製表技術参考資料 23

ISSN 1348－7329

諸外国における最新のデータエディティング事情

～混淆正規分布モデルによる多変量外れ値検出法の検証～

NSTAC

Working Paper No.23

平成 25 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

諸外国における最新のデータエディティング事情
～混淆正規分布モデルによる多変量外れ値検出法の検証～*

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動向を調査・研究したものである。この目的のために、2012 年 9 月にノルウェーの首都オスロにて開催された国連欧州経済委員会(UNECE: United Nations Economic Commission for Europe)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出席し、報告された全 44 論文を調査した。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、イタリア国家統計局による多変量外れ値検出に関する論文を詳しく検討した。本稿では、この調査の結果をもとに、諸外国における最新のデータエディティング事情及び混淆正規分布モデル(Contaminated Normal Model)[1]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、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。

本稿の構成は以下のとおりである。第 1 節では、2012 年 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の議論をまとめた。第 2 節では、外れ値とエラーの関係、そして選択的エディティングの意義などを議論した。第 3 節では、混淆正規分布モデルによる多変量外れ値検出法に基づく選択的エディティングの理論について概説した。第 4 節では、実データとして、EDINET(**E**lectronic **D**isclosure for **I**nvestors' **NET**work)のデータを用いて、混淆正規分布モデルによる多変量外れ値検出プログラムである *SeleMix* パッケージの検証を行った。第 5 節では、経済センサス - 活動調査への応用を目指し、シミュレーションによる巨大データセット（観測数 100 万）のデータを用い、*SeleMix* パッケージの検証を行った。第 6 節では、結語と将来の課題にて締めくくる。付録として、2012 年 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にて報告された全 44 論文の日本語要旨も掲載した。

* 本稿は、高橋(2012)の続編にあたり、平成 24 年度第 1 回統計技術研究会（平成 25 年 1 月 31 日）及び経済統計学会関東支部 7 月例会（平成 25 年 7 月 6 日）における資料を増補・改訂したものである。イタリア国家統計局の Ugo Guarnera 氏には、*SeleMix* パッケージに関して、情報共有をしていただいた。また、坂下信之課長（統計センター統計技術研究課）、野呂竜夫総括研究員（統計センター統計技術研究課）には、本稿の原稿へのコメントをいただいた。ここに感謝の意を表したい。ただし、本稿にあり得るべき誤りはすべて執筆者に属する。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 統計情報・技術部 統計技術研究課 上級研究員

[1] 高橋(2012)に引き続き、本稿においても、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（渡辺, 山口, 2000, pp.57-58）。

諸外国における最新のデータエディティング事情
～混淆正規分布モデルによる多変量外れ値検出法の検証～

高橋　将宜

序論（研究の目的）

日本の全事業所・企業を対象とし、経理項目を網羅的に調査する経済センサス - 活動調査が、2012 年 2 月に我が国で初めて実施された。経済センサス - 活動調査は、様々な公的経済統計の基礎的資料となるものであり、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[2]がますます重要になってきている。

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。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報告された論文にも注目し、情報を収集してきた。

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ついては、高橋(2012)で取り上げたとおりであるが、2012 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は、統計センターとして参加し、報告及び情報収集を行った。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　2012 年 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、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し、討議を行う国際会議である。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、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通算で 18 回目のセッションが開催され、以下の 7 つの事項が討議された：(1) 選択的及

[2] 日本を含めた各国公的統計機関では、データを単に収集するだけではなく、収集したデータにエラーが含まれていないかを審査し、必要に応じてエラーの訂正を行う。このように、エラーを検出し訂正するプロセスを、統計データエディティング(Statistical Data Editing)または単にデータエディティングと呼ぶ(de Waal et al., p.1)。データエディティングによるエラーの訂正に関する研究は、1950 年代以来行われている(Nordbotten, 1955)。また、データエディティングの発展形としての選択的エディティングに関する研究は、1990 年代初頭以来行われている(Latouche and Berthelot, 1990, 1992)。

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と補定；(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 年ワークセッションの参加者は以下の 28 か国、1 団体であった。欧州からの参加国は 20 か国で、アイスランド、アゼルバイジャン、イタリア、エストニア、オーストリア、オランダ、スイス、スウェーデン、スペイン、スロバキア、スロベニア、デンマーク、ドイツ、ノルウェー、ハンガリー、フィンランド、フランス、リトアニア、ロシア、英国であった。欧州以外からの参加国は 8 か国で、アラブ首長国連邦、オーストラリア、カナダ、ニュージーランド、メキシコ、韓国、日本、米国であった。それ以外に、欧州委員会を代表して、欧州統計局(Eurostat)も参加した。

以下、各トピックで行われた議論について、トピックごとにまとめた。

1.1 選択的及びマクロエディティング

人手によるエディティングは、審査、処理、照会など、非常に時間と費用のかかるものである。マクロエディティング及び選択的エディティングは、出力品質の維持を前提条件としつつ、こういった人手によるエディティングの費用をできる限り取り除くことを目的とする。

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（高橋, 2012, p.5）。マクロエディティングも、選択的エディティングも、どちらも選択手法であり、その共通目的は、潜在的に影響力のあるエラーを持つユニットを人手審査のために選び出すことである。

トピック1では、伝統的な選択手法の応用及び発展や、実務上の問題について討議し、イタリア、スウェーデン、スペイン、ドイツ、米国から報告が行われた。スペインの報告では、選択的エディティングに関する2つの汎用的な原則を提案し、ユニット選択が解決策となるような最適化問題を扱った。イタリアは、混合モデル(Mixture Model)による多変量外れ値検出法を応用した選択的エディティングの報告を行った。米国センサス局は、対外貿易データへのスコア関数の適用可能性の検討を行い、擬似バイアスの評価方法を報告した。スウェーデン統計局は、木解析手法について報告した。ドイツ連邦統計局は、自動比較に関する報告を行った。スウェーデンによる2つ目の報告では、選択的エディティングの汎用ツールを用いた閾値設定に関する実装上の課題を取り上げた。イタリアの2つ目の報告では、推定過程としての選択的エディティングについて報告を行った。詳しい報告内容は、付録1

のトピック(1)に記載している要約のとおりである。

本トピックにおけるディスカッションでは、以下のとおり指摘があった。売上高のような複合変数のエディティングでは、個別の構成要素を考慮することが重要である。その際に、半連続変数への選択的エディティングの適用は容易ではないが、二段階手法が役立つであろう。また、投資といった予測不可能な変数の場合、選択的エディティングはうまく機能しない。すなわち、選択的エディティングを行うには、使用するモデルの当てはまりがよくなければならない。たとえば、選択的エディティングで使用するモデル自体が、データ内のエラーの影響を受ける可能性があるため、ロバスト（頑健）な手法が望まれる。さらに、選択的エディティングを標準化し、汎用化することは極めて難しく、特定のデータや変数ごとに、選択的エディティングを適用しなければならない。

本トピックでは、多数の異なる手法が存在するものの、選択的エディティングは、現在、非常に妥当な手法と結論付けられた。

1.2 エディティングに関するグローバルな解決策

公的統計に割り当てられる予算の削減は、日本だけに限られた話ではなく、万国共通の制約として問題となっている。トピック2では、予算が削減される中、統計作成の効率性を高めるためには、国際協力の構築が欠かせないことを示した。この目的を達成するために考えられる方法は様々あるが、概念や手法は、一般的に、ツールやシステムよりも組織間で共有しやすいため、概念や手法に関して、国際的な標準規格を定めることが重要である。また、公的統計の短期的な目的はデータの提供であるが、長期的な責務として、将来の世代のために、標準的に利用可能な手法によってデータを保存する必要がある。

こういった目的で、統計データエディティングに関するワークセッションでは、『統計データエディティングに関する用語集』(*Glossary of Terms on Statistical Data Editing*)を2000年に刊行した（UNECE, 2000）。この用語集は、幾年にもわたる共同作業の賜物であり、統計データエディティングに関する200以上の用語が収録されている。これらの用語の中には、エディティングにより影響を受けるデータの収集、処理、頒布に関する用語を含んでいる。しかし、2000年に刊行され、10数年の歳月が流れたため、現在の状況に対応するために、新たな概念の追加、既存の概念の修正や削除などに関し、今回のワークセッションの参加者からの意見を募った。本用語集の改訂版は、近日刊行予定であり、その有用性から、統計センターにおいても、その動向を注視している。刊行された暁には、日本語版の刊行をUNECEに打診しているところである。

トピック2では、オーストラリア・UNECE、オランダ・ノルウェー、スウェーデン、ニュージーランド、欧州統計局、カナダから報告が行われた。ニュージーランド統計局は、『統計データエディティングに関する用語集』の改訂に向けた検討に関する報告を行った。オランダ統計局とノルウェー統計局の共同研究では、最新のエディティング理論と実践を

考慮に入れ、一般的なエディティング業務の流れに関する報告を行った。オーストラリアとUNECEの共同研究では、公的統計の近代化を支援するために開発されている汎用統計情報モデルについての報告を行った。スウェーデンは、外部世界についてのデータ及び知識にかかわる公的統計作成の2つのパラダイムについて報告した。欧州統計局の報告では、欧州内での共同システムにおけるデータ妥当性検証手法を改良する提案を行った。カナダは、データエディティング及び補定の文脈において、様々な情報源に対応できる手法に必要な要件を提示した。ニュージーランド統計局は、経済・世帯調査の処理基盤の最新状況について報告を行った。詳しい報告内容は、付録1のトピック(2)に記載している要約のとおりである。

ディスカッションでは、以下のとおり指摘があった。データを実際に使用するユーザーに、統計手法に関する情報を与える必要がある。用語集は、この目的にかなったツールであり、各国の事情を考慮に入れるべきものである。そのため、本ワークセッションの参加者は、用語集の改訂に向けて、フィードバックを提供することに合意した。

1.3 複数情報源と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と補定

複数情報源からデータを統合することによって、統計の作成コストを抑え、調査票の回答者にかかる負担を減らすことができ、さらに質の高い情報を提供できる。そのため、各国の公的統計において、複数の情報源からデータを統合して使用する例が増え始めている。複数の情報源から統合されたデータは、混合モードによるデータ収集の特殊事例と言える。混合モードによるデータ収集とは、対象となる変数の情報を様々な手段によって入手するものである。具体的な手段としては、電子調査票、紙の調査票、企業システムへの直接的なアクセス、行政データ[3]の使用などが挙げられる。こういった状況では、様々な情報収集手段における質の違いというものが根本的な課題となる。すなわち、こういった状況におけるエディティングでは、各々の情報収集手段からデータを入手し、統合情報の一貫性を維持する必要がある。

トピック3では、アラブ首長国連邦、イタリア、オランダ、ニュージーランド、ノルウェー、ハンガリー、フランス、英国から報告が行われた。ノルウェーは、2011年にレジスターベースで行ったセンサスについて報告した。ニュージーランド統計局は、様々な政府機関から提供されるデータを1つの環境に統合するために開発したシステムについての報告を行った。英国国家統計局は、電子調査票におけるエディットを検証するために使用した実験計画について報告した。ハンガリーは、保険統計の分野において、調査データを行政データに置き換えるための方法論について報告を行った。アブダビは、混合モードのデータ収集法及び自動エラーデータ検出法を用いた経済調査について報告した。英国国家統計局の2つ目の報告では、付加価値売上高データと標本調査を混合情報源として利用し、月次

[3] 一般的に、行政データは、複数の情報源から統合されたデータの典型例である。

企業調査の推定値を算出する手法について報告を行った。オランダは、行政データを短期統計として利用する際の問題を解決するための共同プロジェクトについて報告した。イタリアは、ビジネスレジスターを改善する手法についての報告を行った。詳しい報告内容は、付録1のトピック(3)に記載している要約のとおりである。

ディスカッションでは、以下のとおり指摘があった。行政情報ファイルにおいて、どのように外れ値に対処するかは重要なことである。行政情報源を利用することにより、とりわけ精度や時宜性といった品質に関して、どのような影響が出るのかは重要なことであるが、時宜性は、行政データの収集方法に依存するものであろう。また、統計機関が、行政データの正確さをどのようにして決めるかという問題もある。つまり、調査データと行政データとの間に違いがある場合、調査データは、必ずしも行政データよりも正確であるとは限らない。

1.4 メタデータ及びパラデータを使用したエディティングプロセスの効率性分析

OECD (2007)の統計用語集によると、メタデータ(Metadata)とは、「他のデータを定義し記述するためのデータ」である。任意のデータがメタデータであるか否かは、特定の目的を持った特定の状況に依存している。すなわち、いかなるデータも、常にメタデータであるわけではなく、文脈（コンテクスト：あるデータがメタデータとして使用される状況と目的）に依存するのである。メタデータの例としては、「ある論文」（データ）に対する「題名」、「著者名」、「発表年」といった書誌情報（メタデータ）などが考えられる。また、参照メタデータ(Reference Metadata)とは、「統計データの内容や品質を記述するメタデータ」である。理想的には、参照メタデータには、概念的なメタデータ、方法論的なメタデータ、品質に関するメタデータのすべての要素が含まれているべきである。

一方、パラデータ(Paradata)については、現在、国際的に合意された定義は存在しないが、Nicolaas (2011, p.3)によれば、オーディット・トレール（データ処理の内容を追跡調査する記録）やコンタクト・ヒストリー（接触歴）を含む自動生成されたプロセスデータであり、調査データを収集するプロセスについてのあらゆる種類のデータを含んでいる。例としては、調査員の電話記録や面接の長さなどが挙げられる。

統計エディティングプロセスにおける手法や自動アプリケーションの進展に伴い、メタデータ及びパラデータの重要性は、日に日に高まってきている。とりわけ、公的統計のプロセスが適切であるかどうか、また効率的であるかどうかを検証し、改善点を検出するための情報源として使用されている。

トピック4では、カナダ、スウェーデン、フィンランド、フランスから報告が行われた。フランスは、2009年より導入した構造的企業統計を作成する新システムに関して、過去3年の経験に基づくメタデータにより、欠点を補う代替案を検証した。フィンランドによる報告では、統計データエディティングのためのプロセスモデルを紹介した。カナダ統計局

は、企業調査の主要な情報収集法として電子調査票を採用するに際して、7つの調査について実験を行った。スウェーデン統計局では、特定の調査のデータエディティングに関わっている職員を一同に会して経験談を報告しあう質的な調査方法を過去5年間にわたって実施してきた。詳しい報告内容は、付録1のトピック(4)に記載している要約のとおりである。

ディスカッションでは、以下のとおり指摘があった。データエディティングプロセスに関する指標について、国際的に共通な枠組みを構築することは望ましく、汎用統計情報モデルは、この目的で有用である。

エディティング担当職員との意見交換会は、同一調査において複数回行われたこともあり、こういった意見交換会の主な目的は、プロセスを改善するために、専門家による検証や認知テストのための情報を得ることである。

諸外国においても、選択的エディティングの採用に前向きではないケースがあるが、パラデータやメタデータを利用し、最も重要なエラーに焦点を当てることにより、エディティング業務がより意義深いものとなる。また、エディティングの重要な目的は、データを単に訂正するだけではなく、エラーから学ぶことである。

いかにして人々が調査に回答しているかなど、人間的要素というものは、重要である。統計機関は、回答者と直接的に対話をし、回答者のニーズに対応すべきである。こうすることによって、情報源の段階でデータの品質を向上させることができる。

様々なデータユーザーごとに、メタデータやパラデータへの需要は異なっている。内部ユーザーについては、統計作成プロセスの効率性改善が主たる目的であるが、外部ユーザーについては、データ品質についての情報を提供することが目的である。

1.5 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

トピック5の報告の多くは、統計基盤Rにおいて開発されたツールに基づいている。報告論文7本のうち、Rを用いた研究が6本、SASを用いた研究が1本であった。また、Rは、他の基盤やデータベースとの対話の仲介役ともなり得る。本トピックでは、オーストリア、オランダ、リトアニア、日本、米国から報告が行われた。

米国センサス局は、生データからエディティング、補定、出力結果の公表まで、人口調査の処理を統一化するための汎用システムについての報告を行った。オランダは、伝統的な視覚化ツールの限界を克服する目的で開発された2つの視覚化ツールについて報告した。オーストリアは、時系列データ分析における季節調整ソフトウェアの煩雑さを克服できるRパッケージについての報告を行った。国連工業開発機関(UNIDO: United Nations Industrial Development Organization)は、エラーデータや不完全なデータを選別して抽出する手法の報告を行った。オランダ統計局では、自動的にデータエディティングを行う手法をRパッケージとして作成した。リトアニア統計局では、エディティング及び補定のためのSASのマクロプログラムを開発した。日本は、経済センサスのデータエディティングへの

適用を目指して、売上高の多重代入法に関する研究を報告した。詳しい報告内容は、付録1のトピック(5)に記載している要約のとおりである。

ディスカッションでは、以下のとおり指摘があった。ソフトウェアの試作段階では、現実的な目標を持ち、明確なスコープを持つことが重要であり、ソフトウェアを用いた統計の作成段階では、ユーザーからの信頼性を高めるために、メンテナンスを定期的に行い、関連した研究活動も活発に行う必要がある。

使用する統計ソフトウェアの選択に関して、人的要因は重要である。統計家の多くはSASに馴染みがあるが、年々、学生の間では、Rの利用者数が増えている。現実的な解決策としては、少なくとも短期的には、RとSASの両方を並行して使用することであろう。Rとは、そもそも、統計環境であり、他の言語で開発されたパッケージと共にRを使用することが可能である。Rの利点としては、現在利用可能な統計パッケージが非常に多くあり、それらすべてが無料である点が挙げられる。

1.6 新たな手法

トピック6では、データエディティング及び補定を改善し、最適化するための手法や技術に関する最新のアイディアや発展について報告が行われた。本トピックでは、オランダ、スウェーデン・エストニア、スロベニア、欧州統計局から報告が行われた。スウェーデンとエストニアの共同研究では、確率に基づくエディティング手法を研究した。欧州統計局は、カテゴリカルデータの補定に関して、機械学習法の分野で開発された手法を検証した。スロベニア統計局は、ベイズ手法に基づく補定の実装について報告した。オランダ統計局では、ソフトエディットを考慮した新しい自動エディティング手法を開発した。詳しい報告内容は、付録1のトピック(6)に記載している要約のとおりである。

ディスカッションでは、以下のとおり指摘があった。補定手法の精度をどのように評価するかは議論の余地があるが、健全な統計的手法は、点推定値だけではなく、標準誤差のような不確実性に関する指標も示しているべきである。変数が多かったり、変数の種類が異なっていたり、補定が複雑になる場合、市販のソフトウェアの既定の設定では不十分だと考えられるが、使用している手法が適切であるかどうかを検証するために、どのような診断（図や数値分析）がなされるべきかについては、将来の課題である。

エディティングに関して、確率的エディティングは、二段階標本抽出法に酷似していると指摘された。また、選択的エディティングは、カテゴリカルデータに対して適切なのかどうか、将来の課題として議論をする必要がある。

1.7 センサスデータのエディティング及び補定

多くの統計機関では、「伝統的な」センサスデータ収集法を徐々に行政データやレジス

ターデータなどに置き換え始めており、結果として、エディティング及び補定の戦略に大きな影響が出ている。したがって、トピック7では、センサスデータに適用するエディティング及び補定技術の方法論的革新を扱う。本トピックでは、アラブ首長国連邦、オーストリア、スロベニア、メキシコ、英国から報告が行われた。スロベニアは、2010年のスロベニア農業センサスにおける複数情報源データのエディティングに関する報告を行った。オーストリアは、レジスターベースのセンサスに移行した2011年のセンサスにおける補定プロセスに関する報告を行った。英国は、2011年センサスで実装した自動作成環境におけるエディット及び補定の長所と短所について概観した。また英国の2つ目の報告では、2011年センサスにおけるエディット及び補定戦略の目標を達成するためのツールの検証を行った。アブダビは、公的統計の近代化のプロセスとして、人口センサスのエディティング及び補定に関する研究を行った。メキシコは、地理情報システムを用いたセンサスデータのエディティングに関して報告を行った。詳しい報告内容は、付録1のトピック(7)に記載している要約のとおりである。

ディスカッションでは、センサスデータに関して、以下のとおり指摘があった。センサスデータの使用目的は様々であり、標本抽出フレームや研究目的でのミクロデータの再利用などが例として挙げられるが、これらにおいて、選択的エディティングは、最適な形で適用できない恐れがある。こういった場合、自動エディティングが代替法として考えられる。しかし、個別の対応を必要とする特殊な部分母集団が常に存在しているため、センサスのデータエディティングシステムを完全に自動化することは、実行可能ではないだろう。エディティング及び補定を完全に自動化することによって、データセットにすでに存在しているノイズを再生産する恐れがある。しかし、スコア関数やマクロエディティングを用いることにより、自動エディティングにおけるエラーを見つけることができる。したがって、エディティング及び補定を自動で行った後、ノイズに対処すればよいであろう。

