金融トラブルに巻き込まれる要因の Lasso による分析

鈴木 明宏
高橋 広雅
竹本 亨

Research Group of Economics and Management
No. 2025-E03
2025.3

Discussion Paper Series

Faculty of Humanities and Social Sciences
Yamagata University
Yamagata, Japan

金融トラブルに巻き込まれる要因の Lasso による分析[1]

鈴木 明宏（山形大学人文社会科学部）
高橋 広雅（広島市立大学国際学部）
竹本 亨（日本大学法学部）

概要

本稿は、金融広報中央委員会の「金融リテラシー調査（2016 年・2019 年・2022 年）」の個票データを利用して、個人属性から金融トラブルに遭う可能性を予測するモデルを構築する。その際に、機械学習の一手法である Lasso を用いる。Lasso では過学習を抑えられるとともに変数選択も行われる。本稿では、それを用いることで設問数の非常に多い金融リテラシー調査の設問数を削減しても予測に問題が生じないかを調査する。

１．はじめに

人々の金融商品に接する機会が増大する中で、金融トラブルに遭遇する危険性は高まっている。そこで、金融リテラシーや行動バイアスなどに関するアンケートから金融トラブルに遭う可能性の高い人を見つけることができれば、金融教育に生かすことができると思われる。ただし、アンケートの回答が困難だと回答数が確保できなくなりがちであり、アンケート分析にも支障が出る。そのため、本稿は質問項目の少ないアンケートの作成を目指す。

政府は閣議決定した「経済財政運営と改革の基本方針 2023」の中で「2,000 兆円の家計金融資産を開放し、持続的成長に貢献する『資産運用立国』を実現する。そのためには、家計の賃金所得とともに、金融資産所得を拡大することが重要であり、iDeCo（個人型確定拠出年金）の拠出限度額及び受給開始年齢の上限引上げについて 2024 年中に結論を得るとともに、NISA（少額投資非課税制度）の抜本的な拡充・恒久化、金融経済教育推進機構の設立、顧客本位の業務運営の推進等、「資産所得倍増プラン」を実行する」とした。そして、その一環として 2024 年には金融経済教育推進機構が設立され、金融経済学習の支援事業が金融広報中央委員会[2]から移管された。

金融トラブルと個人属性の関係について論じた研究として、家森・上山（2018a）、家森・上山（2018b）、鈴木・高橋・竹本（2018）、鄭（2021）、鈴木・高橋・竹本（2024）がある。家森・上山（2018a）は独自にアンケート調査を行い、金融リテラシーや個人特性と金融トラブルの経験との関係を調査している。また、家森・上山（2018b）は同じアンケート調査を元に金融教育の経験と金融行動との関係を調査しており、その分析の中で金融教育の経験と金融トラブルの経験との関係を調査している。鈴木・高橋・竹本（2018）は 2016 年の金融リテラシー調査を用いて個人属性、特に行動バイアスと金融トラブル経験や金融行動との関連を調査した。鄭（2021）は同じ金融リテラシー調査を用いているが、対象を学生と無職を除く 18～34 歳の 2,609 人に限定して金融トラブルに遭遇する人の要因を分析している。鄭（2021）が分析対象を若年層に限定している理由は、金融トラブルを

[1] 本研究は JSPS 科研費 22K01548、並びに日本証券奨学財団の助成を受けたものである。

[2] **47** 都道府県**にも**広報委員会**がある**

経験した時期と金融リテラシー調査への回答時点を近づけることで、金融リテラシーとトラブル遭遇の関係を明確にするためである。鈴木・高橋・竹本（2024）は 2016 年に加えて 2019 年と 2022 年の金融リテラシー調査と金融トラブルの種類を質問した独自のアンケートを用いて、種類ごとの予測精度を調べた。その結果、クレジットカードの延滞や借りすぎなど借金に関するトラブルや計画性の欠如によるトラブルについては高い予測精度を示した。しかし、フィッシング詐欺など「騙される」ことによる金融トラブルについての予測精度は高くなかった。

