{x_i}, \boldsymbol{y_i})$の標本に基づいて、モデルパラメータの推定手順を示す。ここで $i = 1,\ldots,n$ である。下記で示す E ステップと CM ステップの繰り返し適用による ECM アルゴリズムを使用して、最尤推定値(MLE)を求める。

5.2.1 推定すべきパラメータ

ECM アルゴリズムで求めるべきパラメータの最尤推定値は、事前確率である混合重み$\boldsymbol{p}$及び回帰係数$\boldsymbol{\alpha, \beta, \sigma^2, \sigma_c^2}$である。アルゴリズムを初期化するために、全データを用いた通常線形回帰による$\boldsymbol{\alpha, \beta, \sigma^2}$の推定値を初期値として用いる。また、$\boldsymbol{p}$の初期値[17]としては、0 と 0.4 の間のランダムな値を用い、$\boldsymbol{\sigma_c^2}$の初期値は$\boldsymbol{\lambda\sigma^2}$とし、$\boldsymbol{\lambda}$は 5 から 100 までの値である。推定すべきパラメータの一覧は以下のとおりである。

$\boldsymbol{p}$：事前確率（混合重み）
$\boldsymbol{\alpha}$：切片
$\boldsymbol{\beta}$：傾き
$\boldsymbol{\sigma^2}$：真のデータの分散
$\boldsymbol{\sigma_c^2}$：エラーデータの分散

5.2.2 E ステップ

E ステップでは、事後確率$\boldsymbol{\tau_i = \tau(y_i; x_i)}$の更新を行う。$\boldsymbol{\tau(y; x)}$は式(8)のとおり定義される。

$$\boldsymbol{\tau(y; x) = \frac{pN(y; \alpha + \beta x, \sigma_c^2)}{(1-p)N(y; \alpha + \beta x, \sigma^2) + pN(y; \alpha + \beta x, \sigma_c^2)}} \quad \textbf{(8)}$$

すなわち、事後確率とは、説明変数$\boldsymbol{x}$を条件とした場合に事前確率$\boldsymbol{1-p}$により平均$\boldsymbol{\alpha + \beta x}$、分散$\boldsymbol{\sigma^2}$の正規分布により生成されている真のデータと事前確率$\boldsymbol{p}$により平均$\boldsymbol{\alpha + \beta x}$、分散$\boldsymbol{\sigma_c^2}$

[17] Bellisai, Di Zio, Guarnera, and Luzi (2009, p.5)によると、「pの初期値として 0.6 と 1 の間のランダムな値を用いる」とあるが、これは、「$1-p$の初期値として 0.6 と 1 の間のランダムな値を用いる」の誤りである。もし$p \geq 0.5$の場合、正データとエラーデータの区別がつかなくなるので、$p < 0.5$と想定しなければならない(DeGroot and Schervish, 2002, p.577)。次節で紹介する R の *SeleMix* パッケージの既定では 0.05 となっている。

の正規分布により生成されているエラーデータにより求められる被説明変数$\boldsymbol{y}$の分布に占めるエラーデータの割合である。

5.2.3 CM ステップ

CM ステップでは、以下に示す[1]から[3]の手順で更新を行う。

[1] 混合重み$(\boldsymbol{p})$の更新：式(9)に示すとおり、混合重み$(\boldsymbol{p})$を事後確率$(\boldsymbol{\tau_i})$の平均として更新する。

$$\boldsymbol{p} = \frac{1}{n}\sum_{i=1}^{n}\tau_i \tag{9}$$

[2] 切片$(\boldsymbol{\alpha})$と傾き$(\boldsymbol{\beta})$の更新：式(10)を用いて傾き$(\boldsymbol{\beta})$を更新し、式(11)で切片$(\boldsymbol{\alpha})$を求めて更新する。

$$\boldsymbol{\beta} = \frac{\sum_{i=1}^{n}\{(\boldsymbol{\tau_i}/\boldsymbol{\sigma^2})(\boldsymbol{y_i}-\widetilde{\boldsymbol{y}})(\boldsymbol{x_i}-\widetilde{\boldsymbol{x}}) + [(\boldsymbol{1}-\boldsymbol{\tau_i})/\boldsymbol{\sigma_c^2}](\boldsymbol{y_i}-\widetilde{\boldsymbol{y}})(\boldsymbol{x_i}-\widetilde{\boldsymbol{x}})\}}{\sum_{l=1}^{n}\{(\boldsymbol{\tau_l}/\boldsymbol{\sigma^2})(\boldsymbol{x_l}-\widetilde{\boldsymbol{x}})^2 + [(\boldsymbol{1}-\boldsymbol{\tau_l})/\boldsymbol{\sigma_c^2}](\boldsymbol{x_l}-\widetilde{\boldsymbol{x}})^2\}} \tag{10}$$

$$\boldsymbol{\alpha} = \widetilde{\boldsymbol{y}} - \boldsymbol{\beta}\widetilde{\boldsymbol{x}} \tag{11}$$

ここで、$\widetilde{\boldsymbol{y}}$と$\widetilde{\boldsymbol{x}}$の定義は、それぞれ、式(12)と(13)のとおりである。

$$\widetilde{\boldsymbol{y}} = \frac{(\boldsymbol{\tau_i}/\boldsymbol{\sigma^2})\boldsymbol{y_i} + [(\boldsymbol{1}-\boldsymbol{\tau_i})/\boldsymbol{\sigma_c^2}]\boldsymbol{y_i}}{(\boldsymbol{\tau_l}/\boldsymbol{\sigma^2}) + [(\boldsymbol{1}-\boldsymbol{\tau_l})/\boldsymbol{\sigma_c^2}]} \tag{12}$$

$$\widetilde{\boldsymbol{x}} = \frac{\{(\boldsymbol{\tau_i}/\boldsymbol{\sigma^2})\boldsymbol{x_i} + [(\boldsymbol{1}-\boldsymbol{\tau_i})/\boldsymbol{\sigma_c^2}]\boldsymbol{x_i}\}}{(\boldsymbol{\tau_l}/\boldsymbol{\sigma^2}) + [(\boldsymbol{1}-\boldsymbol{\tau_l})/\boldsymbol{\sigma_c^2}]} \tag{13}$$

[3] 残差分散$(\boldsymbol{\sigma^2}, \boldsymbol{\sigma_c^2})$の更新：更新された切片$(\boldsymbol{\alpha})$と傾き$(\boldsymbol{\beta})$をもとに、式(14)と(15)により残差分散$(\boldsymbol{\sigma^2}, \boldsymbol{\sigma_c^2})$を更新する。

$$\boldsymbol{\sigma^2} = \left(\sum_{l=1}^{n}\boldsymbol{\tau_{l1}}\right)^{-1}\sum_{i=1}^{n}\boldsymbol{\tau_i}\,(\boldsymbol{y_i}-\boldsymbol{\alpha}-\boldsymbol{\beta x_i})^2 \tag{14}$$

$$\sigma_c^2 = \left(\sum_{l=1}^{n} \tau_{l2}\right)^{-1} \sum_{i=1}^{n} (1-\tau_i)(y_i - \alpha - \beta x_i)^2 \quad (15)$$

5.2.4 最尤推定値(MLE)

以上の E ステップ及び CM ステップの手続きを繰り返し、収束した値が最尤推定値となる。つまり、初期値から E ステップにおいて期待値を求め、次にその期待値を CM ステップにおいて最大化を行って初期値を更新する。この更新された値を新たな初期値として再び期待値を求め、再び最大化を行って更新を行い、収束するまで繰り返すことで得られた値は、局所的最大値であることが証明されている。複数の峰のある分布においては、得られた解が大局的最大値であるとは限らないので、複数の初期値から ECM アルゴリズムを行い、複数の収束した値の中から最も大きいものを選ぶ必要がある(渡辺, 山口, 2000, p.40; 中村, 小西, 1998, p.168)。しかし、今回の場合、平均値が同じ 2 つの分布を想定しているので、自ずと単峰の分布となるため、収束した値は大局的最大値であると保証できる。

5.3 混淆正規分布による選択的エディティング

本節では、これまでに紹介した混淆正規分布モデルを使うことで、どのようにして選択的エディティングを行うのかについて説明をする。

5.3.1 y_iを条件とするy_i^*の条件付分布

混淆正規分布による選択的エディティングを行うには、$\boldsymbol{X}$を含む観測データを条件として、エラーのないデータ$\boldsymbol{Y}^*$の分布を導かなければならない。5.1.3 節でエラーモデルの式(3)を以下のとおり紹介した。式(3)をもとに、5.1.5 節で説明したとおり、ベイズの公式を用いることで、各々の観測値(x_i, y_i)に関してy_iを条件とするy_i^*の条件付分布を式(16)として求めることができる。

$$f_{Y|Y^*}(y|y^*) = (1-p)\delta_{y^*} + pN(y; y^*, \sigma_\epsilon^2) \quad (3)$$

$$f_{Y^*|Y}(y_i^*|y_i) = [1-\tau(y_i; x_i)]\delta_{y_i} + \tau(y_i; x_i)N(y_i^*; \tilde{\mu}_i, \tilde{\sigma}^2) \quad (16)$$

ここで、$\tilde{\sigma}^2$と$\tilde{\mu}_i$の定義は式(17)と(18)のとおりである。

$$\tilde{\sigma}^2 = (\sigma^{-2} + \sigma_\epsilon^{-2})^{-1} \quad (17)$$

$$\tilde{\mu}_i = \tilde{\sigma}^2(y_i/\sigma_\epsilon^2 + (\alpha + \beta x_i)/\sigma^2) \quad (18)$$

5.3.2 元々の尺度のデータに再変換

5.1.1 節において定義したとおり、変数$\boldsymbol{Z}$を対数変換したものが変数$\boldsymbol{Y}$であった。したがって、元々の尺度のデータ$\boldsymbol{Z}$の分布を導くことが必要である。式(16)に$\boldsymbol{Y^* = \log(Z^*)}$と$\boldsymbol{Y = \log(Z)}$を代入することにより、式(19)を導くことができる。

$$\boldsymbol{f_{Z^*|Z}(z_i^*|z_i) = [1 - \tau(\log(z_i); x_i)]\delta_{\log(z_i)} + \tau(\log(z_i); x_i) LN(z_i^*; \tilde{\mu}_i, \tilde{\sigma}^2)} \quad \textbf{(19)}$$

5.3.3 期待誤差とスコア関数

式(19)より、パラメータ$\boldsymbol{p, \alpha, \beta, \sigma^2, \sigma_\epsilon^2 = \sigma^2 - \sigma_c^2}$を、該当する ECM 推定値に置き換え、観測値$\boldsymbol{z_i}$を条件として、真の値$\boldsymbol{z_i^*}$の予測値$\boldsymbol{\hat{z}_i}$を式(20)のとおり導くことができる。

$$\boldsymbol{\hat{z}_i = E(z_i^*|z_i) = \int z_i^* f_{Z^*|Z}(z_i^*|z_i) dz_i^*} \quad \textbf{(20)}$$

したがって、期待誤差(*EE*: Expected Error)は式(21)のとおりとなる。

$$\boldsymbol{EE = (\hat{z}_i - z_i)} \quad \textbf{(21)}$$

これらの推定値に基づいて、有限母集団量のロバスト推定と選択的エディティングを行うことができる。具体的には、$\boldsymbol{w_i}$を各々のユニットへの標本ウェイトとする。スコア関数$\boldsymbol{SF_i}$は式(22)として定義される。すなわち、スコア関数は、ある値の予測値と観測値の重み付き差異が、重み付き予測値の総和に占める割合の絶対値である。

$$\boldsymbol{SF_i = \left| \frac{w_i(\hat{z}_i - z_i)}{\sum w_i \hat{z}_i} \right|} \quad \textbf{(22)}$$