また、レジスターデータに関して、以下のとおり指摘があった。特定の地域を過小あるいは過大にカバーしているとすれば、政治上の問題となり得るので、レジスターデータのカバーしている範囲を確認する必要がある。また、レジスターデータは、通常、時系列的に変化せず、安定しているので、センサスデータの情報源として適切かどうか考える必要がある。レジスターを管理している当局との緊密な協力関係により、こういったリスクを軽減できる。したがって、レジスターデータを使用する際には、エディティングや補定を綿密に行う必要がある。

1.8 次回のワークセッション

ドイツ、ハンガリー、フランス、欧州統計局の作業グループによる提案をもとに、参加者間で次回ワークセッションの議題などを検討した。その結果、INSEE（フランス国立統計経済研究所）の提案により、次回の統計データエディティングに関するワークセッショ

ンは、2014年春にフランスのパリで開催される予定となった。ただし、次回ワークセッション開催の正式決定には、欧州統計家会議(Conference of European Statisticians)の承認を必要とする。次回のワークセッションで討議される予定の事項は、以下のとおりである。掲載されている国名及び団体名は、2012年9月時点において、これらの議題に参加の意欲を表明したものである。

1. 選択的エディティング/マクロエディティング(Selective Editing/Macro Editing)
   - オランダ、カナダ、スウェーデン、スロベニア、ニュージーランド、英国、米国
2. エディティング手法全般－ビジネスレジスターに関するインフラ及び複数情報源の文脈において(Other Methods of Editing – Business Registers Infrastructures and in the Context of Multiple Sources)
   - カナダ、ドイツ、ノルウェー、フランス
3. データエディティングの実施と関係者の協力(Getting the Support of All People When Implementing Data Editing)
   - カナダ、ノルウェー、フィンランド、フランス
4. 新たな手法(New and Emerging Methods)
   - オランダ、スウェーデン、フランス、メキシコ、英国、米国
5. センサスデータ及び社会データのエディティング(Editing of Census and Social Data)
   - カナダ、スイス、ニュージーランド、フランス、英国
6. 国際協力及びソフトウェアとツール(International Collaboration and Software & Tools)
   - UNECE、オランダ、カナダ、スロベニア、ニュージーランド、欧州統計局

2 選択的エディティング：外れ値とエラー

前節では、データエディティングに関する最新の国際的な動向を概観した。本稿は、データエディティングの中でも特に選択的エディティングに焦点を当てている。よって、本節では、外れ値の検出法を検討し、エラーへの対処法を概観する。

2.1 エラーと外れ値

統計データエディティングの目的は、データ内のエラーを検出し、訂正することにあるが、エラーには概ね、体系的エラー(Systematic Error)とランダムエラー(Random Error)の 2 種類がある（Nordbotten, 1955, p.364; Trochim, 2006; de Waal et al., 2011, p.7）。

体系的エラーとは、複数の回答者（回答ユニット）に共通して頻繁に起こるエラーであり、調査票の誤読や誤解に起因していることが多い。こういったエラーとしては、測定単

位エラー(Unity Measure Error)が最もありふれたものであり、たとえば、売上高を 100 万円単位で報告するべきところを 1 円単位で報告してしまうといったケースである。体系的エラーは、文字どおり体系的に発生するために、平均値の上または下に偏りやすく、平均値に大きな影響を与える（図 2.1 参照）。したがって、バイアス（偏り）を生む重大なエラーとなるが、エラーの発生メカニズムを特定しやすいため、比較的に対処しやすい。

一方、ランダムエラーとは、文字どおり、偶発的な原因で発生するエラーである。具体的な例としては、「10000 円」と入力しようとして、「1000 円」と 0 を少なく入力してしまったり、「100000 円」と 0 を多く入力してしまったりするケースを挙げられる。また、「1234 円」と入力しようとして、「1324 円」と順番を打ち間違えて入力してしまったり、「1204 円」と見間違えて入力してしまったりするなどの例も挙げられる。多くの場合、ランダムエラーの発生原因は不明であることが多い。しかし、ランダムエラーは、分布から外れた値になることがあり、こういった場合には、外れ値検出法を用いることにより、ランダムエラーを検出することができる。ランダムエラーは、偶発的に発生するため、平均値よりも大きなエラーと小さなエラーとが相殺しあうため、平均値への影響は少ないが、分散への影響が大きい（図 2.2 参照）。

図 2.1：体系的エラー

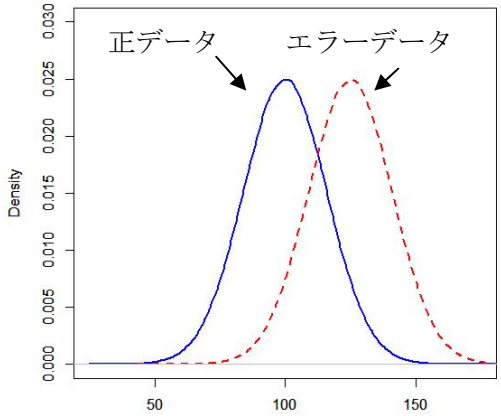
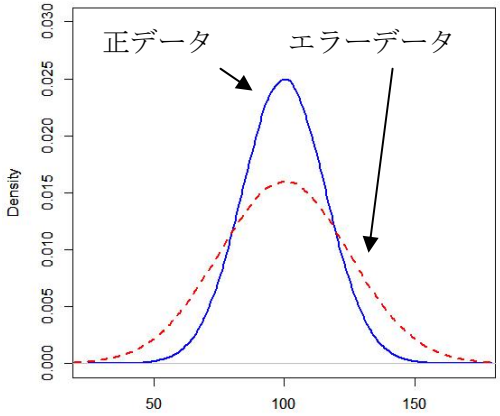


図 2.2：ランダムエラー



注：青線はエラーのない正データの分布、赤点線はエラーを含むデータの分布をそれぞれ表している。

また、一般的に、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と理解されている(Weiss, 2005, p.122)。データセット内の任意の値が単なる外れ値なのか、それともエラーなのか、こういった判断には専門知識や背景知識が不可欠となるが、数理統計的に外れ値を検出することは常に可能である（そもそも外れ値の存在しないデータセットを除く）。たとえば、平均 172、標準偏差 5.5 で正規分布しているデータの中に、220 と 330 という値が存在しているとしよう。220 の z 値は(220-172)/5.5 = 8.727 であり、330 の z 値は(330-172)/5.5 = 28.727 である。一般的に、正規分布では、z 値が 2 を超えると出現率は

2.275%を下回り、z 値が 3 を超えると出現率は 0.135%を下回り、z 値が 4 を超えると出現率は 0.003%を下回る。つまり、z 値 8.727 や z 値 28.727 は、極めて出現率の低い値であり、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ることが分かる。つまり、背景知識がなくとも、これらの値を数理統計的に外れ値として検出できる。しかし、これらの値がエラーであるかどうかは、このデータの背景的な知識を必要とする。

実際には、「平均 172、標準偏差 5.5 で正規分布しているデータ」とは、日本人男性の平均身長とその標準偏差である。この背景から考えれば、身長が 330 センチの人類は存在しないと合理的に考えられるため、330 という値は、明らかにエラーであることが分かる。一方、身長 220 センチの日本人男性は、ほとんど存在しないものの、あり得ない数字ではないため、エラーである可能性は高そうだが精査をする必要のある値だと分かる。

表 2.1 に示すとおり、すべての外れ値がエラーであるわけではなく、すべてのエラーが外れ値であるとも限らない。すべてのデータの中で、外れ値として検出できたものの中には、正データとエラーデータが混在している（①と②）。これらを確認し、正データと確認できたものについてはそのままにし（①のケース）、エラーデータと確認できたものについては訂正を行う（②のケース）。外れ値として検出できなかったものの中で正データであったものは、もともと訂正を行う必要がなかったので問題はない（③のケース）。しかし、エラーの中には外れ値として検出できないものもあり、こういった種類のエラー（多くは上述の体系的エラー）には、外れ値検出法以外の対処法を適用する必要がある（④のケース）。

表 2.1

	正データ	エラーデータ
外れ値	検出可（要確認） ①	検出可（要確認） ②
非外れ値	検出不可 ③	検出不可 ④

したがって、外れ値を検出するということは、その値を即座にエラーとして取り除いたり、他の値に置き換えたりするために行うのではなく、エラーである可能性の高そうな値を探し出し、その値を重点的に入念に精査するために行うのである。また、外れ値検出法のみが、エラーの唯一の検出法であるわけではなく、エラーを効率よく処理する方法の 1 つであることは明記しておきたい。

2.2 影響力

効率的で効果的なデータエディティングを達成するためには、影響力の大きなエラーから対処することが賢明である。しかし、すべての外れ値が、データ全体に大きな影響を与えるわけではないため、外れ値検出法によりエラーの検出を行う際には、影響力の大きな外れ値を重点的に検出し、検討を行う必要がある。

そこで、「影響力とは何か？」ということが問題となる。Fox (1991, pp.21-22)によれば、影響力(Influence)とは、てこ比(Leverage)と乖離(Discrepancy)によって構成される[4]。

影響力 = てこ比 × 乖離

ここで、てこ比とは、「***x***（説明変数）の相関構造を考慮し、***x***の重心(Centroid)からの距離」(Fox, 1991, p.25)と定義でき、乖離とは、データの全体的なパターンから大きく逸脱した特異性と定義できよう。

具体的に例示するために、表 2.2 のデータセットを使用する。主に、ユニット 5 の値に注目する。***y1***と***x1***のペアには外れ値は存在していない。***y2***と***x2***のペアには乖離した値が含まれている。具体的には、***x2***のユニット 5 の値は平均値だが、***y2***のユニット 5 の値は平均値から大きく乖離している。***y3***と***x3***のペアには乖離しており、また、てこ比の高い値が存在している。具体的には、***y3*** と***x3*** の両方において、ユニット 5 の値は平均値から大きく乖離している。***y4***と***x4***のペアには乖離した値が含まれている。具体的には、***y4***のユニット 5 の値は平均値だが、***x4***のユニット 5 の値は平均値から大きく乖離している。仮に、***y1***と***x1***のユニット 5 の値が正しく、***y2***と***x2***、***y3***と***x3***、***y4***と***x4***のユニット 5 の値はエラーであるとしよう。一般的な初級統計学の教科書(Weiss, 2005, p.122)では、IQR (Inter-Quartile Range：四分位範囲)の 1.5 倍を超える値を単変量外れ値としている：上限 = Q3 + 1.5*IQR；下限 = Q1 – 1.5*IQR。したがって、表 2.2 では、IQR の 1.5 倍を超えており、単変量外れ値として認識されるものを便宜的に赤丸で囲んだ。

表 2.2

ユニット	***y1***	***x1***	***y2***	***x2***	***y3***	***x3***	***y4***	***x4***
1	10	10	10	10	10	10	10	10
2	20	20	20	20	20	20	20	20
3	30	30	30	30	30	30	30	30
4	40	40	40	40	40	40	40	40
5	50	50	100	25	240	80	25	100
第 1 四分位	20	20	20	20	20	20	20	20
第 3 四分位	40	40	40	30	40	40	30	40
上限	70	70	70	45	70	70	45	70

表 2.3 は、表 2.2 のデータを用いた回帰分析の結果である。モデル 1 は***y1***と***x1***の回帰分析の結果であり、これを真のモデルとする。真の切片は 0.000 であり、真の傾きは 1.000 であり、真の R^2 は 1.000 である。モデル 2 は***y2***と***x2***の回帰分析の結果であり、モデル 3

[4] 影響力については、Rousseeuw and Leroy (2003, p.13)及び Andersen (2008, pp.8-9)も参照されたい。

は***y3***と***x3***の回帰分析の結果であり、モデル 4 は***y4***と***x4***の回帰分析の結果である。

表 2.3

	モデル 1	モデル 2	モデル 3	モデル 4
切片	0.000	15.000	-54.800	21.000
傾き	1.000	1.000	3.411	0.100
R^2	1.000	0.100	0.906	0.100
n	4	4	4	4

科学的分析においては、通常、傾きの値によって、説明変数と被説明変数の関係性を捉えるため、傾きの値は非常に重要である[5]。モデル 2 では、切片の値が 15.000 となっており、真の値から大きくずれ込んでいるが、傾きは 1.000 であり、真の値と同一になっている。したがって、***x***の値が平均値付近にある場合、てこ比が低いため、***y***の値が乖離していたとしても、影響力が小さいことが分かる。モデル 3 では、切片の値が-54.800 と大幅にずれているだけではなく、傾きも 3.411 と過大推定になっている。***x***の値が平均値から遠いために、てこ比が高く、***y***の値も乖離しているため、影響力が大きいことが分かる。モデル 4 では、切片の値が 21.000 と大幅にずれているだけではなく、傾きも 0.100 と過小推定になっている。***y***の値は平均値から乖離していないが、***x***の値が平均値から遠いために、てこ比が高く、影響力が大きいことが分かる。

これを視覚的に図示したものが、図 2.3 から図 2.6 である。図 2.3 は、真のモデルの散布図を表している。図 2.4 では、ユニット 5 の値が、他の観測値から乖離していることが見て取れるが、***x***の平均値付近に存在しており、てこ比が小さいために傾きへの影響がない。結果として、回帰線を一律に上向きに持ち上げただけで、影響が少なかった。図 2.5 では、ユニット 5 の値が、他の観測値から乖離しているだけではなく、てこ比も大きく、回帰線に大きな影響が出ている。図 2.6 では、ユニット 5 の値は***y***の平均値付近にあるものの、***x***の値が異常に大きく、てこ比が高いため、非常に影響力が大きくなっており、回帰線の傾きが大幅に過小推定されている。

[5] 切片の値は、***x***の値がゼロであった場合に、***y***の値がいくつになるかを示しているだけであり、説明変数と被説明変数の関係性を必ずしも表していないため、科学的分析においては、重要視されないことが多い。

図 2.3：モデル 1 の散布図

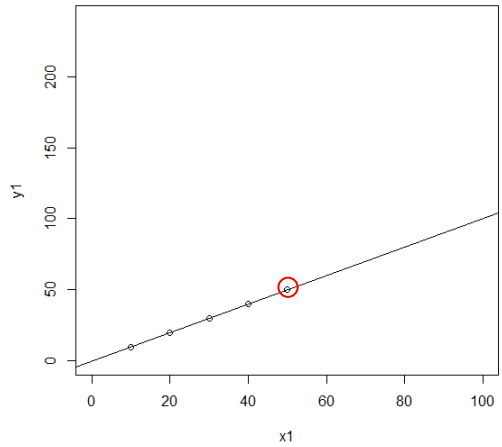


図 2.4：モデル 2 の散布図

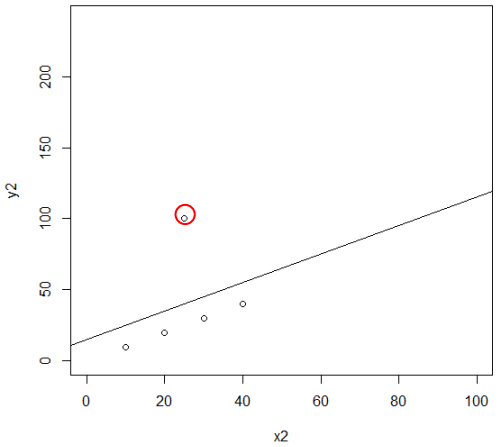


図 2.5：モデル 3 の散布図

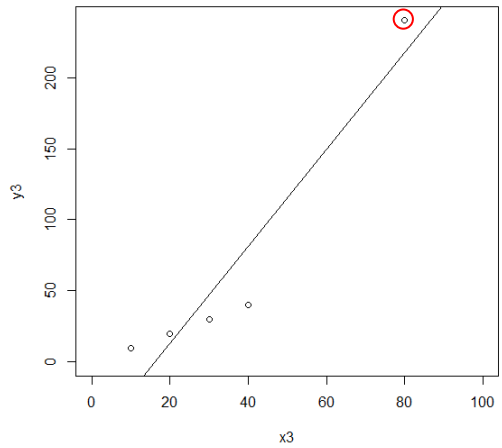
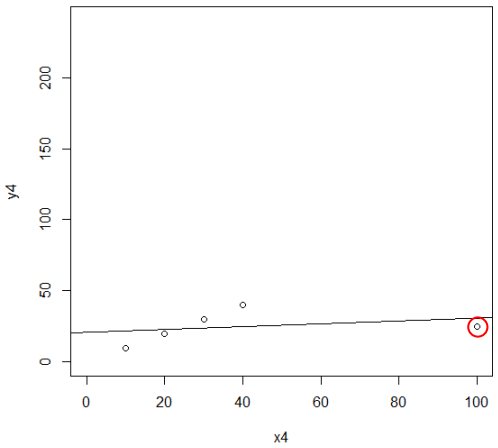


図 2.6：モデル 4 の散布図



結論として、***y2***のユニット 5 の値は、外れ値ではあるが、影響力のない外れ値であった。一方、***y3***と***x3***のユニット 5 の値は、外れ値であり、かつ、影響力があった。また、***x4***のユニット 5 の値も、外れ値であり、かつ、影響力があった。このように、一概に外れ値と言っても、影響力の大きなものと小さなものとがあるため、データエディティングにおいては、影響力の大きな外れ値を重点的に精査することが重要である。

2.3 選択的エディティングを行う意義

図 2.7 と図 2.8 は、de Waal et al. (2011, pp.191-192)にて報告されている実際の統計調査の結果を模した図である。この調査では、350 個のユニットの訂正が行われた。図 2.7 と

図 2.8 では、訂正前の生データの完成度[6]を 0%とし、理想的に完璧な真のデータセットの完成度を 100%とし、350 個を訂正した場合の完成度を 90%であったとしている。縦軸に完成度（%）を、横軸に訂正を行ったユニットの数を図示した。図 2.7 は、選択的エディティングによりエラーの訂正を行った場合の模式図である。すなわち、全ユニットの影響力を事前に算出し、影響力の強いユニットから順番に訂正を行っていった図である。図 2.8 は、人手エディティングによりエラーの訂正を行った場合の模式図である。この図では、エラーの訂正を無作為な順番で行っている。

図 2.7：選択的エディティング

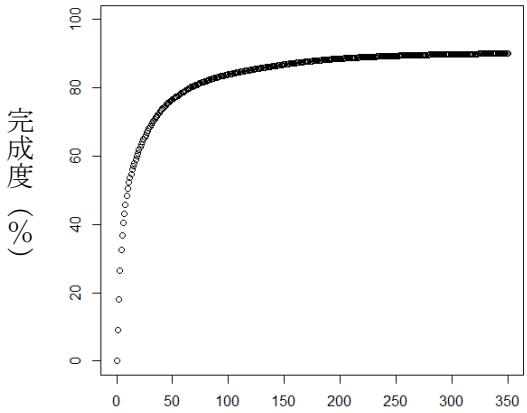


図 2.8：人手によるエディティング

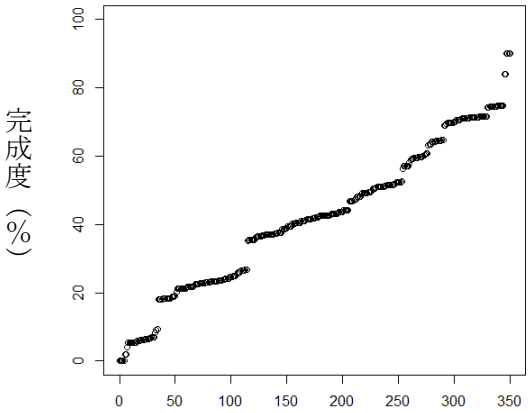


図 2.7 と図 2.8 の具体的な数値を表 2.4 に示す。選択的エディティングでは、最も影響力のある最初の 1 個のエラーを訂正することにより完成度を 9%まで上げ、最も影響力のある最初の 10 個のエラーを訂正することで完成度を 50%に高めることができ、最も影響力のある最初の 100 個の訂正により 84%まで完成度を高められる。一方、人手によるエディティングでは、最初の 1 個のエラーの訂正によって完成度は 0.03%にしかならず、最初の 10 個のエラーを訂正しても完成度は 5%であり、最初の 100 個を訂正しても完成度はたったの 25%である[7]。

[6] ここで「完成度」とは、以下のことを意味している。エディット前の生データには、様々なエラーが含まれており、エラーの氾濫している状態の生データは完成していないという意味で、完成度を 0%としている。一方、神のみぞ知る真のデータセットには、エラーが 1 つも存在しないため、その完成度を 100%としている。

[7] 人手による訂正は、無作為であり、ここでは乱数に基づいている。したがって、具体的な数字は、シミュレーションごとに異なったものになる。

表 2.4

訂正	0	1	10	20	50	100	150	200	250	300	350
選択	0%	9.06%	50.50%	61.72%	76.55%	83.91%	86.83%	88.55%	89.38%	89.83%	89.99%
人手	0%	0.03%	5.37%	6.13%	19.24%	24.68%	38.92%	43.68%	52.42%	69.95%	89.99%