本稿は、金融広報中央委員会の「金融リテラシー調査（2016 年）」と「同（2019 年）」「同（2022 年）」（以下では、これをまとめて「金融リテラシー調査」と略す）の個票データを利用して、個人属性から金融トラブルに遭う可能性を予測するモデルを構築する。その際に、説明変数として使用するのは、金融リテラシー調査のアンケート項目の 44 問である。

構築した予測モデルを使用して、ある人が金融トラブルに遭う可能性を予測する場合にも金融リテラシー調査のアンケートと同じ質問を、この人にする必要があるが、先程も述べたようにその項目数は非常に多い。そのため、すべてのアンケート項目に回答するにはかなりの時間を要する[3]。このことは、この予測モデルを使用して金融トラブルに遭う可能性の高い人を見つける際に障害となる可能性がある。それを回避するためには、予測の精度を高めることに貢献しないアンケート項目を見つけ出し、それを除外する必要がある。それによって、予測には十分であるとともに、回答に時間のかからない現実的なアンケートが完成する。

以上より、本稿は予測精度に貢献しないアンケート項目を除外することで解答に時間を要しないアンケートの作成を目指す。

２．データ

「金融リテラシー調査」は金融広報中央委員会が、日本人の金融リテラシーを調査する目的で全国の 18～79 歳の男女を対象行なった大規模なインターネット調査であり、これまで 2016 年、2019 年、2022 年の計 3 回実施されている。そのサンプル数は 2016 年が 25,000、2019 年が 25,000、2022 年が 30,000 の合計 80,000 となっている。

同調査における「金融リテラシー」とは、お金や金融商品についての知識・判断力を指すもので、金融庁金融研究センターにより開催された金融経済教育研究会は「金融経済教育研究会報告書」の中で「生活スキルとして最低限身に付けるべき金融リテラシー」としている。その具体的な内容は金融経済教育推進会議[4]により作成された「金融リテラシー・マップ」にまとめられており、「家計管理」「生活設計」「金融知識及び金融経済事情の理解と適切な金融商品の利用選択」「外部の知見の適切な活用」の４分野に分かれている。金融リテラシー調査には金融リテラシーに直接関連する設問 25 問以外に、金融教育に対する考え方、金融取引の状況、所得、保有資産、行動バイアス、学歴や居住地などの個人特性など、設問数は大問で 52（2022 年の場合）ある。

本研究にとって重要な設問は「Q47 あなたは、振り込み詐欺や多重債務などの金融トラブルを経験したことがありますか。」という、金融トラブルの遭遇経験を尋ねたものである。この設問は「はい」か「いいえ」の二択となっており、「はい」と回答した人を本稿では金融トラブルに遭遇した人と定義する。

---

[3] 実際、学生に金融リテラシー調査と同じアンケートを回答させると、概ね 30 分ほどかかるようである。

[4] この会議は、金融経済教育研究会報告書を踏まえ、同報告書の方針を推進することを目的として、金融広報中央委員会により設置された。

本研究の分析で用いる変数の数は 416 である。「金融リテラシー調査」の設問数が 52 であるのに対して、本研究で用いる変数が 416 になるのは、5 段階評価で回答する設問のように回答値が 2 つ以上ある設問は、値ごとにダミー変数を作成しているためである。例えば、同調査において「Q49 あなたの世帯は共働きですか。」という設問に対する回答は、「はい」、「いいえ」、「パートナー／配偶者はいない」の選択肢から一つを選ぶ形式になっている。本研究ではこの一つの設問に対して 3 つのダミー変数を作成した。すなわち、「q49_1: 「はい」を選択したら 1、それ以外の場合は 0」、「q49_2: 「いいえ」を選択したら 1、それ以外の場合は 0」、「q49_3: 「パートナー／配偶者はいない」を選択したら 1、それ以外の場合は 0」[5]となるダミーダミー変数を作成した。