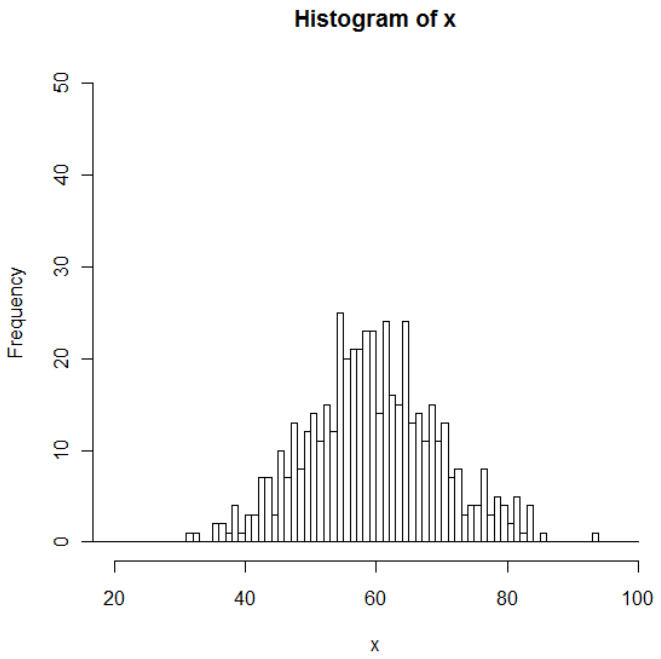
6 *SeleMix* のデモンストレーション：シミュレーションデータ[18]

4 節では、直感的な例として、ある中学校で行われた 100 点満点の中間テストを考えてみた。テスト 1 は平均 60 点、標準偏差 5、受験者数 450 人であり、テスト 2 は平均 60 点、標準偏差 15、受験者数 50 人と想定した。仮にテスト 1 を真のデータと想定し、テスト 2 をエラーデータと想定した。また、具体的なイメージとして、 テスト 1 を 3 年生全体の英語のテスト、テスト 2 を 3 年 1 組の理科のテストと考えてみた。また 5 節では、混淆正規分布モデルによる選択的エディティングの理論を紹介した。

本節では、上記のテストの例を用いて混淆正規分布モデルによる選択的エディティングを実際に行ってみる。統計環境 R の *SeleMix* パッケージを用いた混淆正規分布モデルによる外れ値検出のデモンストレーションを行う。

まず、R の `rnorm` 関数を用いて、正規乱数を生成し、必要なデータセットを作成する。テスト 1、テスト 2、及び混在データは 4 節と同じである。さらに、説明変数$\boldsymbol{x}$を以下のように求める：$\boldsymbol{x} = -\mathbf{60} + \mathbf{2} * \mathbf{test1} + \boldsymbol{e}$であり、 $\boldsymbol{x}$の観測数は 500 である。ここで$\boldsymbol{e}$はランダムエラーであり、$\boldsymbol{x}$の最大値は 94、最小値は 32、中央値は 60、平均値は 60.02、標準偏差は 10.08 である。テスト 1、テスト 2、及び混在データのヒストグラムは 4 節と同じである。$\boldsymbol{x}$のヒストグラムは図 6.1 のとおりである。

図 6.1



[18] *SeleMix* パッケージの関数については、Guarnera and Buglielli (2012)を参照されたい。

これらのデータセットを作成した R のコードは以下のとおりである。まず、再現性を保つために seed を設定する。

```
set.seed(123)
```

次に、正規乱数を 450 個と 50 個ずつ生成する。

```
u1<-rnorm(450)
u2<-rnorm(50)
```

これらの正規乱数を標準化し、平均値を 0 に、標準偏差を 1 にする。

```
v1<-(u1-mean(u1))/sd(u1)
v2<-(u2-mean(u2))/sd(u2)
```

test1 と test2 として、それぞれ平均値 60 と標準偏差 5 または 15 の観測値を生成し、小数点下 0 桁に四捨五入する。

```
test1<-round(60+5*v1, 0)
test2<-round(60+15*v2, 0)
```

そして、これらのデータを total として 1 つのリストにし、totaldata としてデータ化し、便宜のために変数名を y と index1 に変更する。

```
total<-list(Test1=test1,Test2=test2)
totaldata<-stack(total)
names(totaldata)<-c("y","index1")
```

次に、上述したとおり、説明変数 x を生成する。e は 500 個のランダムな正規乱数である。最後に、y を含んでいる totaldata と x を `cbind` で 1 つのデータとしてまとめる。

```
e<-rnorm(500)
x<-round(-60+2*test1+e,0)
ex1.data<-cbind(x, totaldata)
```

出来上がったデータセットの最初の 5 つの観測値を以下に示す。変数 x と変数 y、そし

て、それぞれの観測値が test1 と test2 のどちらに属しているかを示す index1 が付されている。

```
> ex1.data
    x  y   index1
1   53  57  Test1
2   57  59  Test1
3   76  68  Test1
4   60  60  Test1
5   59  61  Test1
```

上述の変数 ***x*** を条件として混淆正規分布による二変量外れ値検出を行う。以下、*SeleMix* の関数を示しながら、具体的な使用法を論ずる。3 節で説明した手順により、既に R に *SeleMix* パッケージをインストール済みであると仮定して議論を進める。

6.1 *SeleMix* 起動

```
library(SeleMix)
```

R に *SeleMix* パッケージをインストールした後、分析を行う前に、上述のコマンドにより *SeleMix* パッケージを起動してから分析を行う。

6.2 混淆正規分布モデルによる外れ値検出

```
ml.par<-ml.est(y, x= variable, model= "N", lambda=3, w=0.05, graph=TRUE)
```

`ml.est` は混淆正規分布モデルの最尤推定値を返す関数である。ここで、`ml.par` は任意の変数名である。`y` は外れ値を含む変数の名前であり、`x=`の右辺は条件とする説明変数である。`model=`の右辺は`"N"`または`"LN"`であり、`N` は正規分布を、`LN` は対数正規分布を指す。`lambda` は分散拡大要因(VIF)の初期値であり、既定では 3 となっている。`w` は汚染データの割合の初期値であり、既定では 0.05 となっている。また、`graph=TRUE` とすると、EM アルゴリズムを視覚的にモニターすることができ、検出された外れ値が表示される。

図 6.2 は、以上のルールにしたがって今回のテストの例を分析した結果である。図 6.2a の横軸は ***x*** であり、縦軸はテストのスコアである。白丸は通常の観測値であり、黒丸は外れ値である。図 6.2b は、収束するまでにかかった繰り返しの回数であり、今回の例では 9

回で収束している。

```
ml.par<-ml.est(ex1.data$y, x=ex1.data$x, model= "N", w=0.1,
graph=TRUE)
```

図 6.2a

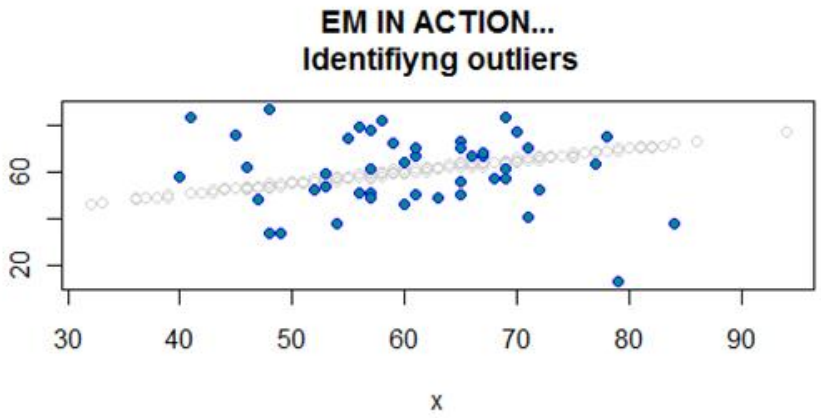
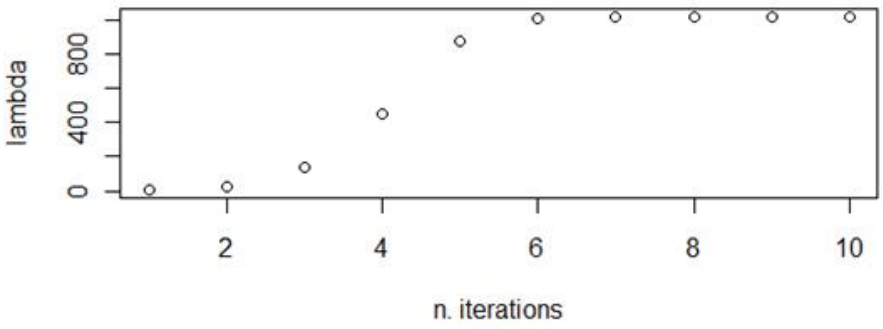


図 6.2b



6.3 パラメータ推定値の表示

```
str(ml.par)
```

上述の関数により、混淆正規分布の最尤推定値を表示することができる。今回の例では、以下の出力結果が得られた。中でも重要なものは、`B`（回帰係数の推定値）、`sigma`（分散共分散行列の推定値）、`lambda`（VIF の推定値）、`w`（エラーデータの割合の推定値）である。

また、`ypred` は変数 y の予測値の行列、`tau` は汚染に関する事後確率のベクトルである。

もし観測値が外れ値として分類されていれば outlier は 1 に、そうでなければ 0 となっている。pattern は非回答項目があるかどうかを示し、0 であれば欠測しており、1 であれば観測されている。is.conv は EM アルゴリズムが収束したかどうかを表しており、収束していれば TRUE となっている。n.iter は EM アルゴリズムが収束するまでに要した回数である。

bic.aic はベイズ情報量基準(BIC)及び赤池情報量基準(AIC)である。今回の例では、通常正規モデルによる BIC が 3059 である一方、混淆正規分布モデルによる BIC は 1394 となっている。また、通常正規モデルによる AIC が 1526 である一方、混淆正規分布モデルによる AIC は 691 となっている。情報量基準は、値の小さい方が優れていることを意味するので、BIC と AIC のいずれの基準においても、混淆正規分布モデルの優位が支持されている。

```
List of 11
 $ ypred  : num [1:500, 1] 57 59 68 60 60.6 ...
 $ B      : num [1:2, 1] 30.428 0.493
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:2] "" "x"
  .. ..$ : NULL
 $ sigma  : num [1, 1] 0.244
 $ lambda : num 1017
 $ w      : num 0.104
 $ tau    : num [1:500] 0.00542 0.00572 0.0037 0.00361 0.25223 ...
 $ outlier: num [1:500] 0 0 0 0 0 0 0 0 0 0 ...
 $ pattern: Factor w/ 1 level "1": 1 1 1 1 1 1 1 1 1 1 ...
 $ is.conv: logi TRUE
 $ n.iter : num 9
 $ bic.aic: Named num [1:4] 3059 1394 1526 691
  ..- attr(*, "names")= chr [1:4] "BIC.norm" "BIC.mix" "AIC.norm"
"AIC.mix"
 - attr(*, "model")= chr "N"
 - attr(*, "class")= chr [1:2] "list" "mlest"
```