どこまでを訂正すればよいかという問題は、主観的な問題となり、筆者から決定的かつ普遍的な指針を示すことはできない。もし時間と予算が無制限に利用可能であるならば、できる限り多くのエラーを訂正するべきであろう。事実、人手によるエディティングでは、350 個の訂正を行わなければ、90%の完成度を維持できない。しかし、現実には時間や予算の制約が存在している。表 2.4 の結果から考えたとき、選択的エディティングでは、200 個の訂正により 89%の完成度を達成することができている。350 個の訂正により 90%の完成度を達成することには、1.75 倍もの時間と予算を割いただけの成果の差があったと言えるであろうか？

また、選択的エディティングを行うことによる副次的なメリットとして、危機管理対策を兼ねることができる点が挙げられる。たとえば、何らかの不測の事態（大規模災害などの突発的な事態）により、100 個のエディティングを終えた段階で業務を終了しなければならないという「想定外」の事態が起きるかもしれない。この場合、従来の人手によるエディティングでは、たったの 25%の完成度しか達成できていないが、選択的エディティングでは、84%の完成度を達成している。このように、従来は「想定外」であった事態も、選択的エディティングを採用することにより、「想定内」とすることが可能なのである。

3 混淆正規分布モデルによる選択的エディティング手法[8]

前節で見たとおり、影響力のある外れ値を検出することで、効率的で効果的なエディティングを行えることが分かった。本節では、影響力のある外れ値を検出する方法として、イタリア国家統計局による混淆正規分布モデル[9]を使用した多変量外れ値検出法に基づく選択的エディティングの理論を示す。

一般的に、2 つの峰を持つデータの混淆正規分布モデルは、以下の式(1)で表すことができる。すなわち、変数$\boldsymbol{x}$が混淆正規分布しているとは、$\boldsymbol{p}$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$\boldsymbol{1-p}$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。

[8] 本節の内容については、高橋(2012, pp.9-19)において、二変量の文脈における詳細な解説を掲載しているので、そちらも合わせて参照されたい。本節は、主に、Buglielli, Di Zio, Guarnera, and Pogelli (2011)、Guarnera, Luzi, Silvestri, Buglielli, Nurra, and Siesto (2012)、Di Zio and Guarnera (2012)に準拠し、多変量の理論について記述する。

[9] 混淆正規分布モデルについては、Barnett and Lewis (1994, pp.43-52)も参照されたい。

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

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く、外れ値と見なせる(DeGroot and Schervish, 2002, p.577)。本節で記述するモデルは、「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」であり、「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」である（高橋, 2012, p.8）。

今回のモデルでは、$\boldsymbol{W}$と$\boldsymbol{Z}$の 2 つの変数群を考える。$\boldsymbol{W}$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群（$n \times q$の行列）であり、$\boldsymbol{Z}$は測定誤差の影響を受けている変数（$n \times 1$のベクトル）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（$n \times 1$のベクトル）である。もし$\boldsymbol{W}$と$\boldsymbol{Z}$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。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（$n \times 1$のベクトル）、$\boldsymbol{Y}$は変数$\boldsymbol{Z}$の対数変換後の変数（$n \times 1$のベクトル）、$\boldsymbol{X}$は変数群$\boldsymbol{W}$の対数変換後の変数群（$n \times q$の行列）である。これらの変数群の関係性は、式(2)のとおりである。$\boldsymbol{\beta}$は$q \times 1$ベクトルの係数であり、$\boldsymbol{U}$は$n \times 1$ベクトルの残差（$\boldsymbol{U \sim N(0, \Sigma)}$）である。

$$\boldsymbol{Y^* = X\beta + U} \quad \textbf{(2)}$$

ここで、$\boldsymbol{\beta}$と$\boldsymbol{\Sigma}$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：$\boldsymbol{Y = Y^* + \epsilon}$であり、$\boldsymbol{\epsilon \sim N(0, \Sigma_\epsilon)}$である。ここで、$\boldsymbol{\Sigma_\epsilon = (\alpha - 1)\Sigma}$である（$\boldsymbol{\alpha > 1}$）。

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。

$$\boldsymbol{f_{Y|Y^*}(y|y^*) = p\delta(y - y^*) + (1-p)N(y; y^*, \Sigma_\epsilon)} \quad \textbf{(3)}$$

$\boldsymbol{X}$を条件とした$\boldsymbol{Y}$の観測値は式(4)の混淆正規分布となる。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、$\boldsymbol{\Sigma}$は正しいデータの分散を表し、$\boldsymbol{\Sigma_c = \Sigma + \Sigma_\epsilon}$は汚染されたデータ、つまり、エラーデータの分散を表す。

$$f_{Y|X}(y|x) = pN(y; X\beta, \Sigma) + (1-p)N(y; X\beta, \Sigma_c) \quad (4)$$

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

$$\frac{(1-p)\, N(y; X\beta, \Sigma_c)}{pN(y; X\beta, \Sigma) + (1-p)\, N(y; X\beta, \Sigma_c)} \quad (5)$$

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（Expectation-Maximization：期待値最大化法）と言う。EM アルゴリズムの発展形である 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（最大化ステップ）を CM ステップ（条件付最大化ステップ）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）。E ステップ（期待値ステップ）と CM ステップ（条件付最大化ステップ）の繰り返し適用による ECM アルゴリズムを使用して、最尤推定値を求める。

混淆正規分布による選択的エディティングを行うには、Xを含む観測データを条件として、エラーのないデータY^*の分布を導かなければならない。ベイズの公式を用いることで、xとyを条件とするy^*の条件付分布を式(6)として求めることができる。

$$f_{Y^*|X,Y}(y^*|x, y) = \tau_1(x, y)\delta(y^* - y) + \tau_2(x, y)N\left(y^*; \tilde{\mu}_{x,y}, \tilde{\Sigma}\right) \quad (6)$$

ここで、τ_1は正データに属する事後確率、τ_2はエラーデータに属する事後確率であり、それぞれ、式(7)と(8)である。

$$\tau_1 = P(y_i = y_i^* | x_i, y_i) \quad (7)$$

$$\tau_2 = P(y_i \neq y_i^* | x_i, y_i) = 1 - \tau_1 \quad (8)$$

また、$\tilde{\mu}_{x,y}$と$\tilde{\Sigma}$の定義は式(9)と(10)のとおりである。

$$\tilde{\mu}_{x,y} = \frac{\mathbf{y} + (\boldsymbol{\alpha} - \mathbf{1})\boldsymbol{\beta}\boldsymbol{x}}{\boldsymbol{\alpha}} \tag{9}$$

$$\tilde{\boldsymbol{\Sigma}} = \left(\mathbf{1} - \frac{\mathbf{1}}{\boldsymbol{\alpha}}\right)\boldsymbol{\Sigma} \tag{10}$$

上記で定義したとおり、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$\boldsymbol{Z}$の分布は、式(11)のとおりである。

$$f_{Z^*|Z}(z^*|z) = \tau_1(\log(z))\delta(z^* - z) + \tau_2(\log(z))LN\left(z^*; \tilde{\mu}_{x,log(z)}, \tilde{\Sigma}\right) \tag{11}$$

式(11)より、パラメータ$\boldsymbol{p}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\Sigma_c}$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$\hat{\boldsymbol{z}}_{\boldsymbol{i}}$を式(12)のとおり導くことができる。

$$\hat{z}_i = E(z_i^*|z_i) = \int z_i^* f_{Z^*|Z}(z_i^*|z_i) dz_i^* \tag{12}$$

したがって、期待誤差(*EE*: Expected Error)は式(13)のとおりとなる。

$$EE = (\hat{z}_i - z_i) \tag{13}$$

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、$\boldsymbol{w_i}$は、標本サイズ n の標本における各ユニットの標本抽出ウェイトである[10]。スコア関数$\boldsymbol{SF_i}$を式(14)として定義する。

$$SF_i = \left|\frac{w_i(\hat{z}_i - z_i)}{\sum w_i \hat{z}_i}\right| \tag{14}$$

こうして算出した$\boldsymbol{SF_i}$の値に応じて、ローカルスコアとグローバルスコアを算出し、観測値を並び替え、影響力の強い順にエディティングを行っていくのである[11]。具体的には、グローバルスコアの値に応じて観測値を降順で並び替え、残差エラーがあらかじめ設定した閾値よりも下の値を影響力のある外れ値として検出する。

[10] 各ユニットの標本抽出方法に違いがある場合に、設定することのできるものであり、既定では 1 となる。

[11] ローカルスコアとは優先的に処理するべきユニットの回答を数値化する指標であり、グローバルスコアとは優先的に処理するべきユニットのレコード全体を数値化する指標のことである(Scarrott, 2007, p.5)。イタリア国家統計局の開発した *SeleMix* では、ローカルスコアは観測値と予測値の差に重み付けを行い、絶対値を取ったものであり、グローバルスコアは各々の変数のローカルスコアの最大値である(Guarnera and Buglielli, 2013, p.7)。

4 *SeleMix* の検証：EDINET データ

Latouche and Berthelot (1992)以来、スコア関数に基づく様々な選択的エディティング手法が提唱されてきた。1.1 項で見たとおり、選択的エディティングは、今日においても日々、進化を続けている手法である。高橋(2012, pp.7-9)で示したとおり、イタリア国家統計局では、2002 年から混合モデルに基づく選択的エディティング手法の開発に取り組んできた。Buglielli, Di Zio, Guarnera, and Pogelli (2011)は、10 年にわたる研究を集大成させるものであり、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[12]。

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り、高橋(2012)に引き続き、EDINET のデータを模擬試験データとして利用し、*SeleMix* の検証を行う。本稿では、4 変量における多変量外れ値検出を行う。人工的にエラーを付置し、それらの検出を行えるかどうかを検証し、実際に選択的エディティングによってエラーの訂正を行うことで、その精度も検証する。

4.1 EDINET データ

EDINET とは、**E**lectronic **D**isclosure for **I**nvestors' **NET**work の略であり、『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』を意味する。これは、「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（金融庁, 2012）。今回の例では、欠測値を除外した 2,871 レコードを使用する[13]。

対象となる変数は、経済センサス - 活動調査における「売上（収入）金額」、「売上原価」、「資本金又は出資金、基金の額」、「従業者数」である。経済センサス - 活動調査のデータを実際に利用する前に、これら 4 つの変数に該当する EDINET のデータを使用して検証を行う。すなわち、「売上高」、「売上原価合計」、「資本金」、「事業従事者数」である。対応関係は表 4.1 に示すとおりである。

---

[12] CRAN (Comprehensive R Archive Network)のウェブサイトより無料でダウンロードし（http://cran.r-project.org/web/packages/SeleMix/index.html）、R に実装することで、誰でも使用可能となっている（2013 年 7 月 11 日アクセス）。また、*SeleMix* パッケージの関数については、高橋(2012, pp.22-28)及び Guarnera and Buglielli (2013)を参照されたい。

[13] *SeleMix* による外れ値検出では、X はエラーのない状態でなければならないため、今回は実験の目的で X の欠測値をすべて除外した。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 4.1：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
売上原価	売上原価合計	cost
資本金又は出資金、基金の額	資本金	capital
従業者数	事業従事者数	worker

以下の 4 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、売上原価への支出が大きければ大きいほど売上も大きくなると考える。資本金が大きければ大きいほど、また、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、大きい事業ほど売上も大きくなると考える。

EDINET におけるこれら 4 つの変数の基本統計量は表 4.2 に示すとおりである。「売上高」、「売上原価合計」、「資本金」の単位は 100 万円、「事業従事者数」の単位は人（1 人）である。すなわち、今回のデータにおける最大の売上高は 8 兆 9810 億円であり、最小の売上高は 800 万円であった。最大の売上原価は 8 兆 8220 億円であり、最小の売上原価は 100 万円であった。最大の資本金は 2 兆 3380 億円であり、最小の資本金は 1 億円であった。最大の事業従事者数は 2 万 2050 人であり、最小の事業従事者数は 1 人であった。

表 4.2：各変数の基本統計量（生データ）

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	7633	19830	106300	59920	8981000	406832
cost	1	5041	14450	86730	44220	8822000	370469
capital	100	1052	2695	15630	8022	2338000	84923
worker	1	65	145	364	327	22050	944

また、各変数間の相関係数は表 4.3 に示すとおりである。今回のデータでは、売上高と売上原価との相関が 0.989 と最も高く、売上高と事業従事者数との相関が 0.502 で続いており、売上高と資本金との相関は 0.367 となっている。また、説明変数間では、売上原価と事業従事者数との相関が 0.475 と最も高く、売上原価と資本金との相関が 0.340、資本金と事業従事者数との相関は 0.216 となっている。

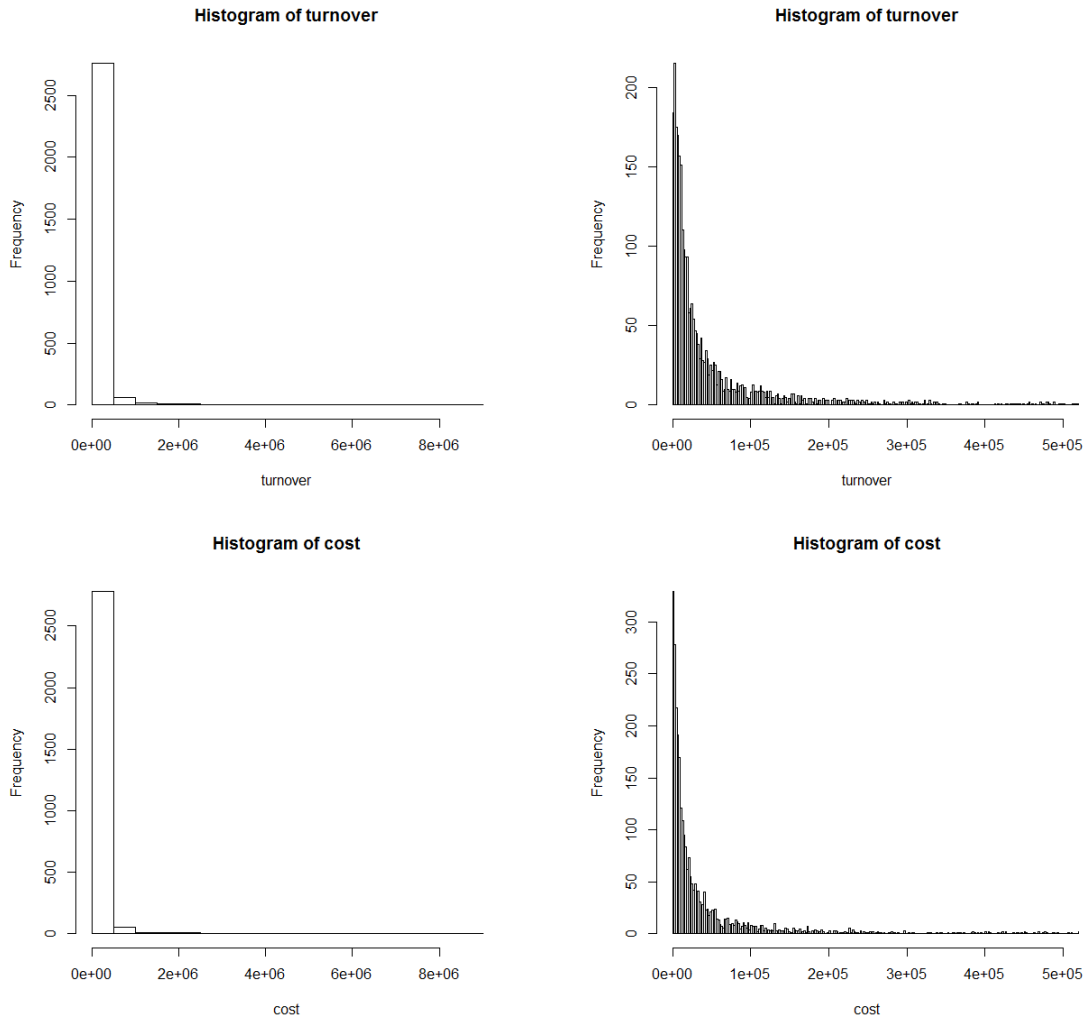
表 4.3：相関係数（生データ）

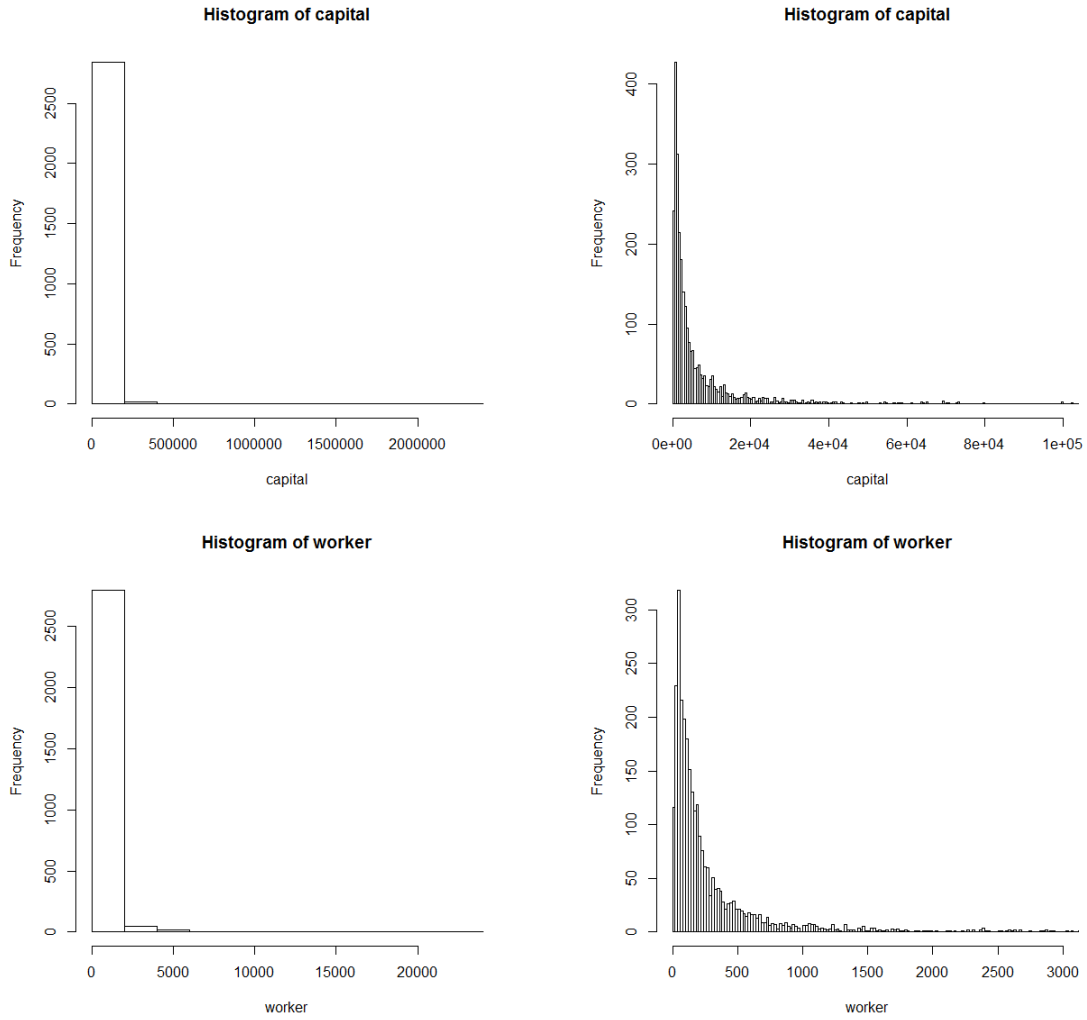
	turnover	cost	capital	worker
turnover	1.000			
cost	0.989	1.000		
capital	0.367	0.340	1.000	
worker	0.502	0.475	0.216	1.000

これらの変数のヒストグラムは図 4.1 のとおりである。経済データによくあるように、売上高、売上原価、資本金、事業従事者数のいずれも偏りがあり、合理的に正規分布に近いとは言えないことが視覚的に分かる。

図 4.1：各変数のヒストグラム（生データ）

全体図　　拡大図





完全な正規分布は、歪度(わいど、S: Skewness) = 0、尖度(せんど、K: Kurtosis) = 3 となり、歪度と尖度は、それぞれ、式(15)と(16)のとおり求められる(Gujarati, 2003, p.886, p.890; Greene, 2003, pp.848-849)。ここで、μは平均値を表し、σは標準偏差を表す。また、$E(X-\mu)^2$は二次積率である分散(σ^2)であり、$E(X-\mu)^3$は三次積率であり、$E(X-\mu)^4$は四次積率である。

$$S = \frac{E(X-\mu)^3}{\left[\sqrt{E(X-\mu)^2}\right]^3} = \frac{E(X-\mu)^3}{\left[\sqrt{\sigma^2}\right]^3} = \frac{E(X-\mu)^3}{\sigma^3} \qquad (15)$$

$$K = \frac{E(X-\mu)^4}{[E(X-\mu)^2]^2} = \frac{E(X-\mu)^4}{[\sigma^2]^2} = \frac{E(X-\mu)^4}{\sigma^4} \qquad (16)$$

それぞれの変数の歪度と尖度を表 4.4 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おり、数値的にも正規分布とは異なっていることが分かる。