3．分析方法

本稿では Lasso という手法を利用して変数選択を行うことで、設問数の多い金融リテラシー調査からどれだけのアンケート項目を削減できるかを調査する。

3.1 Lasso[6]とは

本予測モデルを構築する際に既存データへの当てはまりを重視しすぎると、将来の未知データに対する当てはまりが悪くなる可能性がある。このような現象は「過学習（overtraining）」や「過適合（overfitting）」と呼ばれる。

図 1 は適当にプロットした点を直線（一次）と曲線（24 次）で近似したものである。過学習が発生している場合には回帰係数が極端に大きくなる傾向がある（図 1 のような曲線のあてはめでは接線の傾きが極端に大きい箇所がある）ため、過学習に対処する方法として通常の最小二乗法

$$min(y - \beta_0 - \beta \cdot x)^2$$

ではなく、（問題 L）

$$min(y - \beta_0 - \beta \cdot x)^2$$
$$s.t. \|\beta\| \leq s$$

のように、係数の大きさに制約をかける方法が考えられる。ここで、$\|\cdot\|$は（何らかの）ノルムを表し、s は定数である。このような方法は「罰則付き回帰（penalized regression）」と呼ばれる。

---

[5] 多重共線性を考えれば回帰など行う場合には、Q49 については、3 ではなく 2 つのダミー変数を用いるべきである。しかし、Stata では不要な変数は自動的に省くようになっており、多重共線性の問題は発生しない。

[6] Lasso の詳細については例えば、Hastie, Tibshirani, and Friedman (2014)や James, Witten, Hastie, Tibshirani (2018)を参照。

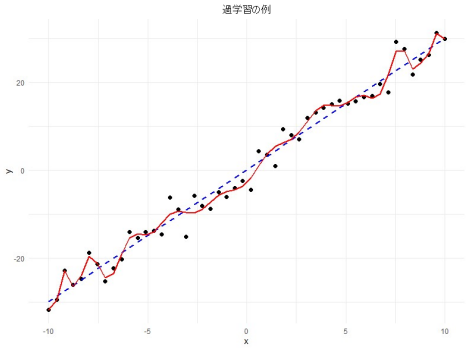

図１：過学習

Lasso は罰則付き回帰の一種で、罰則として L1 ノルム

$$\|\beta\| = \sum_{j=1}^{p} |\beta_j|$$

を採用したものである[7]。ここで、p は回帰で用いられる変数の数を表す。

図２からわかるように、Lasso の制約は滑らかではないため、最小化をした場合に端点解となる可能性が高い。多くの場合、

(1) 解析的に解けない。

(2) いくつかの係数はゼロと評価される。

ということが起こる。

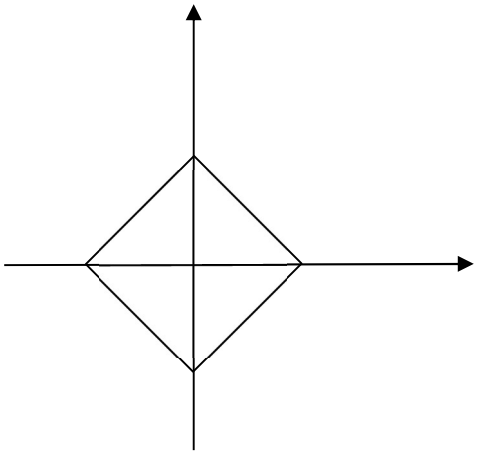
図２：L1 ノルム

（問題 L）を解くときには Lagrange 関数

$$L = (y - \beta_0 - \beta \cdot x)^2 + \lambda\|\beta\|$$

を考えればよいが、ある β_j がゼロのところでは微分不可能となるため、図３のように端点解が発生する。つまり、Lasso ではいくつかの係数がゼロとなる。そのため、Lasso は係数の決定と同時に変数選択[8]も行ってくれる。