6.4 検出した外れ値の数

```
sum(ml.par$outlier)
[1] 47
```

上述の関数を用いることで、検出された外れ値の数を知ることができる。今回の例はシミュレーションなので、外れ値が 50 個存在していることが始めから分かっているが、混淆正規分布モデルに基づく *SeleMix* によって 47 個の外れ値が検出された。検出された 47 個はすべてエラーデータに属すもので、外れ値として正しく検出されたと言える。今回のシミュレーションでは、正データの中から誤って外れ値として検出されたものはなかった。

6.5 外れ値を含む変数(y)の予測値の算出

```
B1<-as.matrix(c(30.428, 0.493))
sigma1<-as.matrix(0.244)
lambda1<-1017
w1<-0.104
ypred<-pred.y(ex1.data$y,x=ex1.data$x,B=B1,sigma=sigma1,lambda=lamb
da1,w=w1,model="N",t.outl=0.5)
```

6.3 節で得られたパラメータ推定値を上記のように入力し、y の予測値を求める。これらの数値は、データセットごとに異なるので、都度、手作業で入力して推定を行う。`pred.y` は、y の予測値を求める関数である。`B` は係数の推定値、`sigma` は分散共分散行列の推定値、`lambda` は VIF の推定値、`w` はエラーデータの割合の推定値である。また、`t.outl` は外れ値検出をする際の事後確率の閾値であり、既定では 0.5 となっている。

6.6 y の予測値と y の観測値の図

```
sel.pairs(cbind(ypred[,1,drop=FALSE],ex1.data[,"y",drop=FALSE]),out
l=ypred[,"outlier"])
```

上述の関数 `sel.pairs` を用いることで、y の予測値と y の観測値の図を作成できる。ここで、`ex1.data` は今回用いたデータセットの任意の名称である。出力した結果は、以下の 3 つの図である。図 6.3 は、y の予測値の箱ひげ図である。図 6.4 は、y の観測値の箱ひげ図である。図 6.5 は、y の観測値と y の予測値の散布図である。ここで、白丸は通常の観測値であり、黒丸は外れ値である。

図 6.3： y の予測値の箱ひげ図（黒丸＝外れ値）

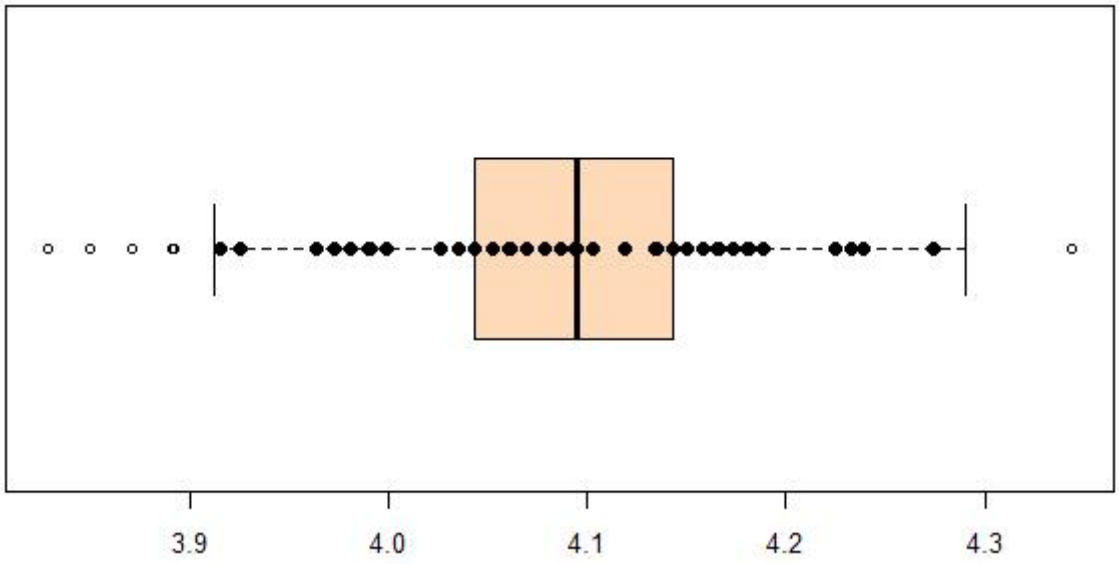


図 6.4：y の観測値の箱ひげ図（黒丸＝外れ値）

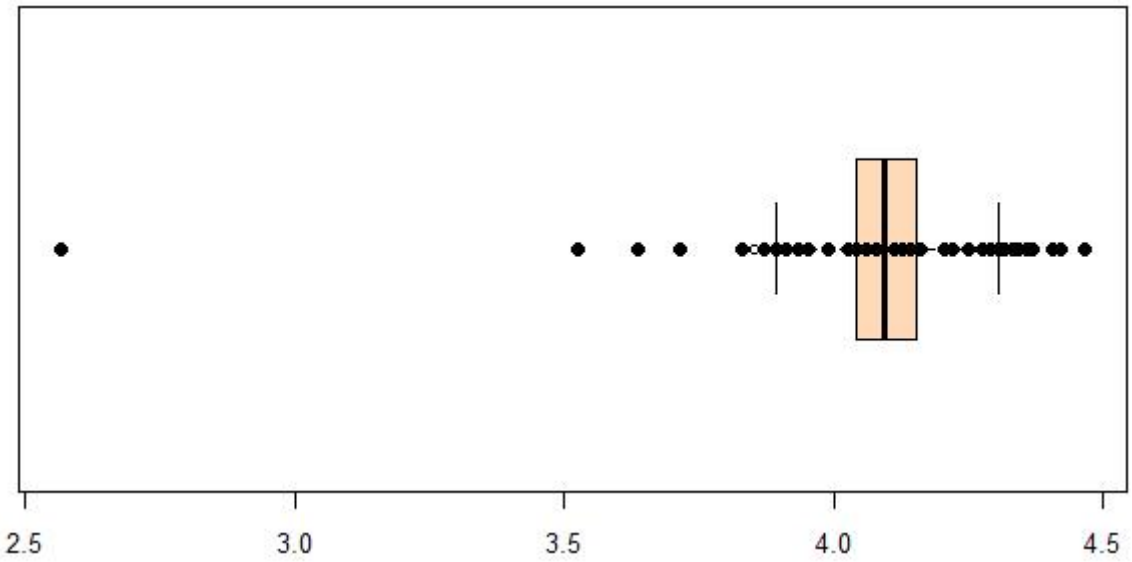
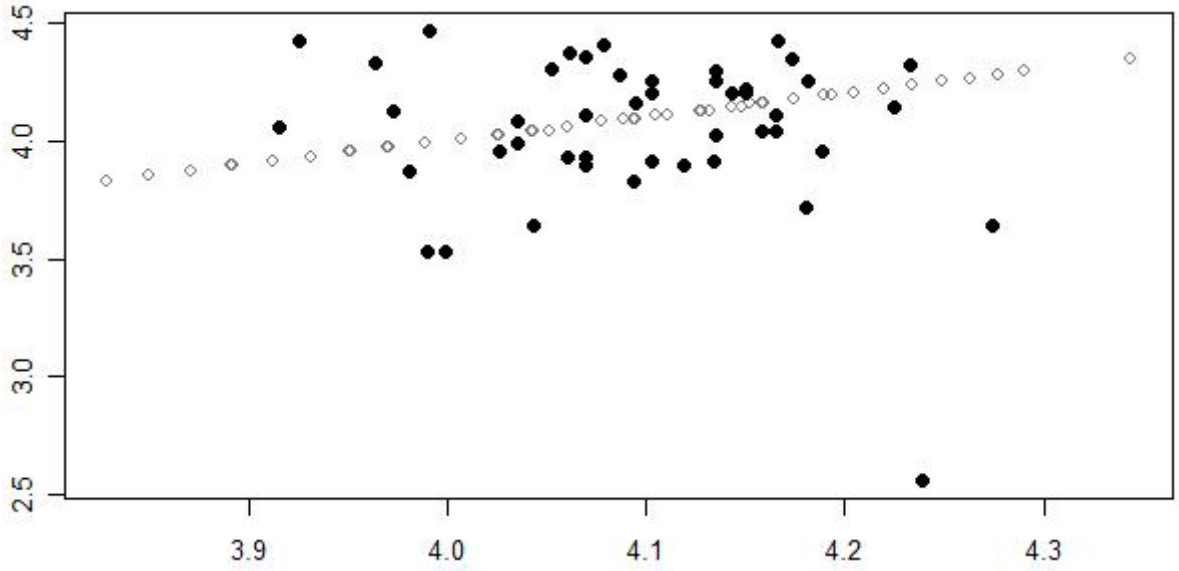


図 6.5：y の観測値と y の予測値の散布図（黒丸＝外れ値）



6.7 影響力のある外れ値検出

```
sel0001<-sel.edit(ex1.data$y,ypred=ml.par$ypred,t.sel=0.001)
sum(sel0001[,"sel"])
[1] 9
```

sel.edit は、影響力のある外れ値を表示する関数である。t.sel は、選択的エディティングのための閾値であり、既定では 0.01 であるが、今回は影響力のある外れ値を検出できるよう、デモ用に 0.001 にセットした。今回の例では、47 個の外れ値が検出されたが、実際に影響力があると判定されたのはそのうちの 9 個であったことが分かる。

6.8 優先スコアの降順で結果を並び替え

以下の手順で、影響力のある外れ値を表示させる（表 6.1）。

```
sel.ord<-sel0001[order(sel0001[,"rank"]),]
sel.ord
```

表 6.1

y1	y1.p	weights	y1.score	global.score	y1.reserr	y1.sel
13	69.31	1	1.88e-03	1.88e-03	-4.49e-04	1
38	71.80	1	1.13e-03	1.13e-03	1.43e-03	1
87	54.12	1	1.10e-03	1.10e-03	2.55e-03	1
83	50.67	1	1.08e-03	1.08e-03	1.46e-03	1
41	65.40	1	8.13e-04	8.13e-03	3.81e-04	1
76	52.63	1	7.79e-04	7.79e-04	1.19e-03	1
82	59.04	1	7.65e-04	7.65e-04	4.15e-04	1
79	58.05	1	6.98e-04	6.98e-04	-3.50e-04	1
34	54.56	1	6.85e-04	6.85e-04	-1.05e-03	1

y1 は観測値であり、y1.p は予測値である。weights は標本ウェイトである。y1.score はローカルスコアであり、global.score はグローバルスコアである。y1.reserr は残差エラーの値であり、y1.sel は影響力のある外れ値に付されるフラグである。4 節で予想したとおり、y1 の値が 45 未満及び 75 以上の部分に影響力のある外れ値が存在している。

6.9 外れ値及び影響力のあるエラーの図示

```
ex1.data<-cbind(ex1.data,tau=ml.par$tau,outlier=ml.par$outlier,sel0
001[,c("rank","sel")])
sel.pairs(ex1.data[,c("x","y")],outl=ml.par$outlier,sel=sel0001[,"s
el"])
```

まず、cbind 関数を用いて、元々のデータに列を追加し、そこに影響力のある外れ値に関する情報を格納する。その後、sel.pairs 関数を用いて図示する。白丸は通常の観測値及び外れ値であり、菱形は影響力のある外れ値である。図 6.6 は x の箱ひげ図であり、図 6.7 は y の箱ひげ図である。図 6.8 は、x と y の散布図である。必ずしもすべての外れ値が、影響力のある外れ値であるとは限らないことが分かる。