表 4.4：歪度と尖度（生データ）

変数	歪度	尖度
turnover	11.200	178.153
cost	12.416	215.010
capital	19.950	491.167
worker	11.690	210.394

注：正規分布の場合、歪度 = 0、尖度 = 3

したがって、分布の歪みを矯正する必要がある。一般的に、経済データの歪みは、自然対数変換により是正できることが多いため、自然対数による変換を行ってみることとする。自然対数変換後の各変数の基本統計量は表 4.5 に示すとおりである。どの変数においても、平均値と中央値がほぼ同じ値になり、平均値から第 1 四分位と平均値から第 3 四分位までの距離がほぼ均等になっている。

表 4.5：各変数の基本統計量（自然対数データ）

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
logturnover	2.092	8.940	9.895	9.982	11.000	16.010	1.700
logcost	0.000	8.525	9.578	9.587	10.700	15.990	1.862
logcapital	4.605	6.958	7.899	8.078	8.990	14.660	1.516
logworker	0.000	4.174	4.977	5.028	5.790	10.000	1.225

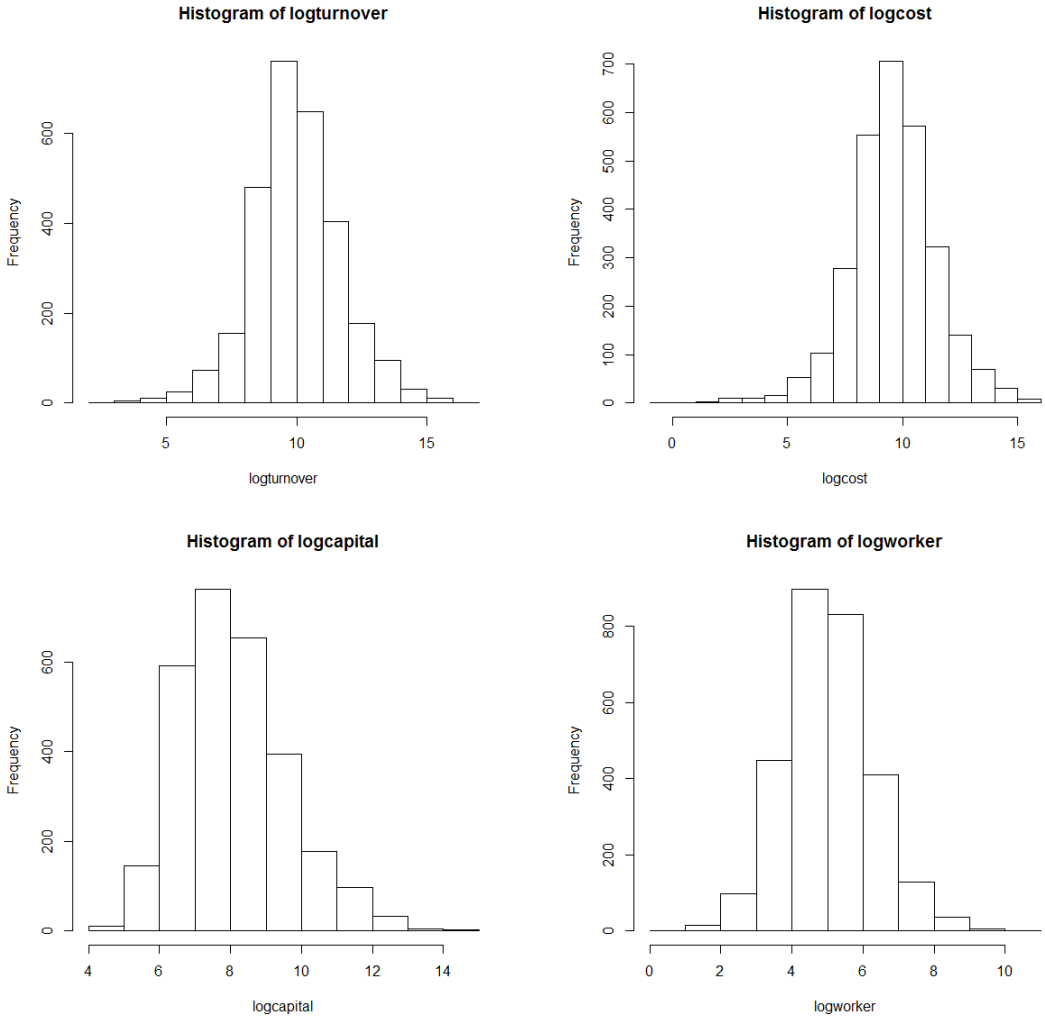
また、各変数間の相関係数は表 4.6 に示すとおりである。最も高い相関は、売上高と売上原価の 0.966 であり、生データからわずかに下がったものの依然として高い。売上高と事業従事者数との相関は 0.606、売上高と資本金の相関は 0.699 と生データよりも高くなっている。また、説明変数間では、売上原価と資本金との相関が 0.641 と最も高く、売上原価と事業従事者数との相関は 0.577、資本金と事業従事者数との相関は 0.501 となっている。

表 4.6：相関係数（自然対数データ）

	logturnover	logcost	logcapital	logworker
logturnover	1.000			
logcost	0.966	1.000		
logcapital	0.699	0.641	1.000	
logworker	0.606	0.577	0.501	1.000

自然対数変換後のヒストグラムは図 4.2 のとおりである。自然対数に変換することにより、合理的に正規分布を近似していることが視覚的に分かる。

図 4.2：ヒストグラム（自然対数データ）



それぞれの変数の歪度と尖度を表 4.7 に示す。自然対数データの歪度と尖度は、生データと比較して、0 と 3 に近づいており、数値的にも正規分布を近似していることが分かる。

表 4.7：歪度と尖度（自然対数データ）

変数	歪度	尖度
logturnover	0.033	3.890
logcost	-0.238	4.382
logcapital	0.671	3.514
logworker	0.270	3.532

注：正規分布の場合、歪度 ＝0、尖度 ＝3

4.2 外れ値（エラー）の生成方法

実データセットとしての EDINET データには、外れ値は存在しているが、非常に稀な虚偽報告の例を除くと、エラーは存在していない。選択的エディティングの最終目標は、単なる外れ値の検出ではなく、エラーを効率的に抽出し対処することにある。そのため、今回の実験では、下記の要領にて、報告された真の値の約 15%を人工的にエラー化し、桁違いによるランダムエラーを模した[14]。

まず、2,871 個の標準正規乱数[15]を発生させ、各々のユニットに割り振る。その後、標準正規乱数の値が 1.44 以上のとき、売上高の値を 10 倍にした（227 個）。また、標準正規乱数の値が-1.44 以下のとき、売上高の値を 1/10 倍にした（234 個）。実際のエラー含有率は、16.057% = ((227 + 234) / 2871 * 100)となった。表 4.8 では、logturnover は正データの基本統計量（表 4.5 と同一）であり、logturnover_cはエラーを含むデータの基本統計量である。平均値は正データとほぼ同じだが、真の標準偏差が 1.700 であるのに対して、エラーのあるデータの標準偏差は 1.943 となっており、1.143 倍に膨れ上がっている。

表 4.8：エラーを含む売上高の基本統計量（自然対数データ）

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
logturnover	2.092	8.940	9.895	9.982	11.000	16.010	1.700
logturnover_c	0.779	8.792	9.898	9.976	11.190	17.400	1.943

表 4.9 では、logturnover は正データと各変数との相関であり（表 4.6 と同一）、logturnover_cはエラーを含むデータと各変数との相関である。いずれの値も正データと比較して低い値となっており、ランダムエラーがノイズとして影響を及ぼしている。

表 4.9：相関係数（自然対数データ）

	logturnover	logturnover_c
logcost	0.966	0.848
logcapital	0.699	0.622
logworker	0.606	0.534

表 4.10 は、エラーデータ（自然対数）の歪度と尖度を表している。歪度は 0 に近く、尖度は若干 3 よりも大きいが、比較的、正規分布に近い値となっている。

[14] 経済の実データでは、そもそも大きな値というものが含まれているため、このような手順で生成したランダムエラーは、通常の手法では検出することが非常に困難なものである。
[15] 使用したシード値は、分析した時刻（10:05）に基づき 1005 とした。

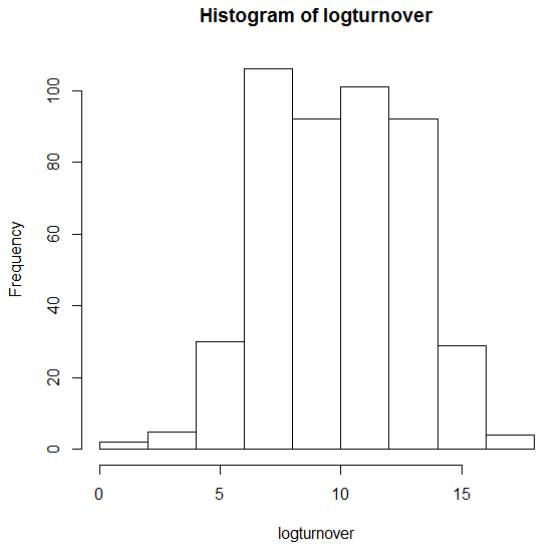
表 4.10：歪度と尖度（自然対数データ）

変数	歪度	尖度
売上エラー	0.007	3.842

注：正規分布の場合、歪度 ＝0、尖度 ＝3

図 4.3 は、エラーデータのヒストグラムである。少々いびつな部分があるものの、中心付近に値が多く見られ、裾に行くにつれて少なくなっていっており、正規分布を近似していると言える。

図 4.3：エラーのヒストグラム



4.3 真のモデルとエラーを含むモデル

表 4.11 では、モデル 1 はエラーのない真のモデルであり、モデル 2 はエラーの存在しているモデルである。つまり、モデル 1 の係数及び標準誤差が正しい値であり、モデル 2 の値はエラーによる影響を受けたものである。

表 4.11

	モデル 1	モデル 2
切片	1.004 (**0.044**)	0.909 (**0.112**)
logcost	0.783 (**0.006**)	0.776 (**0.014**)
logcapital	0.141 (**0.007**)	0.159 (**0.017**)
logworker	0.067 (**0.008**)	0.067 (**0.019**)
R^2	0.945	0.731
Adjusted R^2	0.945	0.730
n	2871	2871

注：被説明変数は logturnover；報告値は係数（標準誤差）

上述したとおり、今回の実験におけるエラーは、ランダムエラーである。すなわち、ランダムなノイズであり、バイアスはほとんどないため、係数への影響は少なかったことが見て取れる。

一方、ランダムエラーは、ばらつき（分散）に影響を与えるため、**標準誤差が平均して 2.421 倍（2.333 倍～2.545 倍）に肥大化**している。エラーのないモデル 1 の分析では、説明変数の値が増加すると、売上高も増えると結論付けられたが、エラーのあるモデル 2 の分析では、説明変数と売上高との間には、統計的な相関が存在しないという誤った結論となってしまう恐れがある。

以上のとおり、エラーが存在することによって、統計分析の結果が異なってしまう恐れがあるため、エラーによる影響は無視できないことが分かる。

4.4 単変量外れ値検出法によるエラーの検出

前項では、エラーによって統計分析に影響が出ることが分かった。そのような影響力のあるエラーを検出する方法として、本稿では、混淆正規分布モデルによる多変量外れ値検出法を推奨している。一方、2.1 項の例では、z 値を用いた単変量外れ値の検出例を示した。また、2.2 項で述べたとおり、一般的に、IQR (四分位範囲)の 1.5 倍を超える値を単変量外れ値としている：上限 = Q3 + 1.5*IQR；下限 = Q1 – 1.5*IQR。

そこで、混淆正規分布モデルのような高度な手法を用いずとも、簡単な単変量外れ値検出法で十分ではないかという疑問があるだろう。本項では、z 値と IQR の 1.5 倍の基準を用いて、エラーの検出を正確に行えるかどうかを試してみる。

今回のエラーは、標準正規乱数の z 値が 1.44 以上または-1.44 以下のときに発生するメカニズムとなっていた。そこで、エラーを含む自然対数データの売上高の z 値が、1.44 以上または-1.44 以下となる値を外れ値として検出した。つまり、分布の上側 7.5%と下側 7.5%

の合計 15%を外れ値として検出するということである。その結果、403 個の外れ値が検出されたが、その中で実際にエラーであったものは 175 個であり、正答率は **43.424%**に過ぎなかった。

```
turnoverz<-(logturnover-mean(logturnover))/sd(logturnover)
```

一方、IQR の 1.5 倍の基準では、エラーを含む自然対数データにおける売上高の上限は 14.776 であり、この値を超えるデータは 32 個検出されたが、その中で実際にエラーであったものは 20 個だった。また、エラーを含む自然対数データにおける売上高の下限は 5.202 であり、この値を下回るデータは 31 個検出されたが、その中で実際にエラーであったものは 15 個だった。トータルで検出した外れ値は 63 個であったが、その中で実際にエラーであったものは 35 個であり、正答率は **55.556%**に過ぎなかった。

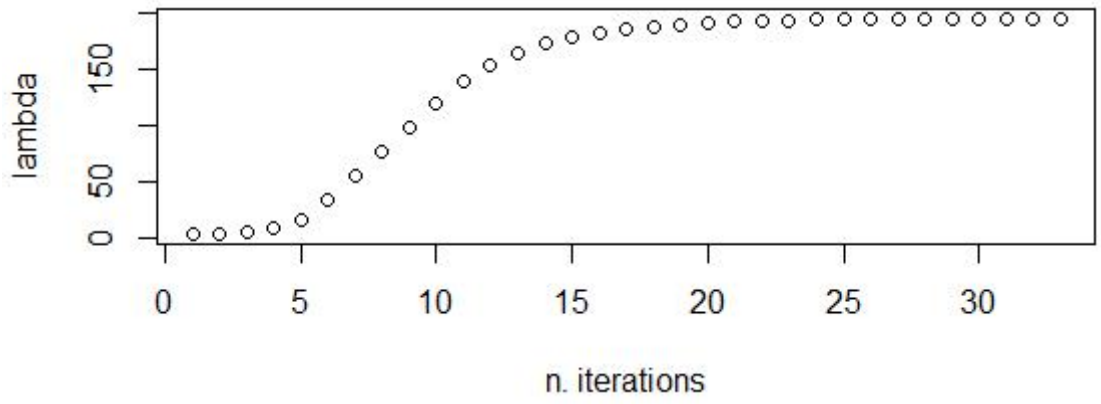
```
UL<-logturnover-(quantile(logturnover,probs=0.75,names=F)+1.5*IQR(l
ogturnover))
LL<-logturnover-(quantile(logturnover,probs=0.25,names=F)-1.5*IQR(l
ogturnover))
```

結論として、単純な単変量外れ値検出法では、エラーの特定を正確に行うことができないということがはっきりと分かる。

4.5 *SeleMix* による外れ値検出の精度評価

前項で見たとおり、単変量外れ値検出法では、多変量エラーを正確に検出することができなかった。そこで、本項では、*SeleMix* を用いて、売上原価、資本金、事業従事者数を条件とした売上高の多変量外れ値検出を行い、エラーの検出を的確に行うことができるかどうかを検証した。図 4.4 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（lambda は分散拡大要因の値）。今回の実験では 32 回の繰り返しの後に収束し、実際に収束するまでにかかった時間は**約 9 秒**と高速であった。

図 4.4：ECM アルゴリズムの収束



混淆正規分布モデルの推定値は表 4.12 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 4.12：モデルの結果

パラメータ	推定値(混淆正規)	推定値(通常 OLS)
$\widehat{\boldsymbol{\beta}}_{\boldsymbol{0}}$	0.592	0.909
$\widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}$	0.927	0.776
$\widehat{\boldsymbol{\beta}}_{\boldsymbol{2}}$	0.037	0.159
$\widehat{\boldsymbol{\beta}}_{\boldsymbol{3}}$	0.017	0.067
sigma	0.016	
lambda	195.000	
w	0.340	
BIC	**4100.000**	**8238.000**
AIC	**2036.000**	**4109.000**

今回の実験では、2,871 観測数のうち 857 個の外れ値が検出された。中でも、238 個は、影響力のある外れ値として検出され、優先的にエディティングをすべきものとして選択された。また、今回の実験では、461 個のエラーを人工的に発生させていた。エラーデータ 461 個のうち、外れ値として検出できたものは 460 個であり、正答率は **99.783%**であった。また、影響力のある外れ値として検出した 238 個のうち、エラーデータであったものは 207 個であり、正答率は **86.975%**であった。

4.6 図による外れ値検出法との比較

図 4.5a は売上高の箱ひげ図であり、図 4.5b は売上原価の箱ひげ図であり、図 4.5c は事業従事者数の箱ひげ図であり、図 4.5d は資本金の箱ひげ図である。ここでは、通常の値を白丸、*SeleMix* により検出した外れ値を**黒丸**で示している。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いる。

図 4.5a

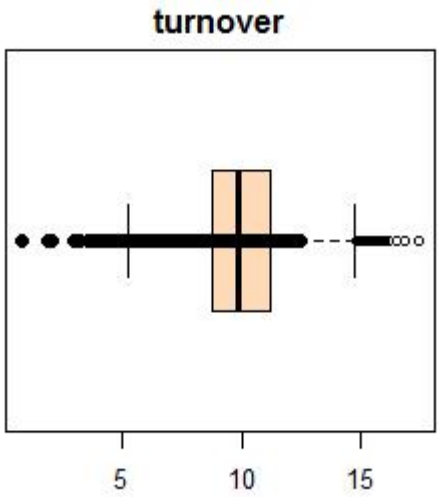


図 4.5b

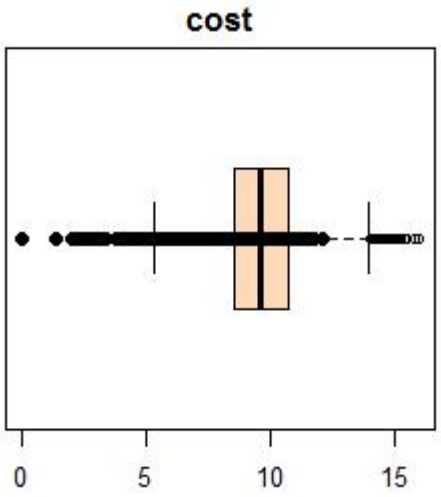


図 4.5c

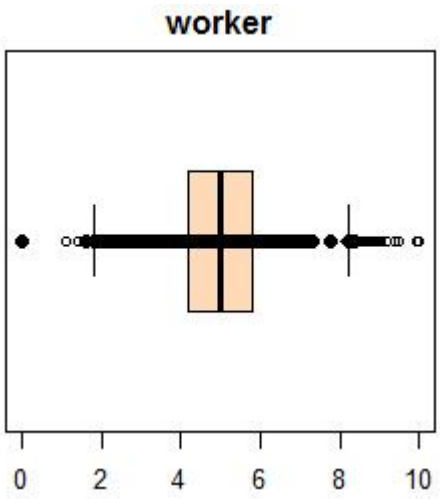


図 4.5d

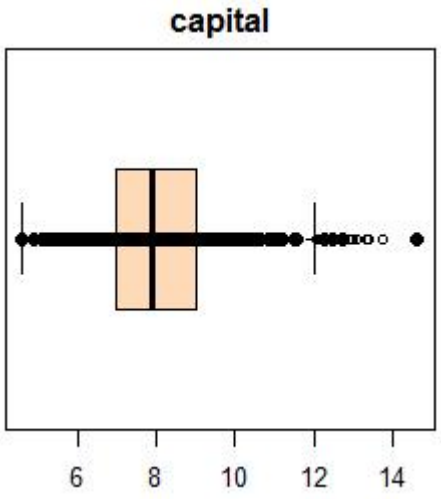


図 4.6a は、売上高（縦軸）と売上原価（横軸）の散布図であり、図 4.6b は、売上高（縦軸）と事業従事者数（横軸）の散布図であり、図 4.6c は、売上高（縦軸）と資本金（横軸）の散布図である。ここでは、通常の値を白丸、*SeleMix* により検出した外れ値を**黒丸**で示している。2 変量散布図では、外れ値の多くが中心付近に埋もれており、検出することができないことが分かる。