---

[7] 同様に、L2 ノルム（係数の平方和の平方根）を採用するものが Ridge 回帰である。これについても、Hastieet al. (2014)や James et al. (2018) などを参照。

[8] 変数選択においては AIC や BIC を用いたステップワイズ法が知られているが、本稿のように元々の変数の個数が多い場合には作業量が多くなりすぎるという問題がある。対して、Lasso や Ridge では、1 つの最適化問題を解くことで使う変数が決定される。Hastie et al. (2014)

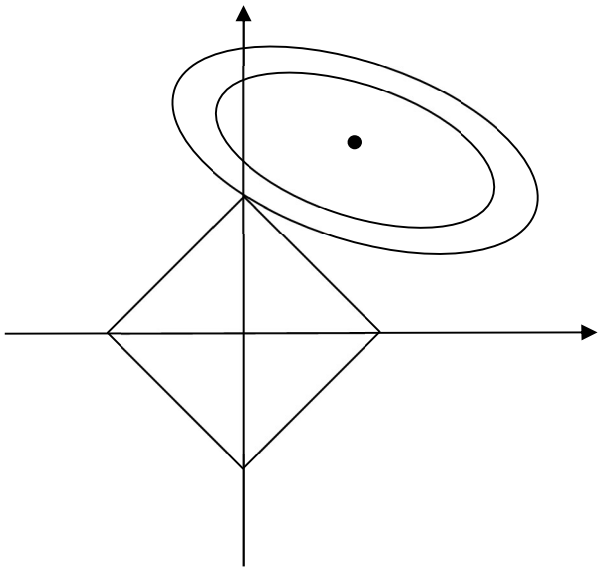

図３：Lasso の最適化問題のイメージ

3.2　モデル

本稿では、金融リテラシー調査の設問から作成したダミー変数を説明変数とする Lasso-Logit モデルを分析する。被説明変数はトラブルに遭遇するかどうかというダミー変数であるため、logit モデルを用いる必要がある。そこで、変数を減らす前のモデルは以下のようになる。

$$ln\left(\frac{トラブルに遭う確率}{遭わない確率}\right) = \beta_0 + \sum_i \beta_i \cdot (各設問から作成したダミー変数) + \varepsilon$$

この式を問題 (L) の目的関数として、係数に L1 ノルムで制約をかけるのが Lasso-Logit モデルである。

通常、Lasso におけるモデル選択では、ハイパー・パラメータ（ここでは、上記の問題 L の λ）の決定にグリッド・サーチを用いる。グリッド・サーチとは、パラメータの範囲を刻んで総当たりで探す方法であり、連続的に探すことができないために用いている。Lasso-Logit のハイパー・パラメータ選択に際しては、対数尤度の－２倍である「逸脱度（deviance）」を基準として、それが最も小さくなるものを採用することが多い。本稿では、10 分割交差検証（10-fold cross validation）によって平均逸脱度を計算し、それを選択の基準として採用する。ここで、10 分割交差検証とは、訓練データを 10 個に分割してそのうちの 1 個を（訓練段階での）テストデータとして扱い残り 9 個で学習するという作業を 10 回行い平均を取れるようにする方法である。

ただし、本稿の目的である設問数（や変数の個数）を減らすことを考えると、十分に設問数を減らすことが難しい可能性がある。そのため、単に逸脱度（の平均）を最小化する場合と「1 標準偏差ルール（one-standard-error rule）」を採用した場合を比較する。ここで 1 標準偏差

では、AIC や BIC を使用した変数選択は「部分集合選択」、Lasso や Ridge は「縮小推定」と分類されている。これらの手法間の比較については例えば、末石 (2019)などを参照されたい。

ルールとは、交差検証時に単に逸脱度を最小化する代わりに最小化された逸脱度から 1 標準偏差分だけ大きくなることを許容する基準である[9]。交差検証において複雑なモデルを選択することを回避する手法として、1 標準偏差ルールは提案されている。