図 6.6：x の箱ひげ図（菱形＝影響力のある外れ値）

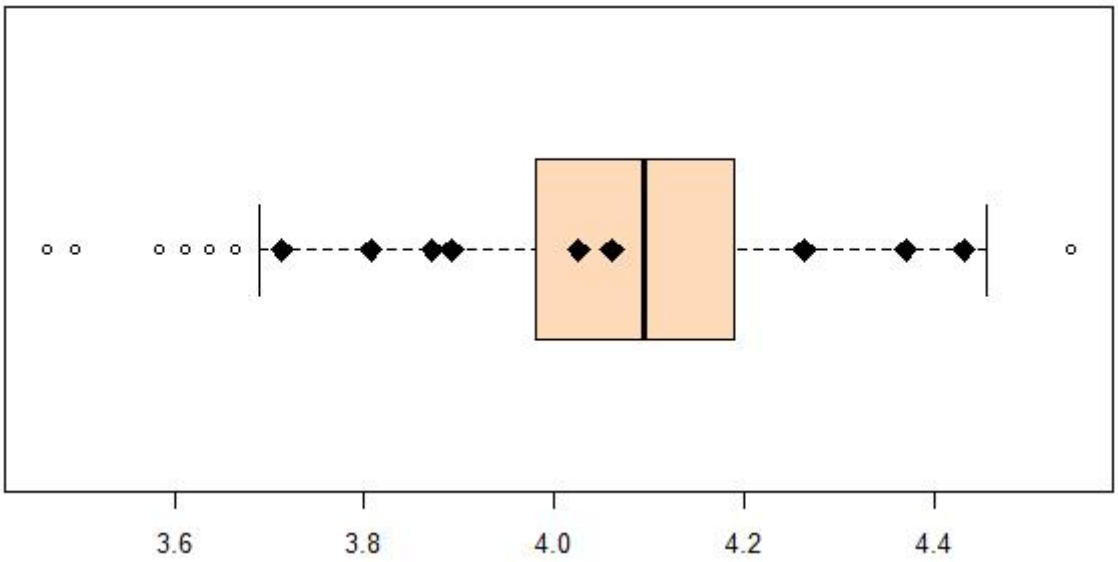


図 6.7：y の箱ひげ図（菱形＝影響力のある外れ値）

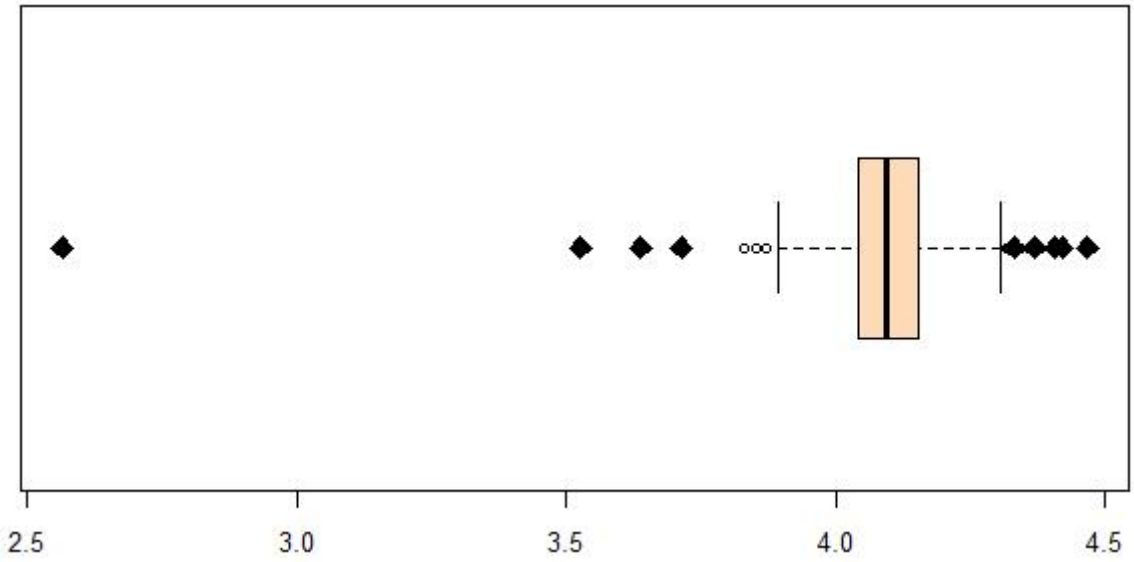
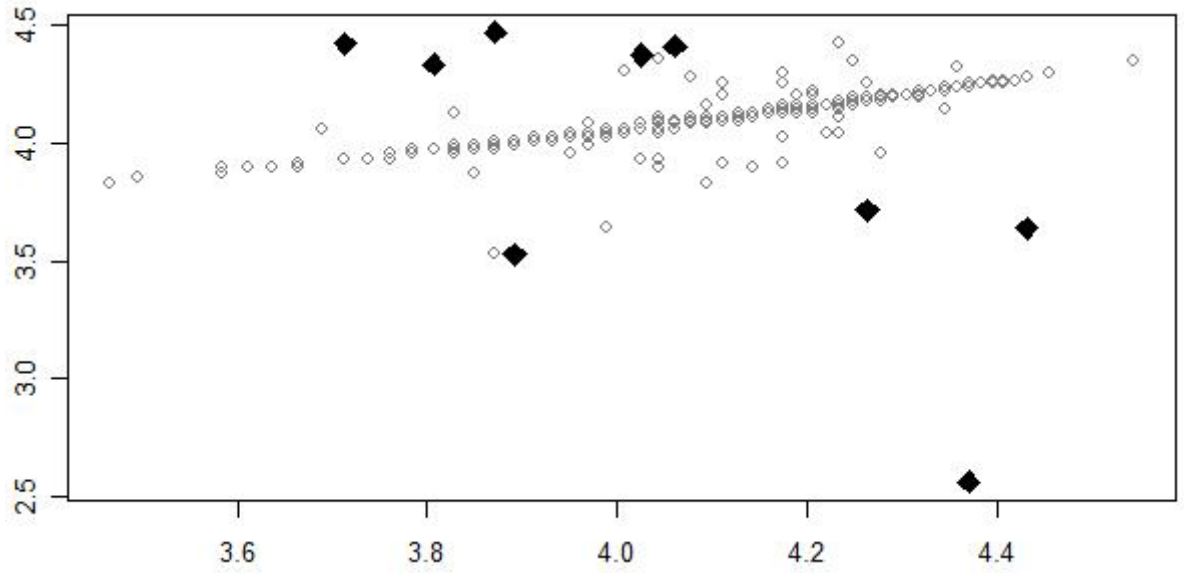


図 6.8：x と y の散布図（菱形＝影響力のある外れ値）



7 *SeleMix* のデモンストレーション：EDINET データ

前節までは、理論及びシミュレーションに基づいていたが、本節では、実データを使用した多変量外れ値の検出を行う。また、本節は、経済センサス-活動調査の経理項目のエディティングに向けた研究の一環である。ただし、経済センサス-活動調査の実データは、まだ利用可能ではないので、EDINET[19]のデータを模擬試験データとして利用する。

今回の例では、欠測値を除外した 3,042 レコードを使用する。*SeleMix* による外れ値検出では、Y の欠測値は自動的に補定を行うことができるため、Y の欠測値は容認できるが、X はエラーのない完全なデータであることが要求されるため、今回は実験のために X の欠測値をすべて除外した[20]。

対象となる変数は、経済センサス-活動調査票における「売上（収入）金額」、「常用雇用者」、「資本金又は出資金、基金の額」である。上述したとおり、経済センサス-活動調査のデータは利用可能ではないので、これら 3 つの変数に該当する EDINET のデータ、すなわち、「売上高」、「事業従事者数」、「資本金」を使用する。対応関係は表 7.1 に示すとおりである。

[19] 「**EDINET**（**E**lectronic **D**isclosure for **I**nvestors' **NET**work）は、『金融商品取引法に基づく有価証券報告書等の開示書類に関する電子開示システム』のことで、提出された開示書類について、インターネット上においても閲覧を可能とするもの」である。http://info.edinet-fsa.go.jp/

[20] これは、実際に選択的エディティングを行う際には、X の欠測値を補定によって埋めておく必要があることを意味している。

表 7.1：使用する変数名

経済センサス-活動調査	EDINET	英語名
売上（収入）金額	売上高	turnover
常用雇用者	事業従事者数	worker
資本金又は出資金、基金の額	資本金	capital

以下の三変量モデルの文脈における多変量外れ値検出の研究を行う。想定として、事業従事者数が増えれば増えるほど、事業規模が大きくなり、また、資本金が大きければ大きいほど、事業規模が大きくなると考える。また、大きい事業ほど売上も大きくなるという前提に基づく考え方である。すなわち、$\mathbf{turnover_i = f(worker_i, capital_i)}$ということである。

EDINET におけるこれら 3 つの変数の基本統計量は表 7.2 に示すとおりである。

表 7.2：各変数の基本統計量

変数名	最小値	第 1 四分位	中央値	平均値	第 3 四分位	最大値	標準偏差
turnover	8.1	7383.4	19830.6	106639.8	60008.2	8980555	411520.9
worker	1.0	64.0	147.0	367.5	338.0	22053	937.8
capital	100.0	1088.0	2887.0	17107.0	9062.0	2337895	86344.6