図 4.6a：売上高と売上原価

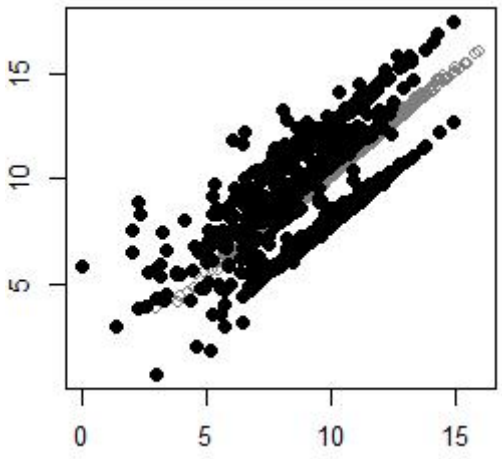


図 4.6b：売上高と事業従事者数

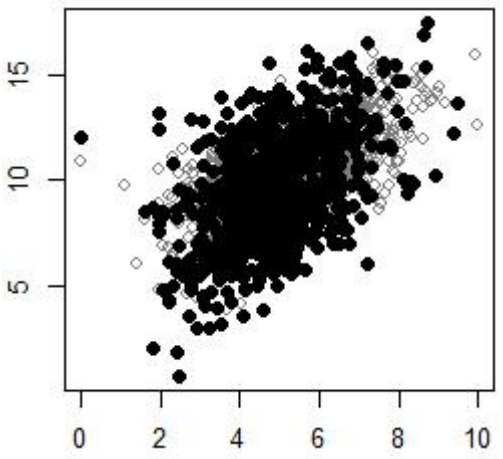
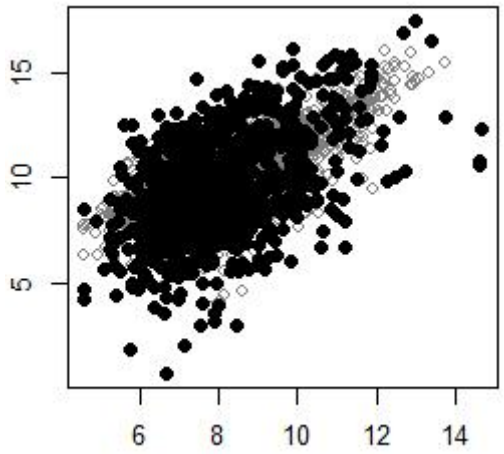


図 4.6c：売上高と資本金



4.7 図による影響力のある外れ値検出法との比較

図 4.7a は売上高の箱ひげ図であり、図 4.7b は売上原価の箱ひげ図であり、図 4.7c は事業従事者数の箱ひげ図であり、図 4.7d は資本金の箱ひげ図である。ここでは、通常の値を白丸、*SeleMix* により検出した影響力の強い外れ値を**菱形**で示している。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いる。

図 4.7a

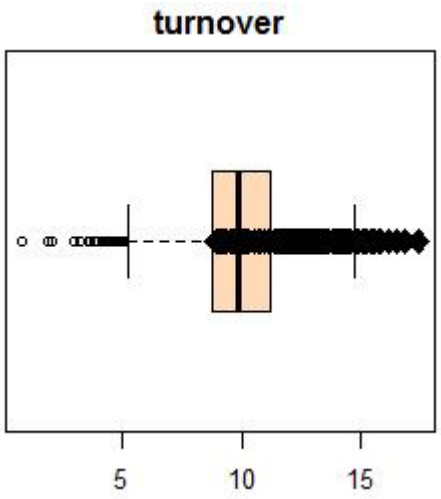


図 4.7b

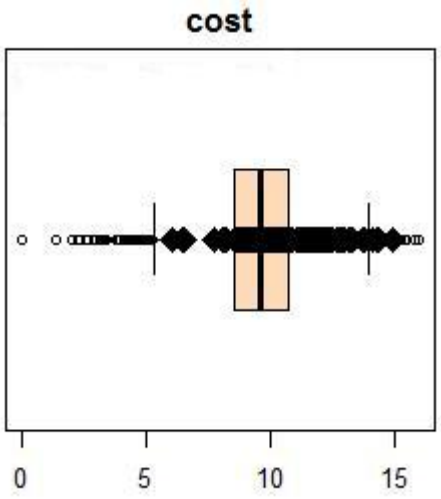


図 4.7c

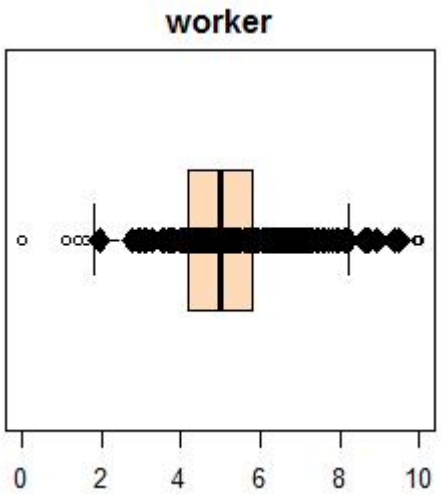


図 4.7d

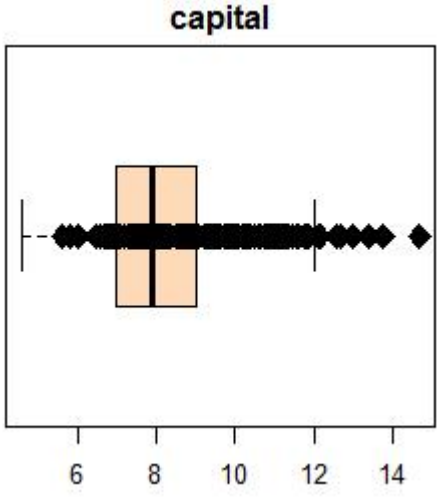


図 4.8a は、売上高（縦軸）と売上原価（横軸）の散布図であり、図 4.8b は、売上高（縦軸）と事業従事者数（横軸）の散布図であり、図 4.8c は、売上高（縦軸）と資本金（横軸）の散布図である。ここでは、通常の値を白丸、*SeleMix* により検出した影響力の強い外れ値を**菱形**で示している。図 4.6 と比較することで、必ずしも外れ値のすべてが影響力ありと判断されている訳ではないことが分かる。

図 4.8a：売上高と売上原価

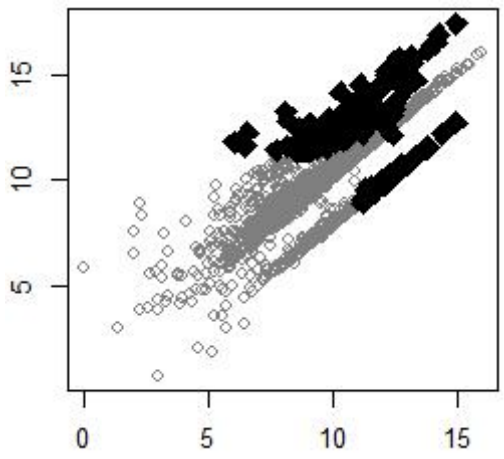


図 4.8b：売上高と事業従事者数

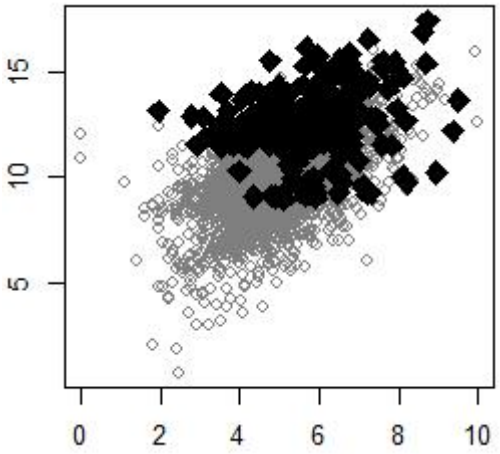
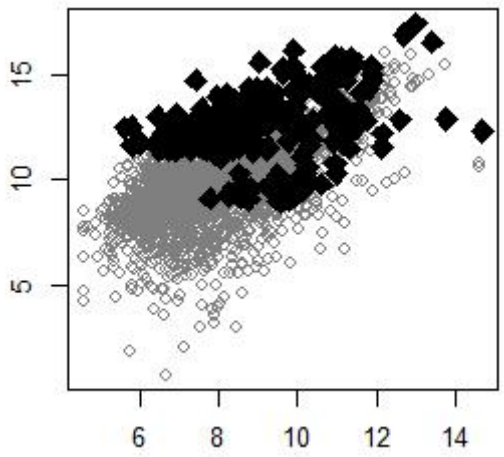


図 4.8c：売上高と資本金



4.8 検出した外れ値（エラー候補）への対処

4.5 項において、多変量外れ値としてエラーの候補を検出した。本項では、検出したエラーの候補を人手訂正と機械訂正によって処理を行う。

表 4.13 に結果を示す。モデル 1 は、真のモデルである。モデル 2 は、エラーのあったモデルである。これら 2 つのモデルは、表 4.11 と同じものである。モデル 3 は、選択的エディティングにより検出した影響力のある外れ値を人手により審査し、検出されたエラー238 個を真値に置き換え、統計分析を行ったものである（選択的エディティング＋人手訂正モデル）。モデル 4 は、選択的エディティングにより検出した影響力のある外れ値を欠測させ、*Amelia* による多重代入法(M = 20)[16]で補定し、統計分析を行ったものである（選択的エディティング＋機械訂正モデル）。

表 4.13

	モデル 1：真のモデル	モデル 2：エラーモデル	モデル 3：人手訂正モデル	モデル 4：機械訂正モデル
切片	1.004 **(0.044)**	0.909 **(0.112)**	0.879 **(0.087)**	0.995 **(0.096)**
logcost	0.783 **(0.006)**	0.776 **(0.014)**	0.768 **(0.011)**	0.779 **(0.012)**
logcapital	0.141 **(0.007)**	0.159 **(0.017)**	0.155 **(0.013)**	0.132 **(0.014)**
logworker	0.067 **(0.008)**	0.067 **(0.019)**	0.077 **(0.015)**	0.062 **(0.016)**
R^2	0.945	0.731	0.814	0.807
Adjusted R^2	0.945	0.730	0.814	0.807
n	2871	2871	2871	2871

注：被説明変数は logturnover；報告値は係数（標準誤差）

影響力のあるエラーを訂正したことにより、モデル 3 及びモデル 4 では、すべての**標準誤差**が、モデル 2 と比べて、モデル 1 の真の値に近づいている。モデル 3 とモデル 4 の比較では、概ね、同等の結果が得られたと言える。

4.9 閾値の設定

4.5 項の分析では、2,871 観測数のうち 857 個の外れ値を検出し、238 個の影響力のある外れ値を検出した。エラーデータ 461 個のうち、外れ値として検出できたものは 460 個（正答率は **99.783%**）であり、影響力のある外れ値として検出した 238 個のうち、エラーデータであったものは 207 個（正答率は **86.975%**）であった。

[16] 多重代入法及び *Amelia* については、高橋, 伊藤(2013)を参照されたい。

しかし、モデルの検出力は、設定した閾値の値に応じて変化してくる。今回は、*SeleMix* プログラムのデフォルト設定に従い、残差エラーが 0.01 未満のとき、影響力のある外れ値として検出した。

表 4.14 に示すとおり、検出できる外れ値及びエラーの絶対数は、閾値の値が大きくなるにつれて少なくなる。一方、正答率は、閾値の値を大きくすればするほど改善していく。これは、統計的検定における第 1 種過誤と第 2 種過誤の関係に対比して考えることができよう。

表 4.14

閾値	0.001	0.005	0.010	0.020
外れ値	641	312	238	176
エラー	399	248	207	154
正答率	62.2%	79.5%	87.0%	87.5%

今回の結果では、デフォルトの 0.010 の結果が、最もバランスが良さそうに思われる。しかし、異なるデータセットにおける様々な状況に応じて、適切な閾値の設定方法を検討する必要があるだろう。

5 *SeleMix* の検証：模擬経済センサスデータ

前節で使用した EDINET データは、事業所・企業の実データではあるが、観測数が数千しか存在しない。経済センサスにおいて対象となる事業所・企業の数は約 580 万であり、このうち、産業分類などの情報を用いて、いくつかの層に分けるため、580 万のデータを一括して処理するわけではないが、潜在的に、最大の層は数十万以上の観測数が存在する可能性がある。そこで、経済センサスのデータ処理を目指し、経済センサスのデータサイズを模したシミュレーションデータ（観測数 100 万、4 変数）に、人工的にエラーを埋め込み、それを正しく探し出せるかどうかを検証する。

5.1 模擬 EDINET データ

EDINET データセットの情報（平均値、分散・共分散）をもとに、シミュレーションデータを下記の要領で作成した。本項では、シミュレーションデータセットが、自然対数変換後の EDINET データセットに近似していることを示す。この情報をもとに、次項では、観測数 100 万の模擬経済センサスデータセットを作成する。

```
set.seed(1223)
library(MASS)
varcov<-matrix(c(
  2.889857,3.057617,1.8015284,1.2613350,
  3.057617,3.468288,1.8089893,1.3160773,
  1.801528,1.808989,2.2984151,0.9302662,
  1.261335,1.316077,0.9302662,1.5014083
  ),4,4)
z<-mvrnorm(
  n=2871,
  mu=c(9.982,9.587,8.078,5.028),
  Sigma=varcov,
  empirical=TRUE
)
```

上記の手順により生成したデータの基本統計量、相関係数、回帰分析の結果を表 5.1～表 5.3 に示す。EDINET データによる結果をほぼ完全に復元しており、相関係数及び回帰分析の結果は、完全に復元[17]されていることが分かる（表 4.5、表 4.6、表 4.9 参照）。

表 5.1：各変数の基本統計量（模擬 EDINET データ：n = 2,871）

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
logturnover	3.410	8.867	9.989	9.982	11.119	15.497	1.700
logcost	3.778	8.361	9.598	9.587	10.836	16.089	1.862
logcapital	2.546	7.035	8.074	8.078	9.105	12.822	1.516
logworker	-1.346	4.248	5.029	5.028	5.845	10.256	1.225

表 5.2：相関係数（模擬 EDINET データ）

	logturnover	logcost	logcapital	logworker
logturnover	1.000			
logcost	0.966	1.000		
logcapital	0.699	0.641	1.000	
logworker	0.606	0.577	0.501	1.000

[17] 平均値及び分散・共分散の情報が同一であるため、相関係数及び回帰分析の結果が同一になることは、統計学的に必然ではあるが、確認のために掲載している。

表 5.3：回帰分析（模擬 EDINET データ）

	モデル 1	
切片	1.005	(0.044)
logcost	0.783	(0.006)
logcapital	0.141	(0.007)
logworker	0.067	(0.008)
R^2	0.945	
Adjusted R^2	0.945	
n	2871	

注：被説明変数は logturnover；報告値は係数（標準誤差）

5.2 模擬経済センサスデータ

5.1 項の生成方法により、事業所・企業のデータを模したデータセットを生成できることが分かった。そこで、本項では、5.1 項の生成方法を用い、`mvrnorm` 関数の `n=`の右辺を `1000000` に変更し、模擬経済センサスデータを作成した。観測数の変更に伴い、基本統計量に変化があったため、表 5.4 に結果を示す。また、観測数が増大したことにより、表 5.5 に示すとおり、回帰分析における標準誤差の値が小さくなっている。それ以外の情報は、5.1 項のデータセットとほぼ同じである。

表 5.4：各変数の基本統計量（模擬経済センサスデータ：n = 1,000,000）

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
logturnover	2.050	8.835	9.983	9.982	11.130	18.224	1.700
logcost	0.904	8.331	9.588	9.587	10.842	18.660	1.862
logcapital	0.373	7.057	8.078	8.078	9.100	15.573	1.516
logworker[18]	0.000	4.200	5.029	5.028	5.855	10.849	1.225

[18] 事業従事者数の最小値は「1 人」のため、シミュレーションにより logworker の値が 0 未満の値になったものはすべて log(1) = 0 として処理した。

表 5.5：回帰分析（模擬経済センサスデータ）

	モデル 1	
切片	1.004	(0.0024)
logcost	0.783	(0.0003)
logcapital	0.141	(0.0003)
logworker	0.067	(0.0004)
R^2	0.945	
Adjusted R^2	0.945	
n	1000000	

注：被説明変数は logturnover；報告値は係数（標準誤差）

5.3 *SeleMix* による外れ値検出の精度評価

エラーの生成方法は、前節に準じ、100 万の標準正規乱数を発生させ、その値が 1.44 以上のとき logturnover の値を 10 倍にし、-1.44 以下のとき logturnover の値を 1/10 倍にした。実際には、対数変換後のデータを模しているので、2.303 を加減して生成した[19]。エラーの数は 150,761 個（= 75624 + 75137）であり、エラー含有率は 15.076%である。表 5.6 では、logturnoverは正データの基本統計量（表 5.4 と同一）であり、logturnover_cはエラーを含むデータの基本統計量である。平均値は正データとほぼ同じだが、真の標準偏差は 1.700 であるのに対して、エラーのあるデータの標準偏差は 1.920 と大きくなっている。

表 5.6：エラーを含む logturnover の基本統計量（模擬経済センサスデータ：n = 1,000,000）

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
logturnover	2.050	8.835	9.983	9.982	11.130	18.224	1.700
logturnover_c	-0.253	8.716	9.984	9.983	11.250	19.570	1.920

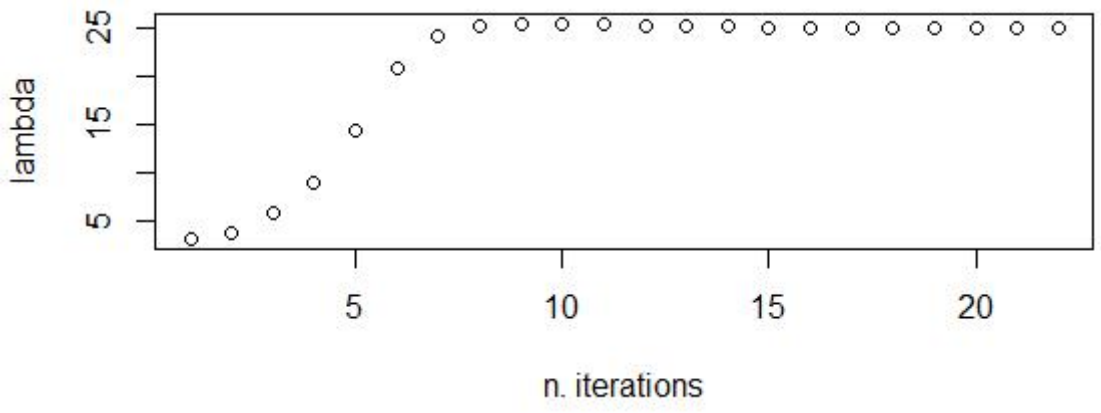
以下、*SeleMix* を用いて、logcost、logcapital、logworker を条件として、logturnover の多変量外れ値検出を行った。図 5.1 は、ECM アルゴリズムが収束するまでにかかった回数を図示している。今回の実験では、21 回の繰り返しの後に収束し、実際に収束するまでにかかった時間は**約 34 分 25 秒**であった[20]。観測数 100 万という非常に巨大なデータセッ

[19] 対数の公式より、$\log(Y \times 10) = \log(Y) + \log(10)$ であり、$\log(Y/10) = \log(Y) - \log(10)$ であり、$\log(10) \approx 2.303$ である。

[20] 検証に用いたパソコンは、Windows Vista を搭載した一般的なノートパソコンである。プロセッサは Intel Core 2 Duo CPU T9400、メモリは 2.00 GB、システムの種類は 32 ビットオペレーティングシステムという性能となっている。

トであるため、収束には時間がかかったが、最大規模のデータセットであっても、十分に機能することが分かった。

図 5.1：ECM アルゴリズムの収束



混淆正規分布モデルの推定値は表 5.7 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 5.7：モデルの結果

パラメータ	推定値(混淆正規)	推定値(通常 OLS)
$\widehat{\boldsymbol{\beta}}_0$	1.002	1.010
$\widehat{\boldsymbol{\beta}}_1$	0.783	0.783
$\widehat{\boldsymbol{\beta}}_2$	0.141	0.141
$\widehat{\boldsymbol{\beta}}_3$	0.067	0.066
sigma	0.122	
lambda	25.000	
w	0.276	
BIC	**2199018**	**2793994**
AIC	**1099475**	**1396972**

今回の実験では、1,000,000 観測数のうち 186,385 個の外れ値が検出された。中でも、167,434 個は、影響力のある外れ値として検出され、優先的にエディティングをすべきものとして選択された。また、今回の実験では、150,761 個のエラーを人工的に発生させていた。エラーデータ 150,761 個のうち、外れ値として検出できたものは 150,745 個であり、正答率は **99.989%**であった。また、影響力のある外れ値として検出した 167,434 個のうち、エラーデータであったものは 150,709 個であり、正答率は **90.011%**であった。

5.4 図による外れ値検出法との比較

4 節と同様に、以下の図では、通常の値を白丸、「外れ値」を**黒丸**で図示する。図 5.2a は logturnover の箱ひげ図であり、図 5.2b は logcost の箱ひげ図であり、図 5.2c は logworker の箱ひげ図であり、図 5.2d は logcapital 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(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。さらに、単変量の文脈で外れている値は、必ずしも多変量外れ値として認定されていない（箱ひげ図の外にある白丸）。