４．分析結果

分析結果の概要は表１の通りである。

		逸脱度	deviance ratio	変数の個数	小問数
min	train	0.420	0.149	150	72
	test	0.418	0.137	-	-
serule	train	0.442	0.103	6	5
	test	0.430	0.111	-	-

表１：結果

逸脱度を見る限り、どちらのモデルでも訓練データとテストデータの間に違いはなく、過学習は回避できていると考えられる。

逸脱度を基準としてその最小化を行った場合には、使用される変数の個数が 150 個と大幅に減ったように見えるが、使用される小問数で数えると 72 個である。小問数は全部で 92 個あるので、Lasso によって 20 個が使われなかったことになる。使われなかった変数の一覧は付録の表３の通りである。表３を見ると、収入・支出面についての設問が多い。これらの変数が除外された理由としては次のようなことが考えられる。収入・支出額はトラブルの経験の有無とは関係なく多くの人が把握しているので、分類[10]には不要であるのだろう。

逸脱度最小化ではそれほど変数が減少していないので、1 標準偏差ルールを適用してみたところ、逸脱度は若干上昇したが、使用される変数の個数が 6 個、小問数では 5 個と劇的に減少した。使用される小問の内容は表２の通りである。

設問番号	内容
Q1_8	お金を借りすぎていると感じている。
Q11	病気、失業、不景気等の万が一の事態に備えて、３か月間分の生活費を確保してありますか。

9 「標準偏差ルール」で許容幅を１標準偏差分にしたものが１標準偏差ルールである。一般的な１標準偏差ルールと標準偏差ルールについては例えば、竹澤 (1999) を参照。

10 被説明変数が連続で値の予測が目的の場合には回帰と呼び、被説明変数が２値（かそれ以上の離散値）で各サンプルのクラス分けの予測が目的の場合には分類と呼ぶ。

Q29	あなたは、過去に金融機関から１か月の生活費を超える金額のお金を借りたことがありますか。最後にお金を借りた際、ご自身の状況に適したローンを選ぶために、他の金融機関あるいは他のローンと比較しましたか。
Q45_2	あなたご自身（あなたの世帯）」は、借入れをしていますか。（消費者ローン）
Q45_3	あなたご自身（あなたの世帯）」は、借入れをしていますか。（その他の借入）

表２：1 標準偏差ルールを適用した場合に使われる小問

Q1_8 は借金の不安についての設問であり、Q45 は借り入れについての設問である。Q1_8 については元々の質問が曖昧すぎたものであるのはいかがなものかということはあるが、詐欺的手法の一部は漠然とした不安につけ込むと考えればそのような人を拾うためには有用なのかもしれない。また、そもそもローンがなければ、借金がらみのトラブルには遭いにくいことが容易に想像される。ただし、住宅ローン（Q45_1）については使われなかった。住宅ローンについては消費者金融など他のローンよりもトラブルに結びつきにくいことが原因と考えられる。

５．終わりに

本稿では金融リテラシー調査データに Lasso を適用することで膨大な数の説明変数をどの程度分析に支障が生じない程度に減らすことが可能か調査した。説明変数の数を減らすことで設問数も減らすことができれば、回答者の負担を減らすことが可能となる。

分析の結果、単純な逸脱度最小化によるモデル選択では小問数が 92 から 72 に減った。回答者の負担軽減を考えると 22%程しか減らせていない。そこで、より削減効果の期待できる 1 標準偏差ルールを適用したところ、小問数は 5 に減少した。大幅に小問数は減少したが、本稿で用いたデータでは逸脱度はそれほど変化していなかった。