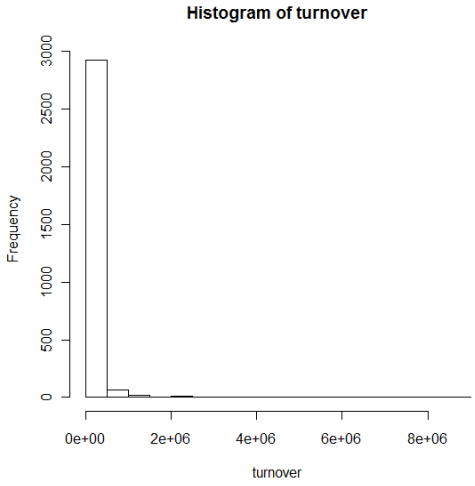
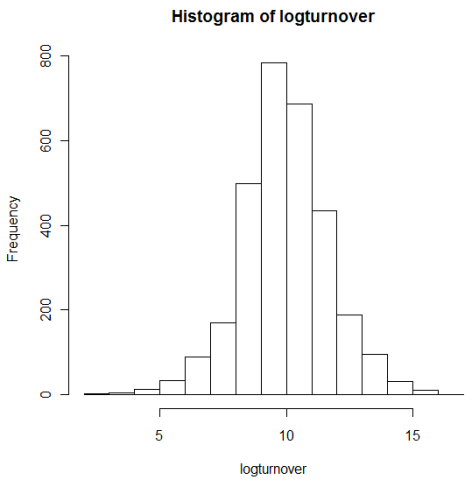
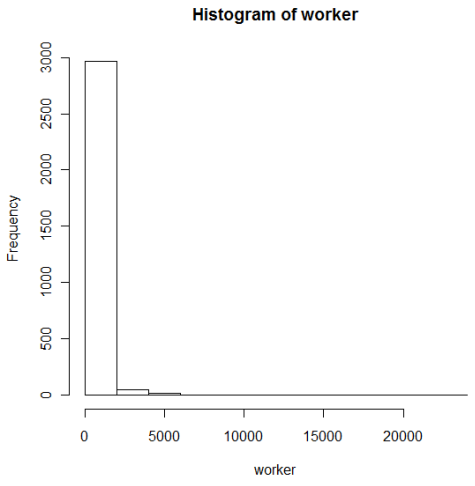
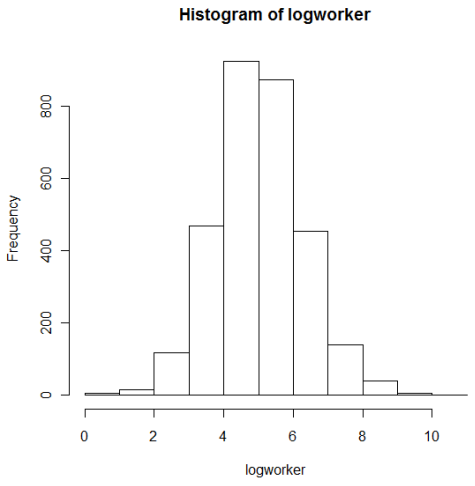
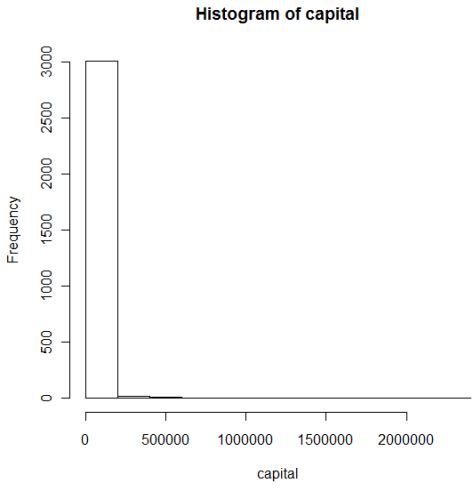
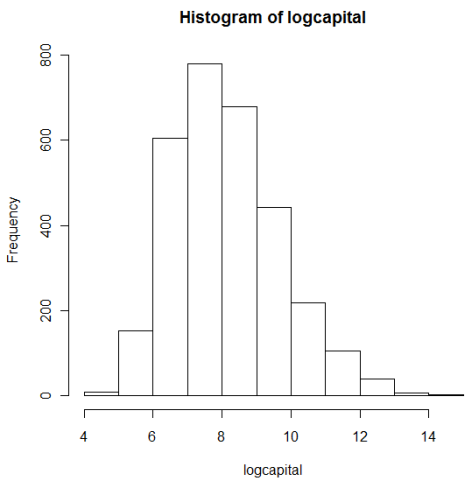
また、各変数間の相関係数は表 7.3 に示すとおりである。

表 7.3：相関係数

	turnover	worker	capital
turnover	1.00		
worker	0.52	1.00	
capital	0.38	0.23	1.00

これらの変数のヒストグラムは図 7.1 のとおりである。売上高、従事者数、資本金のいずれも正規分布ではなく偏りがあるが、いずれも対数変換を行うことで、合理的に正規分布に近くなっていると言える。

図 7.1

元データのヒストグラム	対数変換後のデータのヒストグラム
	
	
	

また、完全な正規分布は、歪度(S: Skewness) = 0、尖度(K: Kurtosis) = 3 となり、歪度と尖度は、それぞれ、式(23)と(24)のとおり求められる(Gujarati, 2003, p.886)。

$$\boldsymbol{S = \frac{E(X-\mu)^3}{\sigma^3}} \tag{23}$$

$$\boldsymbol{K = \frac{E(X-\mu)^4}{[E(X-\mu)^2]^2}} \tag{24}$$

ここで、$\boldsymbol{\mu}$は平均値を表し、$\boldsymbol{\sigma}$は分散を表す。また、$\boldsymbol{E(X-\mu)^3}$は三次積率であり、$\boldsymbol{E(X-\mu)^4}$は四次積率である。

それぞれの変数の歪度と尖度を表 7.4 に示す。生データの歪度と尖度は、0 と 3 からそれぞれ大幅に離れているが、対数変換後の変数は、いずれも 0 と 3 に近い数字となっていることが分かる。したがって、混淆対数正規分布モデルを使用すればよいことが分かる。

表 7.4

変数	歪度 = 0	尖度 = 3
turnover	10.86	166.43
logturnover	-0.08	3.97
worker	11.47	205.24
logworker	0.18	3.56
capital	18.50	438.26
logcapital	0.63	3.36

以下、従事者数及び資本金を条件とした売上高の多変量外れ値検出を行った。図 7.2 は、ECM アルゴリズムが収束するまでにかかった回数を図示している。今回は 214 回の繰り返しの後に収束した。6 節のシミュレーションデータでは、9 回という非常に少ない回数で収束したが、三変量による実データでは、収束までに若干の時間を要したことが分かる。しかし、収束するまでにかかった時間は 1 分 20 秒ほどであり、実用面において大きな障害になるほどではないだろう。

図 7.2：収束するのにかかった回数

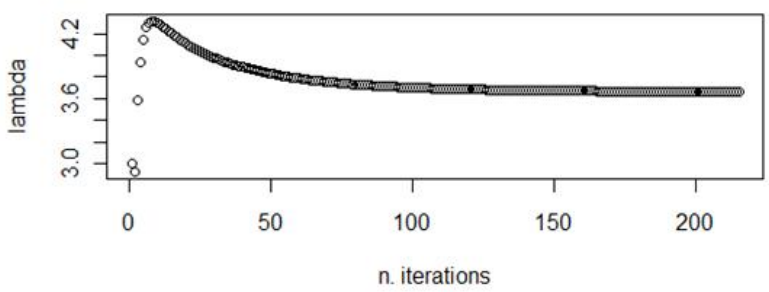


図 7.3 は、売上高を縦軸に、従事者数を横軸にとった散布図である。この散布図上に多変量外れ値を黒丸で示し、正常値は白丸で示している。多変量外れ値は、二変量散布図で示した場合、完全に隠れてしまうものがあることが視覚的に分かる。

図 7.3：売上高と従事者数の散布図（多変量外れ値）

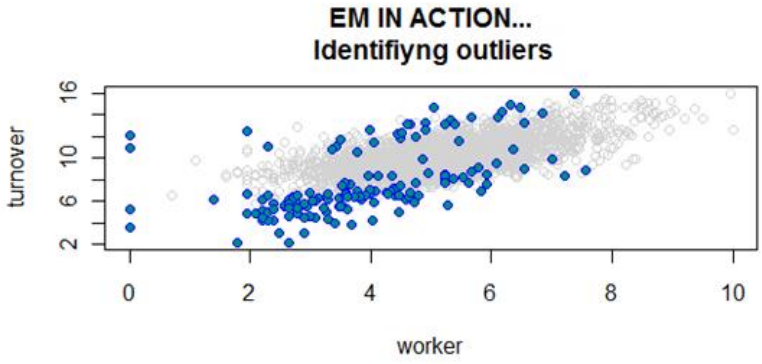
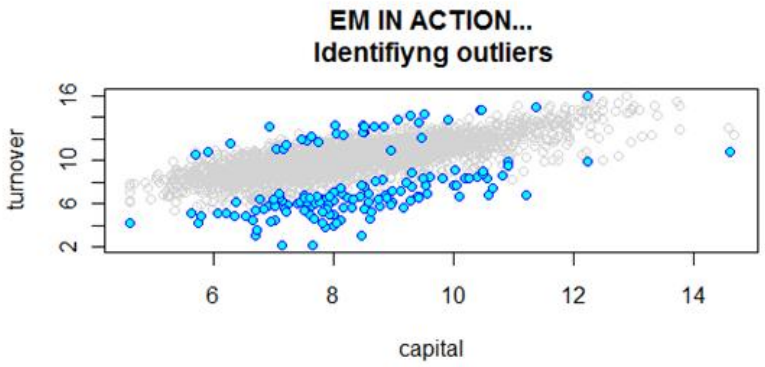


図 7.4 は、売上高を縦軸に、資本金を横軸にとった散布図である。図 7.3 と同様に、この散布図上に多変量外れ値を黒丸で示し、正常値は白丸で示している。この散布図においても、多変量外れ値は、二変量散布図で示した場合、完全に隠れてしまうものがあることが視覚的に分かる。

図 7.4：売上高と資本金の散布図（多変量外れ値）



混淆正規分布モデルの推定値は表 7.5 に示すとおりである。通常 OLS と比較して、混淆正規の BIC 及び AIC の方が小さい数値となっているので、モデルの優位が示されている。

表 7.5：モデルの結果

パラメータ	推定値(混淆正規)	推定値（通常 OLS）
$\hat{\alpha}$	3.006	
$\hat{\beta}_1$	0.586	
$\hat{\beta}_2$	0.444	
sigma	0.863	
lambda	2.660	
w	0.150	
BIC	9378	9512
AIC	4677	4748

以下の図では、通常の値を白丸、「外れ値」を黒丸、「影響力のある外れ値」を菱形で図示する。図 7.5a は売上高における「外れ値」の箱ひげ図であり、図 7.5b は従事者数における「外れ値」の箱ひげ図であり、図 7.5c は資本金における「外れ値」の箱ひげ図である。いずれの図においても、単変量の文脈では、外れ値のほとんどが正常な範囲に収まって隠れており、伝統的な四分位範囲(InterQuartile Range: IQR)の 1.5 倍という単変量外れ値の基準(Weiss, 2005, p.122)では検出できないものが多数あることが分かる。