図 5.2a

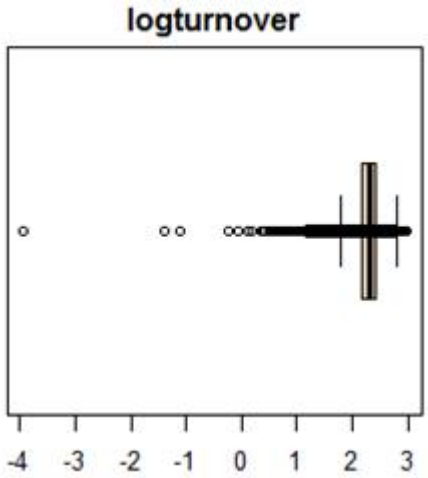


図 5.2b

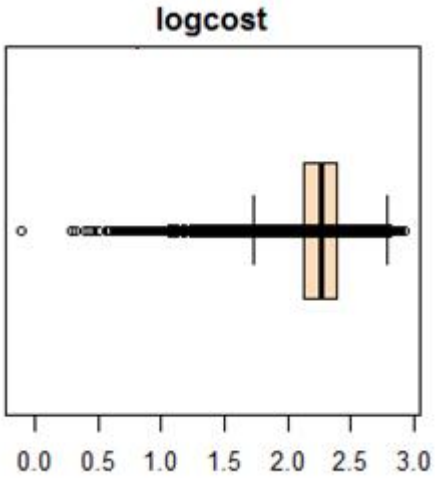


図 5.2c

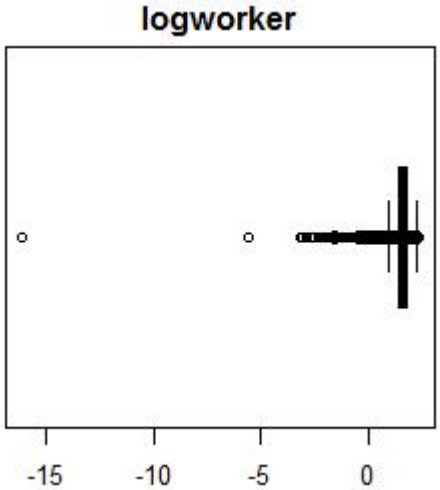


図 5.2d

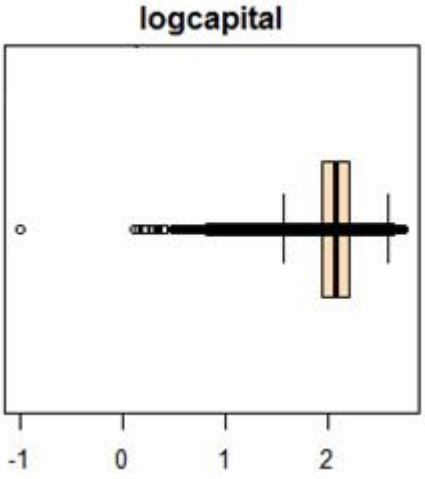


図 5.3a は、logturnover（縦軸）と logcost（横軸）の散布図であり、図 5.3b は、logturnover（縦軸）と logworker（横軸）の散布図であり、図 5.3c は、logturnover（縦軸）と logcapital（横軸）の散布図である。2 変量散布図では、外れ値の多くが中心付近に埋もれており、検出することができないことが分かる。さらに、観測数が 100 万ともなると、もはや、どこにどの値があるのか分からず、図による外れ値の検出はほとんど不可能に近いことも分かる。

図 5.3a：logturnover と logcost

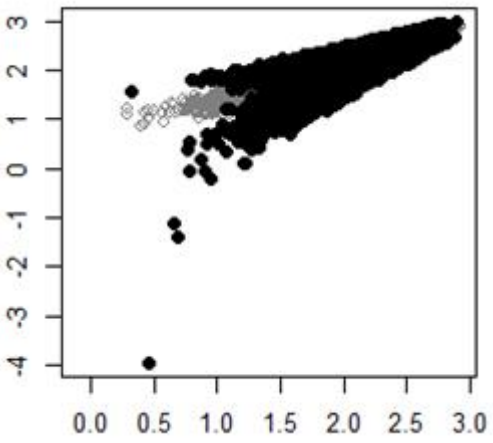


図 5.3b：logturnover と logworker

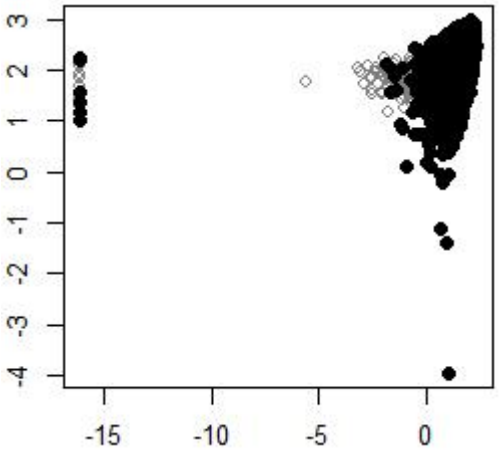
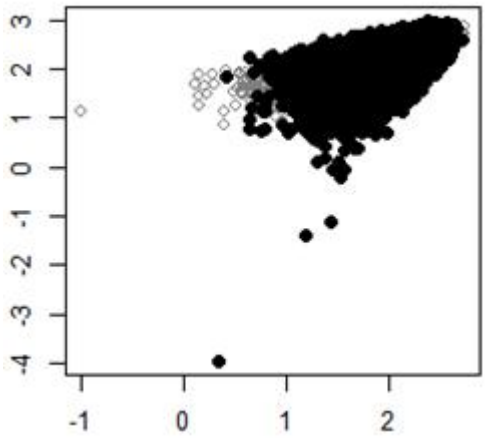


図 5.3c：logturnover と logcapital



5.5 図による影響力のある外れ値検出法との比較

4 節と同様に、以下の図では、通常の値を白丸、「影響力のある外れ値」を**菱形**で図示する。図 5.4a は logturnover の箱ひげ図であり、図 5.4b は logcost の箱ひげ図であり、図 5.4c は logworker の箱ひげ図であり、図 5.4d は logcapital 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 5.4a

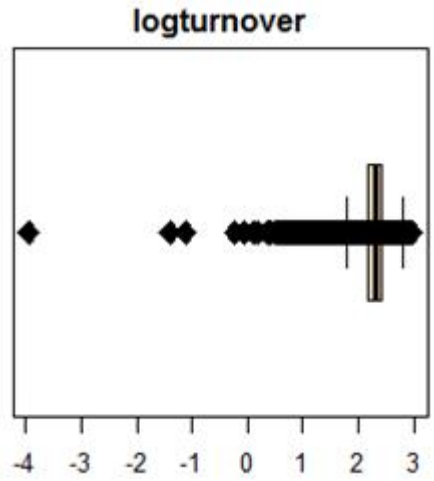


図 5.4b

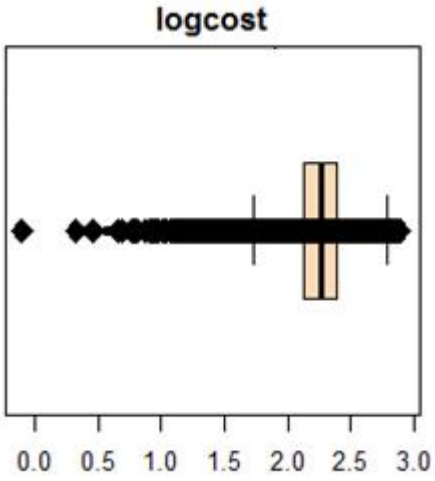


図 5.4c

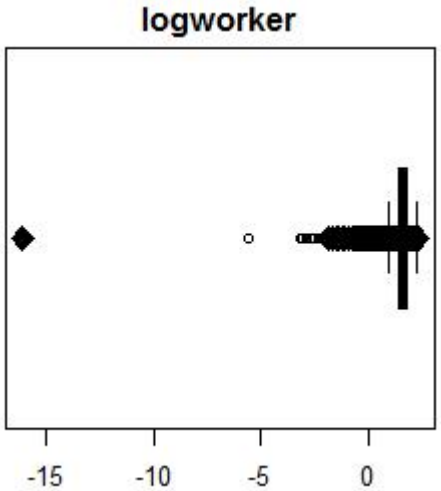


図 5.4d

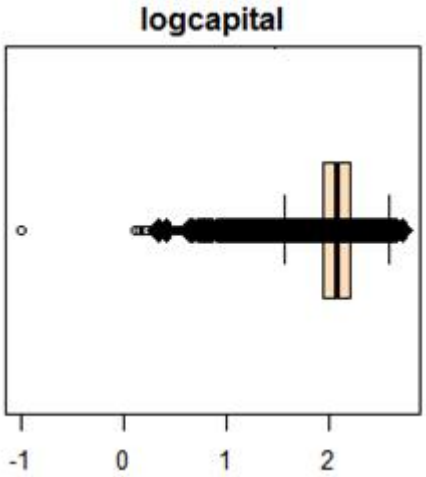


図 5.5a は、logturnover（縦軸）と logcost（横軸）の散布図であり、図 5.5b は、logturnover（縦軸）と logworker（横軸）の散布図であり、図 5.5c は、logturnover（縦軸）と logcapital（横軸）の散布図である。図 5.3 と同様に、観測数が 100 万ともなると、図による検出はほとんど不可能に近いことが分かる。

図 5.5a：logturnover と logcost

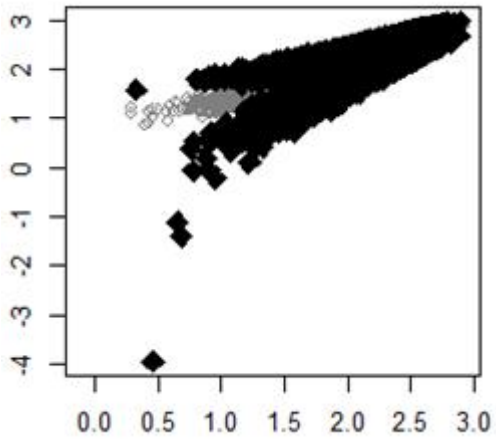


図 5.5b：logturnover と logworker

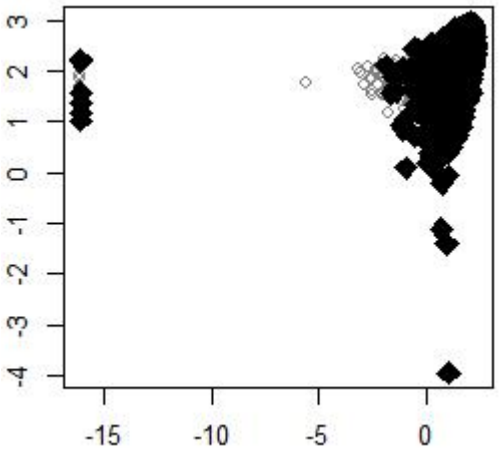
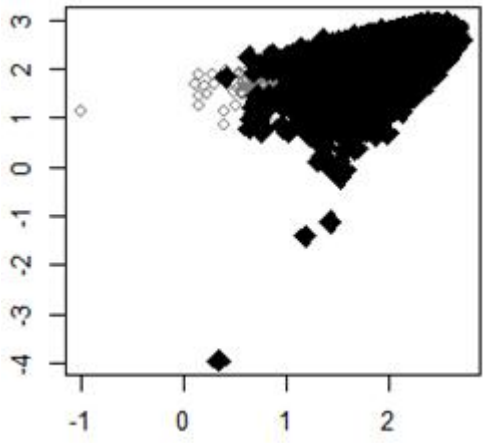


図 5.5c：logturnover と logcapital



6 結語と将来の課題

2012 年 9 月にノルウェーのオスロで開催された UNECE ワークセッションにおいて示されたとおり、各国においてデータエディティングは重要視されており、予算の削減という国際的な流れの中で、選択的エディティングへの注目は、年々、高まってきている。

本稿では、混淆正規分布モデルに基づく選択的エディティングプログラムである *SeleMix* の検証を行った。EDINET を用いた検証により、事業所・企業の経理項目におけるエラーを高い精度で検出できることが分かった。また、選択的エディティングにより、効率的にエラーに対処できることも分かった。シミュレーションデータを用いた検証により、100 万の観測数を持つ巨大データセットにも対応でき、経済センサスへの応用可能性が高いことも分かった。

センサーや IT 技術の発展により、非常に大規模な生データが生産されるビッグデータの時代が到来している。しかし、ビッグデータには、欠測値や外れ値が含まれており、データの使用目的に応じた前処理やクレンジングをどのように行うかといったノウハウがますます重要なものとなってくる（丸山, 2013, p.5）。今回の検証により、*SeleMix* は、巨大データの外れ値検出を行うことができることが分かり、ビッグデータの時代におけるデータエディティング手法としても有用ではないかと期待される。

閾値の設定によって、検出力が変ることから、閾値設定の基準と現実的な値を探索することは将来の課題と言えよう。また、経済センサス - 活動調査の実データを用いた外れ値検出及び選択的エディティングの検証も行いたいと考えている。他にも、検証点として、他の多変量外れ値検出法との比較や破綻点といった頑健性(ロバストネス)に関する検証なども行いたいと考えている(Rousseeuw and Leroy, 2003; Andersen, 2008)。最後に、2014 年にはフランスのパリにて、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が開催される見込みであり、最新の動向に引き続き注視していきたい。

参考文献（英語）

[1] Andersen, Robert. (2008). *Modern Methods for Robust Regression*. Thousand Oaks, CA: Sage Publications.

[2] Barnett, Vic, and Toby Lewis. (1994). *Outliers in Statistical Data*, Third Edition. Chichester: John Wiley & Sons.

[3] 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.

[4] 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.

[5] DeGroot, Morris H. and Mark J. Schervish. (2002). *Probability and Statistics*. Boston: Addison-Wesley.

[6] de Waal, Ton, Jeroen Pannekoek, and Sander Scholtus. (2011). *Handbook of Statistical Data Editing and Imputation*. Hoboken, NJ: John Wiley & Sons.

[7] Di Zio, Marco and Ugo Guarnera. (2012). "Selective Editing as a Part of the Estimation Procedure," *Work Session on Statistical Data Editing, Conference of European Statisticians*, Oslo, Norway, 24-26 September 2012.

[8] Fox, John. (1991). *Regression Diagnostics*. Newburry Park, CA: Sage Publications.

[9] Greene, William H. (2003). *Econometric Analysis*, Fifth Edition. New Delhi: Pearson Education, Inc.

[10] Guarnera, Ugo and M. Teresa Buglielli. (2013). "Selective Editing via Mixture Models," http://cran.r-project.org/web/packages/SeleMix/SeleMix.pdf. Accessed on July 11, 2013.

[11] Guarnera, Ugo, Orietta Luzi, Francesca Silvestri, M. Teresa Buglielli, Alessandra Nurra, and Giampiero Siesto. (2012). "Multivariate Selective Editing via Mixture Models: First Applications to Italian Structural Business Surveys," *Work Session on Statistical Data Editing, United Nations Economic Commission for Europe*, Oslo, Norway, 24-26 September 2012.

[12] Gujarati, Damodar N. (2003). *Basic Econometrics*, Fourth Edition. New York: McGraw-Hill.

[13] 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*.

[14] 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.

[15] Nicolaas, Gerry. (2011). "Survey Paradata: A Review," *ESRC National Centre for Research Methods Review Paper* no.17. National Centre for Social Research (NatCen).

http://eprints.ncrm.ac.uk/1719/1/Nicolaas_review_paper_jan11.pdf. Accessed on July 11, 2013.

[16] Nordbotten, Svein. (1955). “Measuring the Error of Editing Questionnaires in a Census,” *American Statistical Association Journal* vol.55: pp.364-369.

[17] OECD. (2007). *The OECD Glossary of Statistical Terms*. http://stats.oecd.org/glossary/. Accessed on July 11, 2013.

[18] Rousseeuw, Peter J. and Annick M. Leroy. (2003). *Robust Regression and Outlier Detection*. Hoboken, New Jersey: John Wiley & Sons, Inc.

[19] Scarrott, Carl. (2007). “Feasibility Study: A Review of Selective Editing,” *Official Statistics Research Series*, University of Canterbury and Statistics New Zealand.

[20] Trochim, William M. K. (2006). *Research Methods Knowledge Base*. http://www.socialresearchmethods.net/kb/measerr.php. Accessed on July 11, 2013.

[21] UNECE. (2000). *Glossary of Terms on Statistical Data Editing*. New York and Geneva: United Nations Publication.

[22] Weiss, Neil A. (2005). *Introductory Statistics*, Seventh Edition. Boston: Pearson.

参考文献（日本語）

[23] 金融庁. (2012). EDINET 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. http://info.edinet-fsa.go.jp/. 2013 年 7 月 11 日アクセス.

[24] 高橋将宜. (2012). 「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」, 『製表技術参考資料 17』, 独立行政法人統計センター.

[25] 高橋将宜, 伊藤孝之. (2013). 「経済調査における売上高の欠測値補定方法について～多重代入法による精度の評価～」, 『統計研究彙報』第 70 号 no.2, 総務省統計研修所, pp.19-86.

[26] 丸山宏. (2013). 「データに基づく意思決定」, *ESTRELA* no.231: pp.2-7.

[27] 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.

付録 1：2012 年 UNECE ワークセッション報告論文概要

本付録では、2012年9月の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にて報告された全論文を日本語で簡潔に要約して紹介している。実際の全論文（英語）は、UNECEのウェブサイト[21]にて閲覧及びダウンロードが可能である。以下、WPはワーキングペーパー(Working Paper)の番号を表している。その後に英文タイトルを掲載し、括弧の中に著者名と国名を記し、その下に要旨を掲載している[22]。

(0)題目

WP.1 Provisional Agenda and Tentative Timetable (UNECE)

ワーキングペーパー1番は、報告論文ではなく、ワークセッションのタイムテーブルである。本ワークセッションは、ノルウェーのオスロコングレスセンターにおいて、2012年9月24日（月）の午前9時に開幕し、9月26日（水）の午後3時に閉幕した。討議された事項は、以下の7つのトピックであった：(1) 選択的及びマクロエディティング（7論文）；(2) エディティングのグローバルな解決策（7論文）；(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と補定（10論文）；(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（4論文）；(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（7論文）；(6) 新たな手法（4論文）；(7) センサスデータのエディティング及び補定（6論文）。報告された論文の数は44（WP.2～WP.45）であった。

(1)選択的及びマクロエディティング

WP.2 Selective Editing as a Combinatorial Optimization Problem: A General Overview（Ignacio Arbués, Pedro Revilla, and David Salgado, スペイン）

スペインの報告では、選択的エディティングに関する2つの汎用的な原則を提案し、ユニット選択が解決策となるような最適化問題を扱った。この問題の核は、選択されるべきユニットの数を、エディットの対象とならないユニットの測定誤差の二乗平均(Mean Squared Error)の範囲内に制限し、最小化することである。測定誤差をモデル化することによる汎用的な枠組は、観測値-予測値モデル(Observation-Prediction Model)と名づけられ、質的変数や準連続変数などの多用な変数への応用可能性があり、この手法を使用したRパッケージ及びSASマクロを開発中である。

---

[21] http://www.unece.org/stats/documents/2012.09.sde.html（2013 年 7 月 11 日アクセス）

[22] 論文の引用には、下記のフォーマットの使用を推奨する。著者名. (2012). “タイトル,” *Work Session on Statistical Data Editing, United Nations Economic Commission for Europe*, Oslo, Norway, 24-26 September 2012.