しかしながら、残された小問を具体的に見ると、いずれもトラブル遭遇に影響すると思われる変数が残っていることがわかるものの、同じくトラブル遭遇に影響を与えると思われる行動バイアスが全て落ちてしまっている。このことは 1 標準偏差ルールの変数除去効果が強すぎることを示唆する。通常、1 標準偏差ルールより細かく調整するようなオプションは用意されていないため、結果を見ながら必要な変数を残すといった調整を行う必要がある。

また、説明変数の減少が必ずしも設問数の減少につながらない理由の一つは通常の Lasso が変数間の関係を考慮していないことにある。変数間の関係を考慮した拡張として Group Lasso（Yuan and Lin, 2006）や Adaptive Group Lasso（Huang, Horowitz, and Wei, 2010）等の手法が提案されている。これらを用いた分析は今後の課題としたい。

参考文献

Hastie, T., R. Tibshirani, and J. Friedman (2014)『統計的学習の基礎 ―データマイニング・推論・予測―』, 共立出版.

Huang, J., Horowitz, J. L. and Wei, F. (2010). “Variable selection in nonparametric additive models,” *Annals of Statistics*, 38, 2282-2313.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2018)『R による統計的学習入門』, 朝倉書店.

Yuan, M. and Lin, Y. (2006). “Model selection and estimation in regression with grouped variables,” *Journal of the Royal Statistical Society: Series B*, 68(1), 49-67.

家森信善・上山仁恵（2018a）「生活者の金融リテラシーと金融トラブルー 2016 年・金融リテラシーと金融トラブルに関する調査をもとに」『生活経済学研究』47，1-18.
家森信善・上山仁恵（2018b）「学校での金融経済教育の経験が金融リテラシーや金融行動に与える影響」『ファイナンシャル・プランニング研究』17，52-71.
末石直也(2019)「Lasso における正則化パラメータの選択方法について」『國民經濟雜誌』第 220 巻第 4 号，51-65.
鈴木明宏・高橋広雅・竹本亨（2018）「金融教育と行動バイアスが金融行動と金融トラブルへの巻き込まれやすさに与える影響：金融リテラシー調査データを利用した分析」『山形大学紀要（社会科学）』第 49 巻第 1 号，1-13.
鈴木明宏・高橋広雅・竹本亨（2024）「金融リテラシー調査からどのような金融トラブルとの遭遇を予測できるか」『山形大学紀要（社会科学）』第 55 巻第 1 号，1-15.
竹澤邦夫(1999)「Standard Error ルールを用いた予測変数選択」『応用統計学』第 28 巻第 1 号，21-25.
鄭美沙（2021）「若年層のリスク性資産購入経験と金融トラブル経験に関する実証分析」『生活経済学研究』54，45-58.

付録

設問番号	内容
Q1_9	投資や預金をするときには、お金を損することがあってもしかたがないと思う。
Q3_1	1 か月の収入や支出の金額を把握していますか。（収入）
Q3_2	1 か月の収入や支出の金額を把握していますか。（支出）
Q7	次の費用のうち、あなたが今後必要になると意識しているものは、どれですか。あてはまるものをいくつでも選んでください。
Q8	今後必要になると意識している費用について、ご自分の場合の必要額を認識していますか。
Q8_1	年金
Q8_6	車の購入費用
Q8_8	子どもの結婚費用
Q9	今後必要になると意識している費用について、資金計画をたてていますか。
Q9_2	子どもの教育にかかる費用
Q9_3	住宅の購入費用
Q9_5	家族の医療・介護費用
Q9_6	車の購入費用
Q9_7	自分の結婚費用
Q9_8	子どもの結婚費用
Q9_9	その他
Q10	今後必要になると意識している費用について、資金を確保できていますか。
Q10_1	定年退職後の生活費
Q10_7	自分の結婚費用
Q10_8	子どもの結婚費用
Q10_9	その他
Q28	保険に関する以下の記述のうち、適切でないものはどれでしょうか。
Q33	預金保険制度で １千万円まで保護される預金の種類に関する次の記述のうち、適切なものはどれでしょうか。

表３：使用されなかった変数