図 7.5a：売上高の箱ひげ図（黒丸＝外れ値）

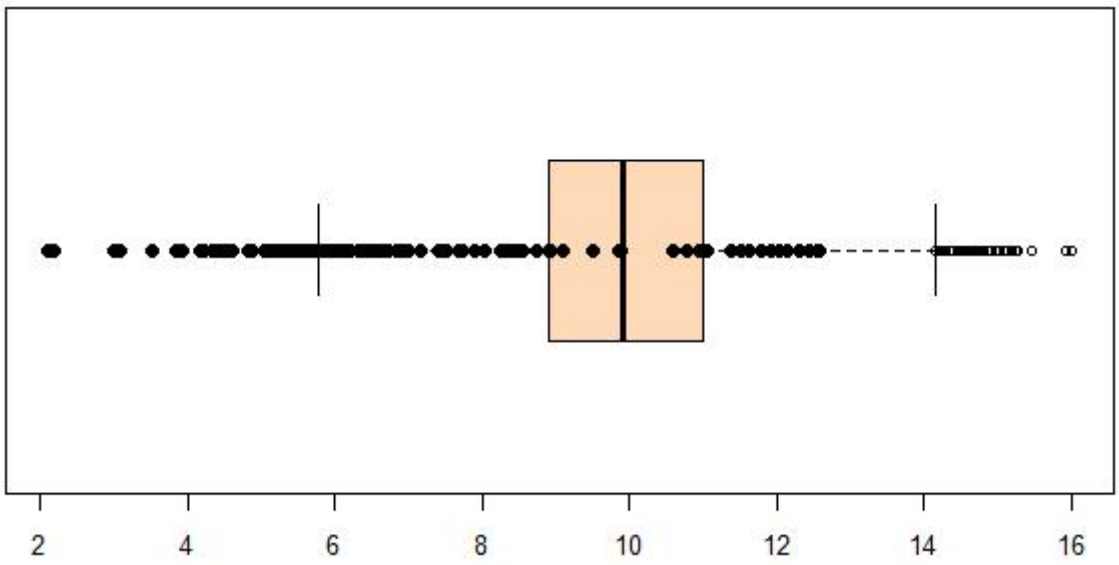


図 7.5b：従事者数の箱ひげ図（黒丸＝外れ値）

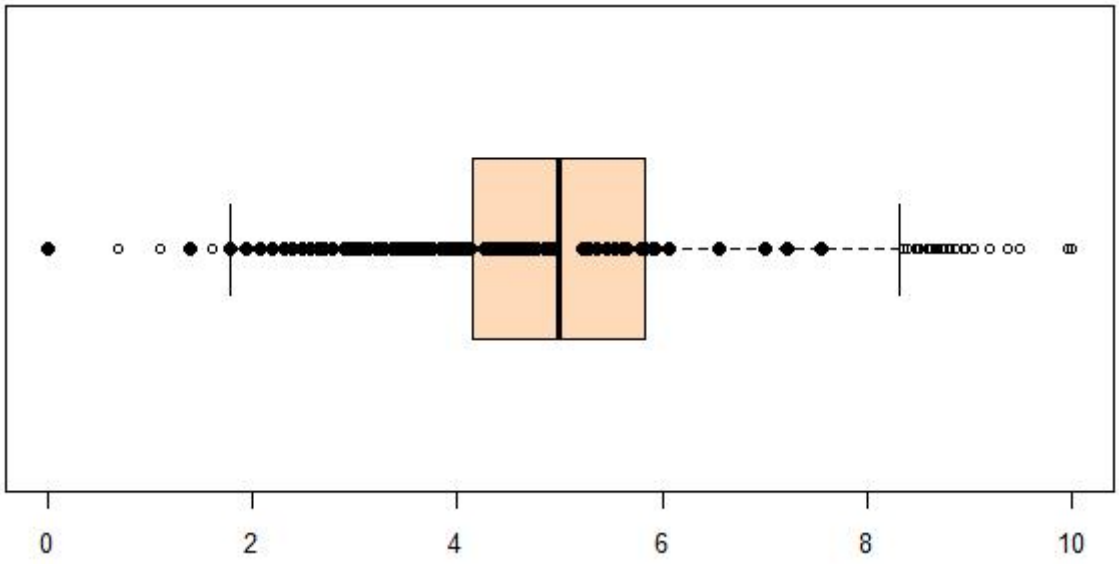


図 7.5c：資本金の箱ひげ図（黒丸＝外れ値）

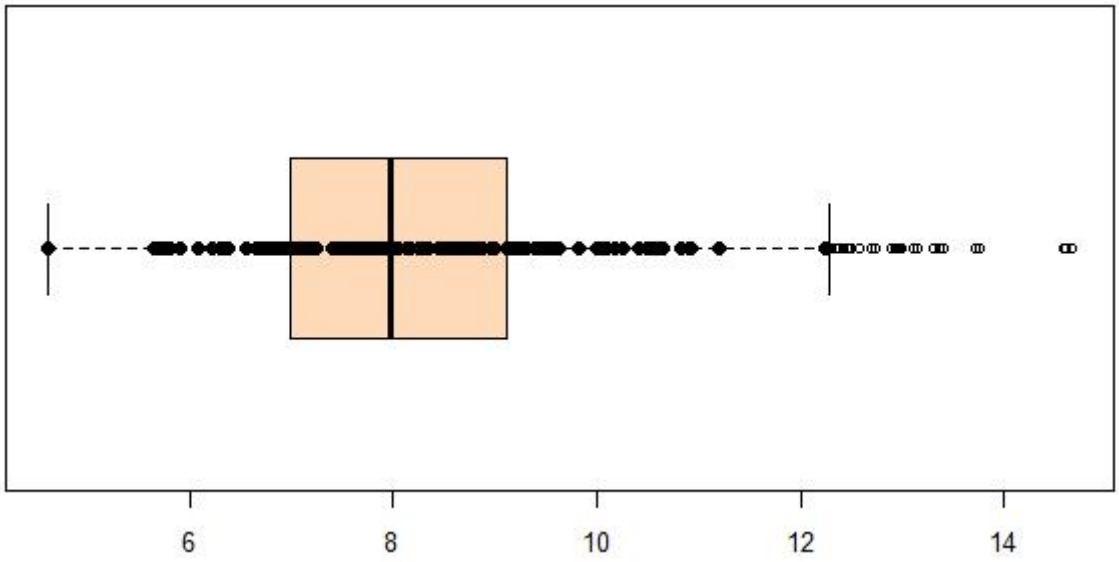


図 7.6a は売上高における「影響力のある外れ値」の箱ひげ図であり、図 7.6b は従事者数における「影響力のある外れ値」の箱ひげ図であり、図 7.6c は資本金における「影響力のある外れ値」の箱ひげ図である。いずれの図においても、単変量の文脈では、影響力のある外れ値のほとんどが正常な範囲に収まって隠れており、伝統的な IQR の 1.5 倍という単変量外れ値の基準では検出できないものが多数あることが分かる。

図 7.6a：売上高の箱ひげ図（菱形＝影響力のある外れ値）

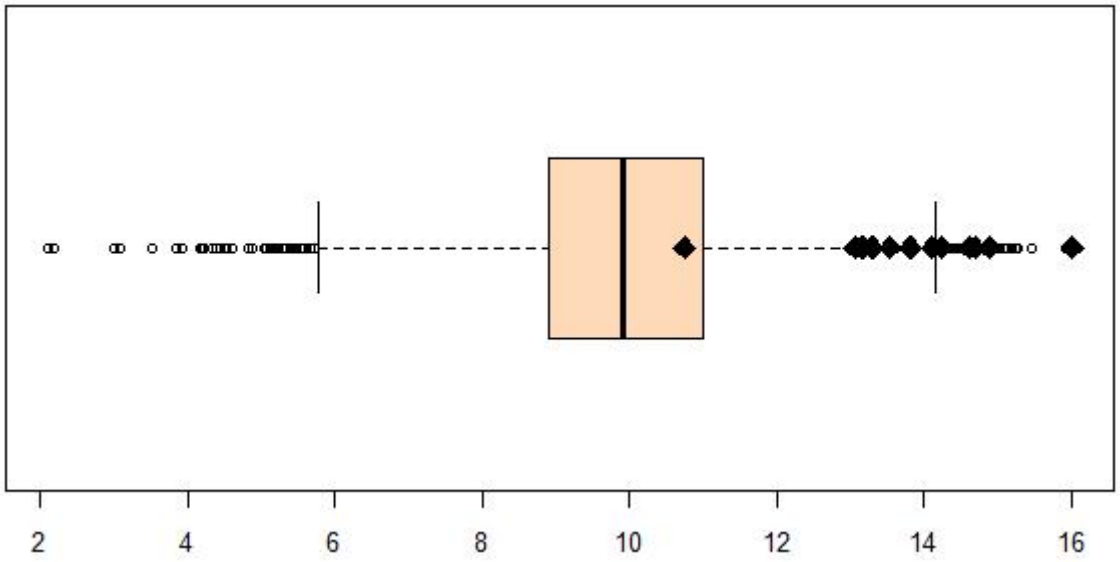


図 7.6b：従事者数の箱ひげ図（菱形＝影響力のある外れ値）

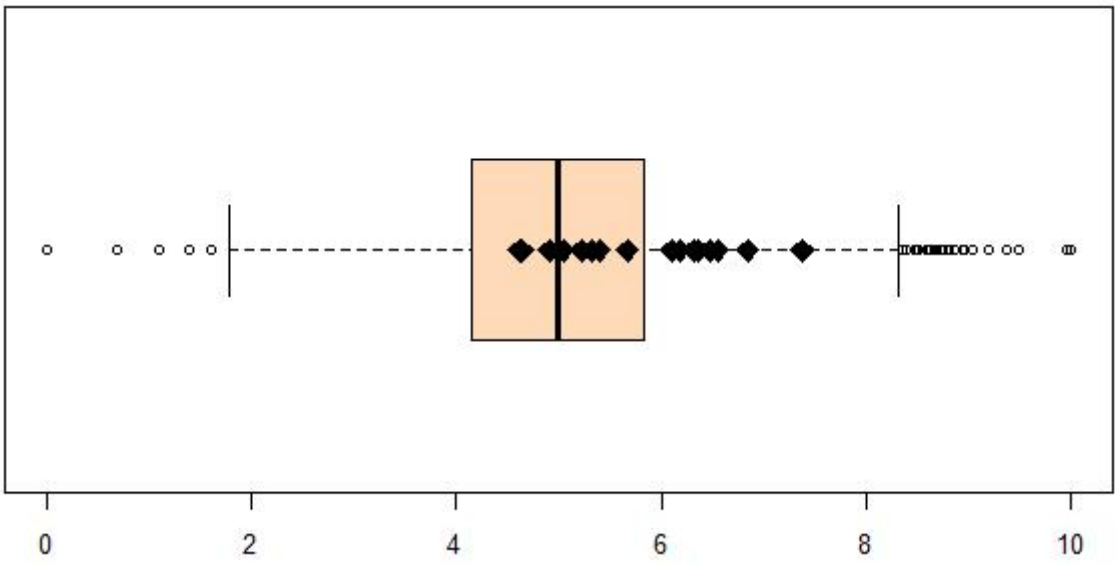


図 7.6c：資本金の箱ひげ図（菱形＝影響力のある外れ値）

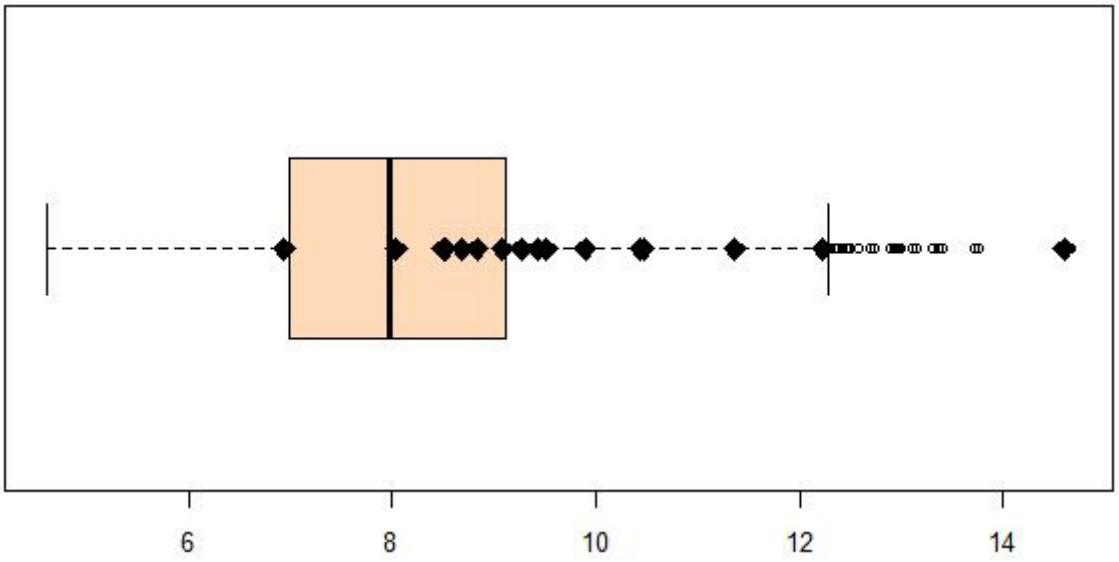


図 7.7a は、売上高（縦軸）と従事者数（横軸）の散布図であり、図 7.7b は、売上高（縦軸）と資本金（横軸）の散布図であり、ここでは、外れ値を黒丸で示している。二変量散布図では、外れ値の多くが中心から逸れてはいるが、中心付近にも多変量外れ値が散見される。