WP.3 Multivariate Selective Editing via Mixture Models: First Applications to Italian Structural Business Surveys（Ugo Guarnera, Orietta Luzi, Francesca Silvestri, M. Teresa Buglielli, Alessandra Nurra, and Giampiero Siesto, イタリア）

イタリアの報告では、*SeleMix*（セレミックス：混合モデルによる選択的エディティングソフトウェア）による多変量外れ値検出法の報告が行われた。*SeleMix*では、連続変数における影響力の強いエラーを検出する。本稿では、外部情報からの補助的な情報（行政データや統計資料）を用いることで、どの程度の費用が軽減できるかを検証することを目的としている。この目的のために、ICT（Information and Communication Technology：情報伝達技術調査）及びSME（Small and Medium Enterprises：中小企業調査）といった構造的企業調査に応用し、対象とした変数は売上高と売上原価である。いずれの調査においても、応用結果は良好なものであったが、SMEに関しては、公表領域が複雑であるため、さらなる分析を必要とする。ICTに関しては、E&Iプロセス（エディティング及び補定プロセス）への統合がすでに進行中である。

WP.4 An application of Selective Editing to the U.S. Census Bureau Trade Data（Maria Garcia, 米国）

米国センサス局は、対外貿易データへの選択的エディティングのスコア関数を適用し、実現可能性の検討を行い、擬似バイアスの評価方法を報告した。伝統的に、選択的エディティングでは、前期データに基づいてスコア関数を作成していたが、貿易データでは期ごとの変動が大きいため、この手法を対外貿易データに応用することはできない。変数の予測値を推定することによって、今期のデータのみで対応できるように、既存のスコア関数を改良した。各々の観測値にスコアを割当、レコードをランキング化する。このランキングは、レコードのエラー可能性とその影響力に基づいている。さらに、この手法を開発した統計家と実際に使用した専門官との間でフィードバックのやり取りも行っている。

WP.5 Tree Analysis – A Method for Constructing Edit Groups（Anders Norberg, スウェーデン）

スウェーデン統計局は、分類回帰樹木(CART: Classification and Regression Trees)と呼ばれる木解析手法の報告を行った。CARTは、被説明変数がカテゴリカルな場合には分類を行い、数値変数の場合には回帰樹木を生成するノンパラメトリックな手法である。木解析とは、エディットグループを構築し、ソフトエディットを用いて疑わしい値を検出する手法である。すなわち、巨大なデータセットを「幹(Branch)」と呼ばれる別々のグループと「葉(Leaves)」と呼ばれる最終グループに分割し、どの変数がどの「葉」に属しているかに基づいて、新たな観測値の値を予測する。この手法は、エディットグループ

の形成に大いに役立つが、これまで、国家統計局によって使用されてきた例はほとんどない。この種の分析を行うことのできるソフトウェアは多くあるが、エディットグループの形成に完全に適合しているものがないからである。

WP.6 An Automated Comparison of Statistics（Elmar Wein, ドイツ）

ドイツ連邦統計局は、自動比較に関する報告を行った。自動比較とは、今期の実測値と妥当な参照値（たとえば前期の値）の間に疑わしい差異が生じた場合に、フラグを立てるものである。理論上、分布の中心や分散といった変数の特定の情報だけではなく構造的な差異を取り除くことができるため、自動比較は、伝統的な人手審査よりも強力である。ドイツ連邦統計局では、SASにおいて、主成分分析を利用した試作版の自動比較プログラムを開発した。自動比較を行うための要件は、レコードの識別子が存在し、同一の数値変数を含む2つのデータセットがあればよく、比較的、緩やかなものである。したがって、自動比較は、汎用的ツールとして使用できる可能性が高く、少なくとも、すべての構造的企業統計に応用できると期待されている。予備的調査の結果によると、これから改良を施していくことにより、構造的企業統計のための有用なツールになると期待される。

WP.7 The Use of Evaluation Data Sets When Implementing Selective Editing（Katrin Lindgren, スウェーデン）

スウェーデンによる報告では、選択的エディティングの汎用ツールSELEKT（セレクト）を用いた閾値設定に関する実装の課題を取り上げている。選択的エディティングでは、各々のユニットのグローバルスコアによって、そのユニットが人手審査に回されるべきどうかを決定する。グローバルスコアは、そのユニットの重要な観測値、疑わしさの度合い、そして公表統計値にどのような影響を及ぼすかによって算出する。影響の度合いは、観測値と予測値の差、そして標本誤差に関連してユニットに割り当てられるデザインウェイトによって推定する。グローバルスコアの算出には、出力における各々の変数の相対的な重要度を含めることもできる。あるユニットのグローバルスコアが高ければ、人手審査用のリストの上位に位置し、実際にエディティングを行う際に、リスト上のユニットを優先化する目的でも使用される。あらかじめ規定された閾値以上のグローバルスコアを持つすべてのユニットは、必ず人手審査に回される。閾値は、統計出力の擬似バイアスを分析することによって設定する。最終的に、選択的エディティングの目的は、擬似バイアスを許容範囲内にしつつ、ミクロエディティングを最小限に抑えることである。

WP.8 Selective Editing as a Part of the Estimation Procedure（Marco di Zio and Ugo Guarnera, イタリア）

イタリアは、推定過程としての選択的エディティングについての報告を行った。選択的エディティングを推定過程とみなす手法の1つとして、二段階手法があり、そこでは、測定誤差が最終推計値に与える影響を減らすことを目的としている。また、モデルベースの手法では、エラーは混淆正規分布モデルにしたがって対数正規データに影響を与えるとされる。イタリアの報告では、*SeleMix*を用い、これら2つの手法を同時に利用した。つまり、混淆正規分布モデルによって算出された期待誤差(Expected Error)の値に応じて、標本デザインを作成し、バイアスを除去するために二段階の標本を抽出した。標本の期待誤差を利用することで、最終推計値からバイアスをより効率的に取り除くことができる。2008年の中小企業調査を利用して評価を行った結果、少数の大規模誤差に関しては、*SeleMix*による選択的エディティングの方が、二段階手法よりもパフォーマンスがよく、多数の小規模誤差に関しては、十分な数のユニットが抽出されさえすれば、バイアス補正により推定値を改善できることが分かった。

⑵エディティングに関するグローバルな解決策

WP.9 Review of the UNECE Glossary of Terms on Statistical Data Editing（Felibel Zabala, Soon Song, Emma Bentley, Val Cox, Catherine Cumpstone, Jane Xu, Joe Luo, Temaleti Tupou, Amanda Hughes, and Anna Lin, ニュージーランド）

ニュージーランド統計局は、『統計データエディティングに関する用語集』(*Glossary of Terms on Statistical Data Editing*)の改訂に向けた検討に関する報告を行った。この用語集には、現在、概念の定義、手法や技術、コンピュータシステムといった統計データエディティングに関する200以上の用語が収録されており、1990年代以来の幾年にもわたる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の参加者間で行われた共同作業の賜物である。2009年のワークセッションにおいて、エディティングの現状及び将来の状況に対応するために、用語集を改訂する必要性があるとの結論にいたった。そこで、当時、エディティング及び補定に関する方法論的基準やガイドラインを作成中であったニュージーランド統計局は、用語集の検討を請け負った。本稿では、用語集の改訂案を提示し、新たな概念の追加、既存の概念の修正や削除などに関し、今回のワークセッションの参加者からの意見を募った。

WP.10 On the General Flow of Editing（Jeroen Pannekoek and Li-Chun Zhang, オランダ・ノルウェー）

オランダ統計局とノルウェー統計局の共同研究では、最新のエディティング理論と実践を考慮に入れ、一般的なエディティング業務の流れに関する報告を行った。この業務の

流れは、全体的なエディティングプロセスをいくつかのプロセスに分割することから構成されている。各々のプロセスには、それぞれの目的に応じた汎用エディティング機能が最大で3種類あり、それぞれに統計関数が割り当てられている。統計関数の観点から各々のプロセス内でどのような活動が行われるのかを記述し、結果として、全体的なプロセスが、効率性、正確性、時宜性といった品質基準を満たしていることを示す。こういった一般的なエディティング作業の流れを、オランダとノルウェーにおける構造的企業統計の例を用いて詳論している。

WP.11 Update on the Development of the Generic Statistical Information Model (GSIM)（Thérèse Lalor and Steven Vale, オーストラリア・UNECE）

公的統計は、一般的にどの国においても、ストーブパイプモデルによって作成されている。ストーブパイプモデルとは、各々の個別領域の統計値が独立して作成されるプロセスのことである。このように統合されていないプロセスでは、効率性が低下し、共通のツールや手法の開発を行うことも困難である。しかし、一般的に、世界中のどこの統計機関においても、概ね同一の情報を作成し消費している。たとえば、すべての統計機関は、分類を行い、データセットを作成し、結果を公表する。このように、各国統計機関の使用する情報は、根本的には同じであるにもかかわらず、それぞれ微妙に異なるやり方で記述されてきた（また、同一の機関内でさえも、そういったことが起きることがある）。そこで、公的統計の近代化を支援するために、基準を設ける必要があり、汎用統計情報モデル(GSIM＝ジーシム)を開発するにいたった。GSIMは、公的統計の近代化にとって礎となるものであり、2011年に欧州統計化会議によって承認された。GSIMは、統計作成における情報オブジェクト及び情報の流れを記述するモデルである。これまでのところ、150以上の情報オブジェクトが識別されてきた。GSIM version 0.8は、ワークセッション終了後、数日以内に公開され、2012年12月にversion 1.0が公開された。

WP.12 Two Paradigms for Official Statistics Production（Boris Lorenc, Jakob Engdahl and Klas Blomqvist, スウェーデン）

スウェーデンは、外部世界についてのデータ及び知識にかかわる公的統計作成の2つのパラダイムについて報告を行った。第1パラダイムは、ストーブパイプモデルに関連するものであり、ここでは、エディティングは特定の目的をもって行われる。第2パラダイムは、汎用目的の自動エディティングに関連するものである。本稿では、知識システム（認知システム）という観点から、統計作成の近代化システムについて論じた。プロセス及び情報に関してモデルを定義することによって統計作成システムのモデル化がどのように進んでいくかということは、認知科学や人口知能によって認知システムがどのように概念化されたかということになぞらえることができる。本稿において、知識システム（認知システム）とは、専門分野の事実に関する情報の蓄積量のことを意味する。統計作成

はユーザーのニーズを考慮に入れるべきであり、知識システムや巨大データベースといったいわゆるビッグデータなどに対処する近代的なアプローチがどのように統計作成に関連しているかを示す。

WP.13 Proposal of a Revised Approach for Data Validation within the European Statistical System（Michel Henrard, 欧州統計局）

欧州統計局は、欧州統計システム(ESS: European Statistical System)内におけるデータ妥当性検証手法を改良する提案を行った。ESSにおける統計作成は、加盟国と欧州統計局の間で共有されている。データの収集、処理、予備的な製表などは加盟国によって行われる。その後、所定の様式にしたがって、データは欧州統計局に伝送される。ESS加盟国からデータを収集する際に、欧州統計局では、公表や分析を行う前に、受け取ったデータの審査と検証を行う。このプロセスには、多大な費用と労力がかかる。2010年の末より、欧州統計局では、妥当性検証に関してビジョン・インフラ・プロジェクト(Vision Infrastructure Project)を行っている。このプロジェクトは、妥当性検証プロセスを改善することによって、加盟国から欧州統計局への統計作成プロセスにおける効率性の向上を目指すものである。このプロジェクトでは、効率性の向上を達成するために、以下の手法を用いている：妥当性ソリューションの実行；妥当性検証の役割分担；政策決定とガイドラインの作成。

WP.14 The Development of a Data Editing and Imputation Tool Set（Claude Poirier, カナダ）

費用の効率化を達成するためには、ローカルなニーズではなく、グローバルなニーズからの視点に立つ必要がある。データエディティング及び補定の文脈では、様々なデータの情報源に対応できるロバスト（頑健）な手法が必要である。本稿では、調査により収集したデータと行政データの双方に共通の要件を提示する。グローバルな視点から客観性を達成するために、ツールセットの形成のたたき台として、実務上のデフォルトとして使用できるような手法を提案する。ツールの望ましい性質としては、以下のものが挙げられる：機能性、妥当性、利用可能性、解釈可能性、一貫性、正確性、時宜性、順応性、信頼性、保全性、相互運用性。提案された基礎的なツールは、以下のものである：BANFF（バンフ）、CANCEIS（キャンサイス）、SELEKT（セレクト）。機能性の溝を埋める目的で、他のツールも、今後、検討する予定である。

WP.15 On Tap: Developments in Statistical Data Editing at Statistics New Zealand（Allyson Seyb, Felipa Zabala, Les Cochran, and Chris Seymour, ニュージーランド）

ニュージーランド統計局は、2011年の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における招待論文をフォローアップし、ニュージーランド統計局における

経済・世帯調査の処理基盤の最新状況について報告をした。ニュージーランド統計局では、費用対効果の高い、持続可能な方法で、目的に合致した統計の作成を目指している。この目的を達成するために、統計2020プログラム(Stats 2020 Programme)を通じて、近代的な統計作成システムの構築と実装を目指している。これまでの課題と教訓は以下のとおりである：汎用的なニーズと個別のニーズのバランスを取ること；共通サービスとして実装されるにふさわしい処理要素を決めること；受身的な文化からより能動的な文化へ変容すること；IT関係のプロジェクトに関し、敏しょうなプロジェクト遂行を採用すること。将来の課題としては以下のものが挙げられる：汎用処理モデルの検証；最新の動向に基づく利点に関する検証；データ収集プロセスの改善。

(3)複数情報源と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と補定

WP.16 Micro Integration of Register-Based Census Data for Dwelling and Household（Li-Chung Yhang and Coen Hendriks, ノルウェー）

ノルウェーの2011年センサスは、他の欧州各国と同様に、レジスターベースで行われた。住居と世帯に関するデータを様々な行政情報源から取得した。中でも、中央人口レジスター(CPR: Central Population Register)と建物レジスター(GAB: Ground Parcel, Address and Building Register)は、最も重要な情報源である。既存の世帯統計は主にCPRに基づいて作成されており、住居統計はGABに基づいている。世帯と住居という明示的な結びつきをもってミクロレベルで2つの情報源を統合することは、詳細なセンサス統計情報を作成するために重要である。しかし、GAB及びCPRには様々なエラーが存在する。したがって、住居と世帯を結びつけた完全なセンサスデータを作成するためには、ミクロレベルでの統合を行う統計手法が必要となる。ダブル最近隣補定手法は、直接的なマッチングが不可能な様々な種類のユニット間のミクロリンクの問題に対する解決策となる。

WP.17 All the Answers? Statistics New Zealand's Integrated Data Infrastructure（Felibel Zabala, Rodney Jer, Jamas Enright, and Allyson Seyb, ニュージーランド）

2002年以来、ニュージーランド統計局では、様々な政府機関から提供されるデータの統合に取り組んでいる。こういったデータセットの結びつけに成功したプロジェクトも複数あり、ここでは、各々のデータセットは特定の質問票に対応している。しかし、各々のデータセットが作成され、保存されている環境が異なっているために、これらのデータセットを統合する際には様々な困難が伴う。ニュージーランド統計局では、こういったデータセットを1つの環境に統合するために、統合データインフラ(IDI: Integrated Data Infrastructure)を開発している。IDI環境におけるデータは、様々な情報源から提供されている。したがって、各々のデータセットの品質も様々であり、こういったデー

タセットを結びつけることには困難が伴うため、効果的で効率的なエディティング及び補定の戦略が不可欠となる。IDIでは、異なる情報源から得られた同じユニットのデータに関する矛盾点やレコード内のエラー及び欠測値に対処し、レコード間の変数の一貫性を保証している。統合データインフラをさらに拡張し、標準的な品質指標の開発を計画している。

WP.18 Editing Challenges for New Data Collection Methods（Rachel Skentelbery and Carys Davies, 英国）

英国国家統計局では、現在、データ収集方法の改善を目指し、オンライン電子調査票(eQuestionnaires)などの様々な収集法を利用している。しかし、こういった異なる収集法を用いることには、エディティングに関して多くの課題がある。そこで、英国国家統計局では、電子データ収集プログラム(EDC: Electronic Data Collection Programme)の開発を進めており、ここでは、データの収集から統合や分析まで、システム、手法、プロセスを改善することを目的としている。二段階手法が提唱されており、第1段階では、核となる枠組みと機能を構築する。次の段階では、複雑な電子調査票に対応するために、手法の開発、システムとインターフェースの改善を目指す。本稿では、電子調査票におけるエディットを検証するために使用した実験計画について報告している。

WP.19 Methodological Questions Raised by the Combined Use of Administrative and Survey Data for French Structural Business Statistics（Philippe Brion, フランス）

ESANE（イザーン：**É**laboration des **S**tatistiques **AN**nuelles d'**E**ntreprises、年次企業統計の精密化）は、複数情報源を利用して構造的企業統計を作成するための新システムである。このシステムは、行政情報データと統計調査データを統合して使用する。行政情報は、主に、金融データ、税務データ、社会保障データなどである。統計調査データは、主に、行政情報に含まれていない情報を補う目的で、一部の企業を標本として抽出して行う。フィールド定義の問題や複数情報源を使用することによる推定量への影響に関して報告を行った。また、ビジネスレジスターの品質に関して議論をし、複合指標に関する問題を検討した。

WP.20 Studying the Options of Substituting a Regular Statistical Survey with Administrative Data（Gergely Horváth and Zoltán Csereháti, ハンガリー）

ハンガリーは、保健統計の分野において、調査データを行政データに置き換えるための方法論について報告を行った。調査データを2次データに置き換えた場合の影響を検証するために、本稿では、小規模かつシンプルな年次調査を対象として選んだ。結論としては、調査データを行政データに置き換える手法がどのようなものであったとしても、決して簡単なことではなく、一筋縄でいくものではない。今回の実験では、データの情報

源は同一であったにもかかわらず、行政データを用いることにより、データ提供者の事務的な負担を軽減することができ、さらにデータの品質を改善することができた。したがって、より正確なデータをより早く提供することができるだろう。しかし、同時に、バイアスが増えるリスクも常に存在することも分かった。

WP.21 Evolving Data Processing in the Statistics Centre（Dragica Sarich and Maitha Al Junaibi, アブダビ）

公的統計機関では、一般的に、データ品質を改善するために、データエディティングなどの統計手法を用いている。こういった統計手法を利用することには、公的統計機関にとって、いくつかの利点がある。たとえば、欠測値のない「完全」データを提供したり、高度な統計分析を行ったりすることなどが挙げられる。アブダビでは、混合モードのデータ収集法及び自動エラーデータ検出法を用いた経済調査を初めて行った。本稿では、データエディティングを行うメリットとデメリットを議論し、データプロセスの一環として自動エラーデータ検出法を用いた経験談について報告を行う。混合データの品質を改善するために、自動データエディティングといった新しい統計手法を用いた結果、調査の品質及び効率性を高められることが分かった。また、実務において直面した問題を克服するための戦略についても報告を行う。事業所調査における欠測値及び特異値に対処する方法も検討中である。

WP.22 Editing and Imputing VAT Data for the Purpose of Producing Mixed-Source Turnover Estimates（Daniel Lewis and Hannah Finselbach, 英国）

英国国家統計局では、付加価値税売上高データと標本調査を混合情報源として利用し、月次企業調査の推定値を算出する手法を開発している。付加価値税データが英国国家統計局に届けられる段階では、税務目的のクリーニングのみ行われた状態で、統計目的のクリーニングは、まだ行われていない状態となっている。したがって、月次企業調査における推定値の品質を保証するためには、付加価値税売上高データにおけるエラーに対処しておかなければならない。行政データにおけるエラーを検出し訂正するためにどのような手法を用いるべきか、それは、特にカバー率、時宜性、精度という点で、混合情報源の統計値を算出するために使用するデータに依拠するであろう。

WP.23 Imputing Missing Values When Using Administrative Data for Short-Term Enterprise Statistics（Pieter Vlag, オランダ）

売上高の月次推定値や四半期推定値を提供する目的では、行政データは不完全であることが多い。こういった行政データの不完全性は、回答が遅れているといった時間の問題のこともあれば、ある種のユニットは特定の時期のみ調査の対象となっているなどの構造的な問題のこともある。こういった推定に関する問題は、非常にありふれたことであ

るため、欧州各国統計機関と共同でプロジェクトを開始した。このプロジェクトでは、まず第1段階として、不完全な行政データを審査し補定する現行の手法について調査を行った。第2段階として、現在、最も有力と考えられる手法を試験的に導入し、比較を行った。イタリア、ドイツ、フィンランドにおける分析では、短期企業統計の行政データ推定値に関する主な訂正は、母集団におけるユニットに関する不確実性によって引き起こされていることが分かった。したがって、実際の母集団を推定するために最も適切な推定法を検証する必要がある。

WP.24 Improvement of the Timeliness of the Italian Business Register via Imputation of Missing Data（Davide Di Cecco and Danila Filipponi, イタリア）

イタリアのビジネスレジスターでは、行政情報及び統計調査情報を統合することによって、すべての産業において現在活動中の企業に関して、その構造的な特徴を記録している。現在のところ、参照年 t の年のビジネスレジスターを構築するプロセスは、以下のとおりである：主要な情報源から年次データが提供され始める$t + 1$の年（つまり参照年の翌年）の第4四半期からプロセスを開始する。正常化(Normalization)と標準化(Standardization)のプロセスを終えた後、データを統合し、各々の統合したユニットに関して、主だった構造的な変数と識別変数（id変数）を推定する。本報告の目的は、$t+1$の年の第1四半期において利用可能な行政情報及び統計情報のみを利用することで、参照年の6か月後には企業人口の構造に関する情報を提供し、ビジネスレジスターを改善することである。明らかに、時宜性に関する改善と情報の正確さは、反比例の関係にある。本稿では、ビジネスレジスターの早期公表に向けて、欠測データの検出と補定方法に関して検討した結果、欠測情報を補定するのに十分なメカニズムがあれば、最終データの代わりに暫定データを用いることができることが分かった。使用した手法を検証するために、2種類の行政データを用意して比較した結果、精度の高さが確認できた。

(4)メタデータ及びパラデータを使用したエディティングプロセスの効率性分析

WP.25 Assessment and Improvement of the Selective Editing Process in Esane (French SBS)（Emmanuel Gros, フランス）

フランス国立統計経済研究所(INSEE)では、2009年より、構造的企業統計を作成する新システムとしてESANEを用いてきた。この新システムでは、選択的エディティングを採用することで、データエディティングプロセスの改善を図っている。しかし、過去3年間の経験より、ESANEに実装されている選択的エディティングにはいくつかの欠点が存在することが分かってきた。たとえば、ローカルスコアの不安定さ、ローカルスコアを統合するグローバルスコアに関する問題、特定の種類の変数（正の値を取らない変数など）に関する問題である。過去3年の経験に基づくメタデータにより、代替案を検証し、これ

らの諸問題を解決するための方法論的な改善策を実践している。改善策の例として、警告メッセージに関する情報、疑わしいユニットの数、エディティング担当職員からのフィードバックなどが挙げられる。また、メタデータ及びパラデータの情報は、調査票の内容や構成を変更するなど、他の手法の改善にも使用できる。国民経済計算の担当職員など、データユーザーにプロセスを説明する際にも有用である。

WP.26 Outlining a Process Model for Editing with Quality Indicators（Pauli Ollila, Outi Ahti-Miettinen, and Saara Oinonen, フィンランド）

フィンランドによる報告では、統計データエディティングのためのプロセスモデルを紹介した。公的統計のエディティングプロセスに関する指標を収集し、これらの指標を以下の3つの機能別に分類した：生データに関する指標；エラーの検出に関する指標；エラーの訂正に関する指標。また、これらは、エディティングモデルの3つの段階にそれぞれ対応している。本稿において紹介した指標の数は、統計作成プロセスの多様性を鑑み、非常に多いものとなっている。しかしながら、すべての指標が、必ずしもすべての種類の統計に適しているわけではない。したがって、各々のプロセスにおいて、どの指標を適用するのかについて考慮することは非常に重要である。指標を算出するために用いる部分集団や変数を選定するには、確固たる専門知識を必要とする。本稿で示した指標の中には、重要ではあるが特定の状況においてのみ有用であるものも含まれている。よって、すべての統計値に関して、詳細な指標を定義し公表することは可能ではないが、エディティングプロセスに関する標準的な指標は、常に算出するべきである。たとえば、データの欠測に関する指標は、統計作成プロセスの各々の段階におけるカバー率を示すもので、重要なツールである。何らかのエディティングが行われた場合には、データユーザーにとって、データのエディット率に関する情報が分かるようになっているべきである。

WP.27 An Embedded Experiment to Test Non-Response Follow-Up Strategies when Using Electronic Questionnaires（Jeannine Claveau and Claude Turmelle, カナダ）

2010年より、カナダ統計局では、企業調査の主要な情報収集法として、電子調査票を導入し始めた。電子調査票を採用するに際して、非回答に対する様々な対応法を評価する目的で、電子調査票を用いて行われている7つの調査について実験を行った。収集に関して、年次調査における収集の最初の数か月間においては、電話照会を行わなくても、多数の回答が期待できる。加えて、督促状をメールで頻繁に送ることによって、回答率を増加させることができる。収集段階の初期において、2週間ごとにメールによる督促状を送ることで、わずかな費用で45%の回答率増加につながった。一方、ある時点以降は、電話による照会を始めなければならないが、このプロセスを遅くから始めても、調査の最終的な回答率に変化は見られないことが分かった。

WP.28 Editing Staff Debriefings at Statistics Sweden（Jörgen Svensson, スウェーデン）

エディティング担当職員(Editing Staff)へのデブリーフィング(Debriefing：意見交換会)は、特定の調査のデータエディティングに関わっている職員が一堂に会して経験談を報告し、議論しあう質的な調査方法である。デブリーフィングとは、フォーカスグループ(Focus Group)のようなものであり、その主な目的は、調査票の質問にどのような問題点があるかを探り、エラーの原因を探ることである。エディティング担当職員は、問題やエラーがどうして起こったのかに関して、有用な考えを持っていることが多い。デブリーフィングでは、調査の回答者の間で繰り返されている反応や問題などの洗い出しを行う目的も持っている。また、デブリーフィングでは、データ収集やエディティングに関するパラデータや記録簿により得られた情報を補いながら、わずかな費用で大量の情報を入手できる。スウェーデン統計局では、過去5年間にわたって、エディティング担当職員へのデブリーフィングを実施してきた。

(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

WP.29 TEA for Survey Processing（Ben Klemens, 米国）

TEAは、生データからエディティング、補定、出力結果の公表まで、人口調査の処理を統一化するための汎用システムである。TEAでは、エディティングや補定の手順を独立させつつ、単一の基盤で行える。TEAは、非公式だがRのパッケージとして利用可能であり、Rの基盤によって様々な視覚化の技術も利用できる。TEAは、アメリカンコミュニティサーベイ及び2010年センサスにおける集団居住住宅データの処理に使用されてきた。ユーザーは、エディット規則、補定モデル、開示抑制法など、各々の調査の仕様書を入力して使用する。TEAでは、まず、エディット規則を満たさないフィールドを検出し、それらをブランク（空白）にする。開示抑制のステップでは、ユーザーの指定どおりに基本的な集計を行い、集計を行うことにより識別されてしまうようなレコードをブランクにする。補定では、ホットデック法やOLSなど様々な手法を多重代入法の枠組みで行うことができる。さらに、補定値は、エディット規則に照らして審査されるので、すべての補定値はエディット規則を満たすことが保証される。調査データに対して補定モデルが構築されると、そのモデルを用いて完全合成データを作成し、ユニット欠測のデータの補定を行う。

WP.30 Innovative Visual Tools for Data Editing（Martijn Tennekes, Edwin de Jonge, and Piet Daas, オランダ）

公的統計において、データ品質を調べるために最もよく使用されている手法は、表、棒グラフ、ヒストグラム、散布図といった視覚化ツールである。しかし、これらの手法で

は、表示できる観測数が約1,000までであり、表示できる変数の数が2までしかなく、複数の集計レベルの値を同時に表示できないといった欠点がある。したがって、オランダは、これらの手法の限界を克服する目的で開発された2つの視覚化ツールについての報告を行った。ツリーマップ(Treemap)とは、ハードディスクの空き容量を研究するために1990年代の初頭に開発されたものであり、経済データにおいて経済活動の様々なレベルにおける分類を行うなど、階層データを視覚化するツールである。テーブルプロット(Tableplot)では、多変量のデータを1つの散布図に要約することができる。テーブルプロットを用いることで、外れ値を検出したり、特異なデータパターンを検出したり、データエディティングや補定を行っている際にデータ品質の管理を行うことができる。これらの手法は、どちらも、Rに実装されており、CRANを通じて自由に入手することができる。

WP.31 Interactive Adjustment and Outlier Detection of Time Dependent Data in R（Alexander Kowarik, Angelika Meraner, Daniel Schopfhauser, Matthias Templ, and Tu Wien, オーストリア）

時系列データ分析は、ビジネス、経済学、自然科学、計量経済学、公的統計などの応用分野において、重大な役割を担っている。米国センサス局によって開発されたX12-ARIMAの季節調整ソフトウェアは、RのGUIにおいてデフォルトで利用可能であり、非常に有用であるが、使い方が複雑であり不便でもある。Rパッケージのx12を用いることで、こういった問題を克服でき、x12-arimaのパラメータや出力を管理し、診断結果を分かりやすく提示するために、Rにおいて時系列データの前処理を行うことができる。さらに、x12のGUIでは、時系列プロット内から直接的に外れ値を手作業で選択することが可能である。

WP.32 Screening Methods and Tools for the UNIDO Industrial Statistics (INDSTAT) Databases（Matthias Templ and Valentin Todorov , UNIDO）

国連工業開発機関(UNIDO)は、産業、国、年ごとに主だった指標に関して、工業統計データベース(INDSTAT)を作成している。エラーデータや不完全なデータをスクリーニング（選別）して抽出することは重要である。UNIDOでは、たとえば、賃金や給与は負の値になり得ないなど、データ内の論理関係に基づくスクリーニングを行っている。また、それ以外にも、高度な外れ値検出法も使用している。本研究では、そういった手法の理論的背景を考察し、Rにおけるプログラミングを行って実装し、統計作成プロセスへの導入の可能性を検討している。データが多変量であることを鑑みれば、外れ値の検出はとりわけ課題となることであるが、様々な品質のデータセットにおいて、非常に単純なスクリーニング手法が、より高度な選別手法よりも効果的な場合がある。

WP.33 Automatic Data Editing with Open Source R（Mark van der Loo and Edwin de Jonge, オランダ）

オランダ統計局では、2010年より、統計作成の戦略的なツールとしてRを使用している。Rでは統計分析とデータ解析に関して様々な手法が提供されているものの、現在のところ、規則にしたがって自動的にデータエディティングを行える手法は存在していない。そこで、こういった穴を埋めるために、オランダ統計局ではeditrules及びdeducorrectというRパッケージを作成した。パッケージeditrulesにより、データエディティングの規則を定義し、視覚化することができ、Fellegi and Holtの理論に基づいて、エディット規則に合わないデータやエラーを検出することができる。パッケージdeducorrectにより、タイプミス、四捨五入による誤差、符号のエラーといったものを解決することができる。このパッケージでは、エディット規則とデータに基づいて、妥当な補定値を算出する演繹的な補定を行うこともできる。これらのパッケージは、いずれも、CRANを通じて公開している。本稿では、Rコードを付属させた簡単な例を使用して、これらのパッケージの核となる機能についてデモンストレーションを行っている。

WP.34 Application of Developed SAS Macro for Editing and Imputation at Statistics Lithuania（Vilma Nekrašaitė-Liegė and Jurga Rukšėnaitė, リトアニア）

リトアニア統計局におけるミクロエディティングは、現在、個別の部署によって独立して行われている。様々な調査における変数を審査するエディティング規則や補定規則の多くは、統計ソフトウェアSASにおいてプログラムしている。しかし、非常に高度な手法を用いている場合もあれば、そうでない場合もあり、部署や人材によっては、人手によってエディティングや補定を行っている場合もある。人手によるエラーの検出には、多大な時間がかかるため、データエディティング及び補定を、具体的に自動化されたプロセスへと変換し、すべての操作を標準化することが決定された。この目的のために、リトアニア統計局では､エディティング及び補定のためのSASのマクロプログラムを開発した。このプログラムには、エラー検出法、外れ値検出法、最近隣法を用いた補定、モデルを用いた補定、分布に基づく補定といった機能がある。

WP.35 Multiple Imputation of Turnover in EDINET Data: Toward the Improvement of Imputation for the Economic Census（Masayoshi Takahashi and Takayuki Ito, 日本）

日本は、経済センサスのデータエディティングへの適用を目指して、売上高の多重代入法に関する研究を報告した。標準的な単一代入法の限界を克服するために、多重代入法のメカニズムと利点を示し、多重代入法の最新アルゴリズムであるEMBアルゴリズムを応用したRパッケージAmelia IIの検証を行った。このアルゴリズムは、期待値最大化アルゴリズムにブートストラップを応用したものであり、巨大データセットにおける多重代入に対応できるものである。本稿の研究段階では、2012年経済センサス - 活動調査の

実データが利用可能ではなかったため、検証にはEDINETデータを用いた。多重代入法の当てはまりは、概して、単一代入法よりも優れており、多重代入法としてのAmeliaは有用なツールであることが分かった。研究成果は、2013年3月に『統計研究彙報』第70号にて刊行された。

⑹新たな手法

WP.36 Probability Editing（Thomas Laitila and Maiki Ilves, スウェーデン・エストニア）

観測値を条件とし、伝統的な標本調査手法を用いることで、観測値におけるエラーを推測でき、条件付けを緩和することで、結果を母集団推定値に一般化することができる。スウェーデンとエストニアでは、こういった確率に基づくエディティング手法を研究してきた。本研究では、確率的抽出フレームワークを用いたエディティングにおけるユニットの選択方法を提案した。確率的エディティングを実例に応用した研究では、よい結果が示されている。この手法は、あらゆる種類のデータに適用可能であり、この手法を用いることにより、推定量の統計的性質を提示することが可能となる。これは、選択的エディティングのみを用いた場合には行うことができないことである。さらに、確率的エディティングを用いることで、測定誤差によるバイアスを除去することができる。

WP.37 Use of Machine Learning Methods to Impute Categorical Data（Pilar Rey del Castillo, 欧州統計局）

カテゴリカルデータの補定に関して、数値変数用に開発された標準的な統計手法では十分ではないことが多い。欧州統計局では、ニューラルネットワーク分類法とベイジアンネットワーク分類法といった機械学習法の分野で開発された手法を検証した。ニューラルネットワーク分類法は、数値変数とカテゴリカルな変数が混在するデータを扱う手法として、近年、発展してきたものである。本稿では、これら2つの手法を用いて、世論調査のミクロデータファイルにおけるカテゴリカルデータの補定を行い、その結果を伝統的な補定法による結果と比較した。ロジスティック回帰分析や多重代入法から得られた結果との比較では、機械学習法は自動化しやすく、伝統的な手法と比べて大幅な改善が見られた。機械学習法は、巨大データセットにも拡張可能である。

WP.38 Implementation of the Bayesian Approach to Imputation at SORS（Zvone Klun and Rudi Seljak, スロベニア）

スロベニア統計局(SORS: Statistical Office of the Republic of Slovenia)は、ベイズ手法に基づく補定の実装について報告した。所得と生活状況に関する統計(EU-SILC: Statistics on Income and Living Conditions)の調査データを用いたエディティングプロセスにおいて、この手法を初めて実装した。本稿では、EU-SILCデータの中でも、事前

分布と総年収の補定を実装するためのベイズ手法を主に扱った。また、回帰モデルに基づく多重代入法を用いたベイズ手法の理論的な基盤も示した。つまり、補定による分散を適切に説明するために、この手法を複数回にわたり複製して検討をした。結論として、もし条件が満たされるならば、今回の手法はとても効果的であることが分かった。しかし、モデルによってデータが適切に記述できない場合には、結果が芳しくなかった。本稿で用いた手法は、SASを用いており、最新のバージョン(9.3)では、MCMC(Markov chain Monte Carlo：マルコフ連鎖モンテカルロ)プロシージャの一部としてすでに含められているものである。

WP.39 Automatic Editing with Hard and Soft Edits – Some First Experiences（Sander Scholtus and Sevinç Göksen, オランダ）

エディット規則は、データが満たさなければならない要件を示したものであり、人手審査によるエディティングも、自動エディティング手法も、どちらもエディット規則の情報にしたがって行われるものである。とりわけ、エディット規則を満たさないレコードに注意を払うものだが、実際には、ハードエディットとソフトエディットの区別が存在する。ハードエディットとは、ある値がエラーであったためにエディット規則を満たさなかった場合である。一方、ソフトエディットとは、ある値が必ずしもエラーではないが、疑わしい値とされたためにエディット規則を満たさなかった場合である。人手審査によるエディティングにおいては、ソフトエディットは重要視されるものであり、自動エディティングにおいても、ソフトエディットを扱うべきである。しかし、現在使用されている自動エディティングのアルゴリズムでは、すべてのエディットはハードエディットとして扱われている。そこで近年、オランダ統計局では、ソフトエディットを考慮した新しい自動エディティング手法を開発した。試作版のアプリケーションは、Rにおいて開発され、エラー検出には、既存のRパッケージを使用している。ソフトエディットを用いることにより、エラー検出に改善が見られたが、付随的に複雑となった現象の影響をさらに調査し、品質への影響を調べる必要がある。

(7)センサスデータのエディティング及び補定

WP.40 Editing of Multiple Source Data in the Case of the Slovenian Agricultural Census 2010（Aleš Krajnc and Rudi Seljak, スロベニア）

スロベニアは、2010年のスロベニア農業センサスにおける複数情報源データのエディティングに関する報告を行った。このセンサスでは、2010年6月1日から7月15日まで、約95,000の農業事業体を標本データとして抽出し、CAPI(Computer-Assisted Personal Interview)を用いることで情報を収集した。また、データの大部分は、様々な行政情報源から入手した。行政情報を利用したことにより、回答者負担が軽減されたことは疑いも

なく、調査費用も軽減することができた。しかし、行政データを使用したことにより、エディティング量が著しく増加し、すべてのデータを調査により入手した方がプロセス全体としては早く完結させることができたのも事実である。だが、正確性と信頼性という点において、行政データを使用し、またそのエディティングを行ったことにより、データ品質を著しく改善することに成功した。近代エディティングの手法として、選択的エディティングの手法が提案されており、この手法を用いていれば、データエディティングにまつわる費用を軽減し、さらなる効率性を達成できたことであろう。

WP.41 The Data Imputation Process of the Austrian Register-Based Census（Alexander Krausl, オーストリア）

オーストリアでは、2001年まで行われていた伝統的なセンサスから、2011年にはレジスターベースのセンサスに移行した。このセンサスは、オーストリア史上初の完全にレジスターに基づいたセンサスである。その結果として、データ収集、データエディティング、補定、品質管理という点で、様々な新しい課題に直面しているが、以前のセンサスと比較して、大幅な費用の削減を達成した。本稿では、オーストリアのセンサスにおける補定プロセスに関する報告を行った。主要な課題の1つは、構造的な補定手順を満たすように、変数の階層的な推定順序を確立することである。これは、データの収集されるタイミングが異なっており、補定ステップの品質を評価するために、必要なことである。レジスター内の様々な変数内に含まれている欠測値は、様々な補定手法によって推定される。ミクロデータレベルにおいて欠測値の補定は、確定的なエディティング手法や統計手法（ホットデック手法、ロジスティック回帰）を用いている。

WP.42 Item Imputation of Census Data in an Automated Production Environment: Advantages, Disadvantages and Diagnostics（Leone Wardman, Stephanie Aldrich, and Steven Rogers, 英国）

2011年英国センサスで実装した自動統計作成環境におけるエディット及び補定の長所と短所について概観する。本研究の結果では、汎用的なパラメータを設定することで、多変量の補定を自動化することが可能であり、また、補定の品質基準を満たすことができることを示した。しかし、この目的を達成するためには、非常に重要な要因がいくつか存在する。まず、実際の調査において補定を行う前に、調整とデータ分析に十分な時間が割り当てられなければならない。確かに、システムの開発段階において、実データを用いて実験を行うことにより、実用段階で必要となる調整を大幅に減らすことができるが、実際に起きる様々な回答誤差のすべてを事前に予測することは可能ではない。したがって、補定を行う前に、データの分析を十分に行うことで、体系的なエラーを検出し、可能な限り高い検出率を達成することが必要である。

WP.43 The Practical Implementation of the 2011 UK Census Imputation Methodology（Stephanie Aldrich, Leone Wardman, and Steven Rogers, 英国）

2011年のセンサスにおけるエディット及び補定の戦略では、観測データの変更を最小限にしデータ品質を維持しつつ、項目レベルの欠測値をすべて補定しすべての不完全性を是正することを重要な目標としていた。2001年センサスの後、CANCEIS（Canadian Census Edit and Imputation System)は、2011年のエディット及び補定戦略の目標を達成するために、潜在的に適切なツールであることが分かった。CANCEISは、カテゴリカルデータ、数値データ、英数字データを同時に扱えるように、センサスデータ専用に開発された最近隣ドナー補定手法を使用している。2001年センサスの合成データを用いた検証では、CANCEISは未観測の分布の推定に優れており、英国内の別々の国（イングランド、ウェールズ、スコットランド、アイルランド）や異なる種類の調査票からのデータに対して、柔軟に一貫した手法の適用を保証できることが分かった。

WP.44 Edit and Imputation of the 2011 Abu Dhabi Census（Glenn Hui and Hanan Ibrahim Al Darmaki, アブダビ）

アブダビの公的統計を近代化する目的で、アブダビ統計センターは、2008年に創設された。2011年10月には、創設後初の人口センサスを行ったが、このセンサスと以前のセンサスとの関係性は最小限のものであり、ほとんどゼロに近い状態から始めなければならなかった。公的統計の近代化のプロセスとして、エディティング及び補定は、主にCANCEISを実装することによって達成した。CANCEISの実装は、成功裏に行われたが、中東地域の人口上の特徴のため、つまり、世帯構成の社会的な差異（大家族、多妻制など）により、複数のエディット規則を変更する必要があった。全体的には、ドナー補定及び確定的補定を使用して、人手によるエディティングの必要性を最小限に抑えながらも、データセットの完全性や一貫性を達成するよう努めた。

WP.45 Editing Census Data: Mexico's Experience（Oswaldo Palma and Carole Schmitz, メキシコ）

メキシコは、地理情報システム(GIS: Geographic Information System)を用いたセンサスデータのエディティングに関して報告を行った。2010 年のセンサスでは、6 種類の調査票（建物リスト、ショートフォーム、ロングフォーム、自己申告フォーム、都市用フォーム、地方用フォーム）による伝統的な収集法を用いている。都市用フォームは、都市問題にかかわるものであり、空間的状況に基づき、潜在的なエラーや矛盾を識別するために、地図技術を用いるのにとりわけ適していた。結論として、2010 年センサスにおける自動エディティングは、以前のセンサスと比較して大幅な改善を見せ、成功裏に行われた。

付録 2: *SeleMix* の使用法（改訂版）

高橋(2012)の 6.5 節で示した手法を、より汎用的にするために、以下の方法に改訂する。ただし、「**係数の数**」の部分は、該当のデータセットに応じて変更を要する。ここで、係数の数とは、切片と傾きの数である。

```
B1<-as.matrix(c(ml.par$B[1:係数の数]))
sigma1<-as.matrix(c(ml.par$sigma[1]))
lambda1<-ml.par$lambda
w1<-ml.par$w
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)
```

`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（切片と傾き）、`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。

製 表 技 術 参 考 資 料 23

平成 25 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

掲載論文を引用する場合は、事前に下記まで連絡してください

統計情報・技術部統計技術研究課　　TEL ： 03－5273－1368

E-mail ： research@nstac.go.jp