図 7.7a：売上高と従事者数の散布図（黒丸＝外れ値）

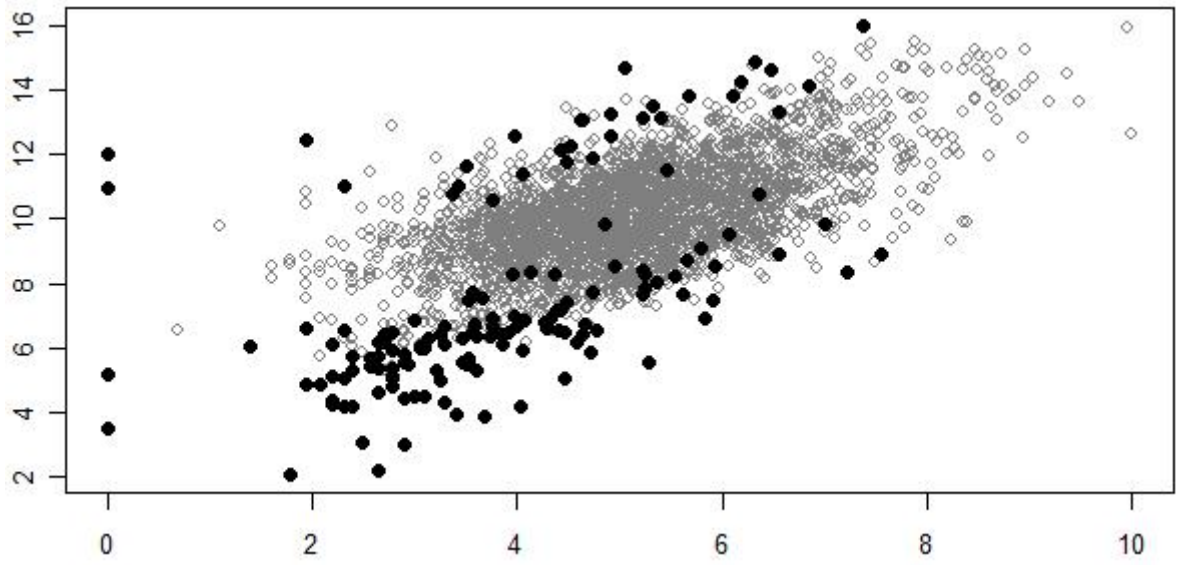


図 7.7b：売上高と資本金の散布図（黒丸＝外れ値）

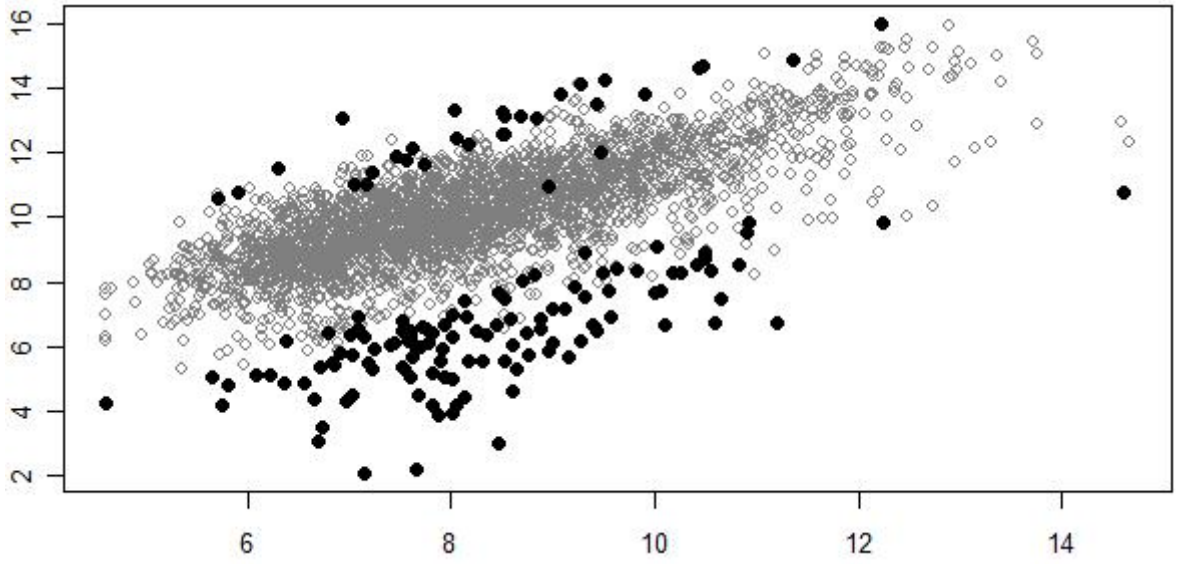


図 7.8a は、売上高（縦軸）と従事者数（横軸）の散布図であり、図 7.8b は、売上高（縦軸）と資本金（横軸）の散布図であり、ここでは、影響力の強い外れ値を菱形で示している。図 7.7 と比較することで、必ずしも外れ値のすべてが影響力ありと判断されている訳ではないことが分かる。

図 7.8a：売上高と従事者数の散布図（菱形＝影響力のある外れ値）

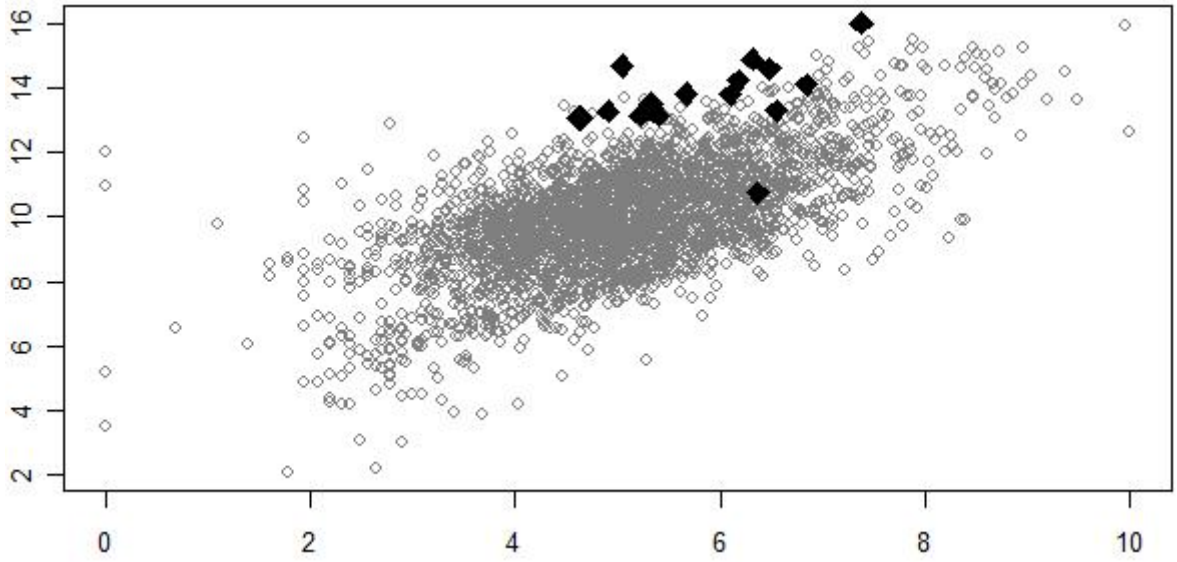
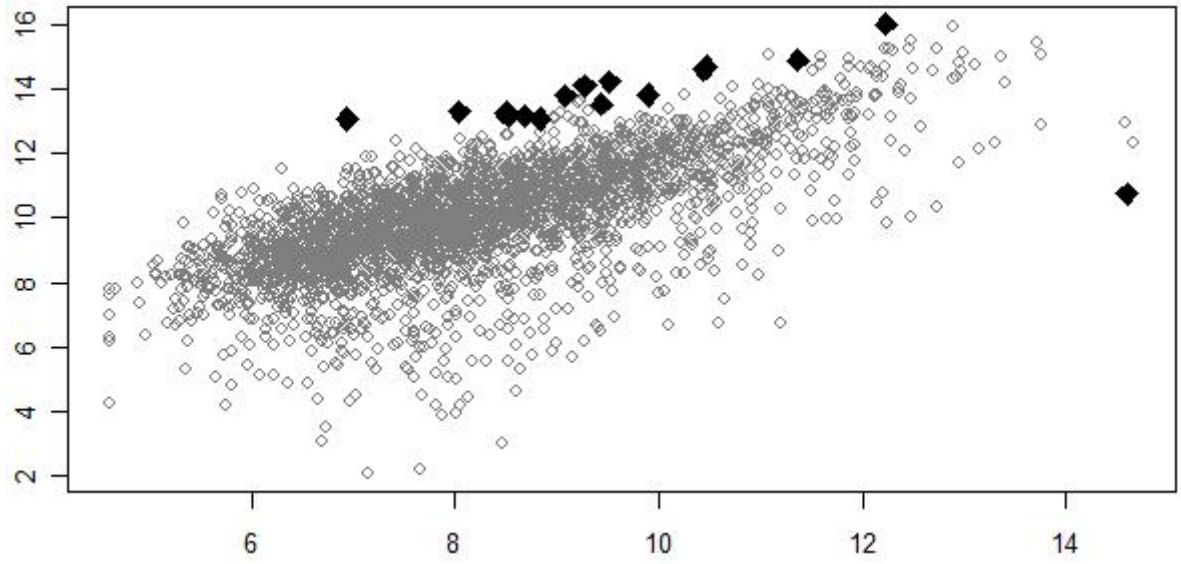


図 7.8b：売上高と資本金の散布図（菱形＝影響力のある外れ値）



今回は、148 個の外れ値が検出されたが、実際に影響力のある外れ値は、表 7.6 に示すとおり、29 個である。

表 7.6：影響力のある外れ値一覧

turnover	turnover.p	turnover.score	global.score	turnover.reserr
8980555.0	4462913.0	0.0149	0.0149	0.0713
2392460.0	416647.1	0.0065	0.0065	0.0563
2915416.0	1678918.0	0.0041	0.0041	0.0498
2203807.0	1047623.0	0.0038	0.0038	0.0457
3392623.0	2240801.0	0.0038	0.0038	0.0419
46422.0	1110127.0	0.0035	0.0035	0.0381
1535183.0	483621.8	0.0035	0.0035	0.0416
3275611.0	2346510.0	0.0031	0.0031	0.0381
2551325.0	1786187.0	0.0025	0.0025	0.0351
1341571.0	619744.1	0.0024	0.0024	0.0325
5240208.0	4526709.0	0.0024	0.0024	0.0302
993850.0	382958.1	0.0020	0.0020	0.0278
4143023.0	3572495.0	0.0019	0.0019	0.0258
4310432.0	3802528.0	0.0017	0.0017	0.0239
1001477.0	536370.4	0.0015	0.0015	0.0222
590274.0	137355.0	0.0015	0.0015	0.0207
472993.6	34361.9	0.0014	0.0014	0.0192
413611.0	845493.7	0.0014	0.0014	0.0177
757557.0	345646.7	0.0014	0.0014	0.0192
3953315.0	3548214.0	0.0013	0.0013	0.0178
1315275.0	977403.3	0.0011	0.0011	0.0165
595217.0	261708.5	0.0011	0.0011	0.0154
514988.0	183454.5	0.0011	0.0011	0.0143
475259.0	170808.3	0.0010	0.0010	0.0132
8242830.0	7945994.0	0.0010	0.0010	0.0122
522515.0	229777.5	0.0010	0.0010	0.0112
2066340.0	1782249.0	0.0009	0.0009	0.0102
124277.0	405252.5	0.0009	0.0009	0.0093
5146318.0	4883897.0	0.0009	0.0009	0.0102

8 将来の可能性と課題

国連欧州経済委員会(UNECE)の統計データエディティングに関するワークセッションにおいて、データエディティング及び選択的エディティングについて活発な議論が行われている。中でも、イタリア国家統計局による混淆正規分布モデルを用いた多変量外れ値検出法は、シミュレーション及び実データのデモンストレーションを通じて、我が国における将来の統計業務への応用可能性が高いことが分かった。さらに、*SeleMix* パッケージは、観測変数と非観測変数の両方の各々のユニットに関して予測値を提供するので、このパッケージは補定の文脈においても有用である。予測値を算出するために使用するモデルは、データ内にエラーが存在していることを考慮しているという点で、この補定は「ロバスト（外れ値などのエラーの影響を受けにくい)」と言える。

正規分布していないデータが懸念材料として考えられるが、EDINET データに見られるとおり、企業データの多くは対数変換を行うことで、対数正規分布となるものが多い。幸いなことに、*SeleMix* において、対数正規分布はすでに対応可能である。正規分布及び対数正規分布以外の分布への対応は、将来の課題と言える。混合モデル自体は 2 つの正規分布の混合を扱うことが最も一般的だが、他の種類の混合を使用することもできる(Everitt, 1996)。したがって、正規分布以外の分布にも理論的には十分に対応可能である。

今回の研究では、推定値の算出アルゴリズムとして EM アルゴリズムの一種である ECM アルゴリズムを紹介した。近年、ECM アルゴリズムをさらに拡張した ECME (Expectation Conditional Maximization Either)アルゴリズムがある。ECME アルゴリズムでは、CM ステップにおいて、不完全データによる尤度の最大化を行うことができ、ECM よりも速く収束することが多い（渡辺, 2008, pp.248-250）。

また、EM アルゴリズムと双璧をなすアルゴリズムとして、ベイズ統計学に基づく MCMC（Markov Chain Monte Carlo：マルコフ連鎖モンテカルロ法）が代表的なものとして挙げられる（金田, 新居, 2009, p.10; 渡辺, 山口, 2000, 第 9 章)。こういった異なるアルゴリズムに基づいた混淆正規分布モデルによる外れ値検出法に関する研究も将来の課題である。

こういった新たなアルゴリズムの研究とともに、様々な種類の実データを使用して、収束までに要する回数や時間の実験を行い、より実務に適したモデルを構築することも将来の課題と言えよう。

今後、イタリアのみならず、諸外国の先行文献を収集・分析することにより、特に企業を対象とした金額データに対する効率的かつ効果的なデータエディティング手法を開発し、製表技術の発展に資することを考えている。

参考文献（英語）

1. Arbués, Ignacio, Pedro Revilla and Soledad Saldaña. (2011). “Selective Editing as a Stochastic Optimization Problem,” *Work Session on Statistical Data Editing, Conference of European Statisticians*, Ljubljana, Slovenia, 9-11 May 2011.
2. Barnett, Vic, and Toby Lewis. (1994). *Outliers in Statistical Data*, Third Edition. Chichester: John Wiley & Sons.
3. Bellisai, Diego, Marco Di Zio, Ugo Guarnera and Orietta Luzi. (2009). “A Selective Editing Approach Based on Contamination Models: An Application to an Istat Business Survey,” *Work Session on Statistical Data Editing, Conference of European Statisticians*, Neuchâtel, Switzerland, 5-7 October 2009.
4. Breiman, Leo, Jerome Friedman, Charles J. Stone, and R.A. Olshen. (1984). *Classification and Regression Trees.* Pacific Grove, CA: Wadsworth.
5. Buglielli, M. Teresa, Marco Di Zio, and Ugo Guarnera. (2010). “Use of Contamination Models for Selective Editing,” *Q2010, European Conference on Quality in Survey Statistics*, Helsinki, 4-6 May 2010.
6. Buglielli, M. Teresa, Marco Di Zio, and Ugo Guarnera. (2011). “Selective Editing of Business Survey Data Based on Contamination Models: an Experimental Application,” *NTTS 2011 New Techniques and Technologies for Statistics*, Bruxelles, 22-24 February 2011.
7. Buglielli, M. Teresa, Marco Di Zio, Ugo Guarnera, and Francesca R. Pogelli. (2011). “An R Package for Selective Editing Based on a Latent Class Model,” *Work Session on Statistical Data Editing, Conference of European Statisticians*, Ljubljana, Slovenia, 9-11 May 2011.
8. DeGroot, Morris H. and Mark J. Schervish. (2002). *Probability and Statistics*. Boston: Addison-Wesley.
9. Di Zio, Marco and Ugo Guarnera. (2006). “A Semiparametric Predictive Mean Matching: An Empirical Evaluation,” *Work Session on Statistical Data Editing, Conference of European Statisticians*, Bonn, Germany, 25-27 September 2006.
10. Di Zio, Marco, Orietta Luzi, and Antonia Manzari. (2002). “Evaluating Editing and Imputation Processes: The Italian Experience,” *Work Session on Statistical Data Editing, Conference of European Statisticians*, Helsinki, Finland, 27 -29 May 2002.
11. Di Zio, Marco, Ugo Guarnera, and Orietta Luzi. (2003). “Using Mixture Modelling to Deal with Unity Measure Error,” *Work Session on Statistical Data Editing, Conference of European Statisticians*, Madrid, Spain, 20-22 October 2003.
12. Di Zio, Marco, Ugo Guarnera, and Orietta Luzi. (2008). “Contamination Models for the Detection of Outliers and Influential Errors in Continuous Multivariate Data,” *Work Session on Statistical Data Editing, Conference of European Statisticians*, Vienna, Austria, 21-23 April 2008.

13. Di Zio, Marco, Ugo Guarnera, Orietta Luzi, and Antonia Manzari. (2005). “Methods and Software for Editing and Imputation: Recent Advancements at Istat,” *Work Session on Statistical Data Editing, Conference of European Statisticians*, Ottawa, Canada, 16-18 May 2005.
14. Everitt, Brian. S. (1996). “An Introduction to Finite Mixture Distributions,” *Statistical Methods in Medical Research* vol.5.
15. Farwell, Keith. (2005). “The General Application of Significance Editing to Economic Collections,” Methodology Advisory Committee, Australian Bureau of Statistics.
16. Ghosh-Dastidar, Bonnie. and J. L. Schafer. (2006). “Outlier Detection and Editing Procedures for Continuous Multivariate Data,” *Journal of Official Statistics* vol.22, no.3: 487-506.
17. Gill, Jeff. (2008). *Bayesian Methods—A Social Sciences Approach*, Second Edition. London: Chapman & Hall/CRC.
18. Granquist, Leopold. (1990). “A Review of Some Macro-Editing Methods for Rationalizing the Editing Process,” *Proceedings of the Statistics Canada Symposium* 225-234.
19. Guarnera, Ugo and M. Teresa Buglielli. (2012). “Selective Editing via Mixture Models,” http://cran.r-project.org/web/packages/SeleMix/SeleMix.pdf.
20. Gujarati, Damodar N. (2003). *Basic Econometrics*, Fourth Edition. New York: McGraw-Hill.
21. Hidiroglou, Michael A. and Jean-Marie Berthelot. (1986). “Statistical Editing and Imputation for Periodic Business Surveys,” *Survey Methodology* vol.12: 73-78.
22. Latouche, Michel and Jean-Marie Berthelot. (1990). “Use of A Score Function for Error Correction in Business Surveys at Statistics Canada,” *Proceedings of the International Conference on Measurement Errors in Surveys*.
23. Latouche, Michel and Jean-Marie Berthelot. (1992). “Use of A Score Function to Prioritize and Limit Recontacts in Editing Business Surveys,” *Journal of Official Statistics vol.8, no.3: 389-400.*
24. Scarrott, Carl. (2007). “Feasibility Study: Selective Editing,” *Official Statistics Research Series* vol.1.
25. United Nations Economic Commission for Europe (UNECE). (2006). *Statistical Data Editing Volume No. 3 Impact on Data Quality*. New York and Geneva: United Nations Publication.
26. Waal, Ton de, Jeroen Pannekoek, and Sander Scholtus. (2011). *Handbook of Statistical Data Editing and Imputation*. Hoboken, NJ: John Wiley & Sons.
27. Weiss, Neil A. (2005). *Introductory Statistics*, Seventh Edition. Boston: Pearson.

参考文献（日本語）

28. 新井貴子，伊藤孝之，阿部穂日，佐々木高浩，袈岩晶，土井満喜. (2010).「国勢調査におけるセレクティブ・エディティングの可能性-その手法と精度に関する基礎的研究」，『製表技術参考資料 15』，独立行政法人統計センター.（非公開）
29. 岡本政人. (2004).「多変量外れ値検出法の研究動向及びカナダ卸売・小売業調査における多変量外れ

値検出法」，『製表技術研究レポート 1』，独立行政法人統計センター.（非公開）

30. 金田尚久，新居玄武. (2009). 「混合分布問題―その基礎からカーネル降下法まで」，『学習院大学経済論集』 vol.46, no.1.
31. 金融庁. (2012). EDINET 概要書. https://info.edinet-fsa.go.jp/download/ESE140001.pdf.
32. 中村永友，小西貞則. (1998). 「情報量基準に基づく多変量正規混合分布モデルのコンポーネント数の推定」，『応用統計学』 vol.27, no.3: 165-180.
33. 畠山昌子. (2008).「『統計データ・エディティング(Vol.3)データ品質への影響』の内容紹介」，平成 20 年度第 1 回統計技術研究会，独立行政法人統計センター.（非公開）
34. 藤原香織. (2009). 「ベイズ検定の漸近理論と特異モデルへの応用」，博士論文，東京工業大学.
35. 堀内泰志. (2006). 「諸外国におけるセレクティブ・エディティングの概要」，『製表技術研究レポート 2』，独立行政法人統計センター研究センター.（非公開）
36. 渡辺美智子，山口和範　編著. (2000). 『EM アルゴリズムと不完全データの諸問題』，東京，多賀出版.
37. 渡辺美智子. (2008). EM アルゴリズム，「21 世紀の統計科学」第 III 巻，日本統計学会 HP 版，第 III 部 統計計算の展開と統計科学.
38. 和田かず美. (2010). 「多変量外れ値の検出～MSD 法とその改良手法について～」，『統計研究彙報』第 67 号 no.4，総務省統計研修所.

製 表 技 術 参 考 資 料 17

平成 24 年 8 月発行

編集・発行　　独立行政法人　統計センター

〒162－8668

東京都新宿区若松町 19－1

電　話　代表　03（ 5273 ）1200

情報技術部統計技術研究課　　TEL ： 03－5273－1368

E-mail ： research@nstac.go.jp