

# 資産リターンの予測可能性と機械学習 —危険なデータマイニング—

内山 朋規, 瀧澤 秀明, 菊川 匡

資産のリターンが予測可能であることは、現在のファイナンス理論における標準的な考え方である。さらに近年では、機械学習やビッグデータに注目が集まっている。実際の市場は単純なモデルの世界よりも複雑であろうことから、リターンを予測するうえでこれらは有効なツールとなるだろう。一方で、より有意な実証的証拠を得ようと変数選択やモデル選択を繰り返すと、見かけ上は予測力があるモデルを得られてしまう厄介な問題を引き起こす。本稿では、オーバーフィッティングの影響はどの程度生じるのか、こうした問題を考慮したうえでリターンの予測可能性をどのように検証するべきか、について論じる。

キーワード：リターンの予測可能性、機械学習、ビッグデータ、多重検定、オーバーフィッティング、偽発見

## 1. はじめに

資産価格は将来キャッシュフローの期待割引価値として定まり、その割引率（言い換えれば、期待リターンあるいは資本コスト）はリスクフリーレートにプレミアムが加わったものであることから、プレミアムの特徴を解明することは、資産価格理論における中心的テーマであり続けてきた。プレミアムはクロスセクションで異なり、かつ時系列に変動（時間変動）しているため、何らかの予測変数（ファクターや特性と呼ばれる）によって、クロスセクションと時系列の双方でリターンが予測可能であることは、現在の学術界における標準的な考え方である（たとえば Cochrane [1] 参照）。株式市場を例にとると、純資産株価倍率 [2] や配当利回り [3] は、それぞれクロスセクションと時系列におけるよく知られた予測変数である。

プレミアムの実証的な特徴を満足に記述できるモデルはまだ存在せず、多くの研究者がさまざまな資産（株式、債券、クレジット、デリバティブ、通貨、コモディティ、不動産）を対象に、クロスセクションあるいは時系列における予測変数（ファクター）の研究を競い合い、新たな予測変数を「発見した」とする報告が続いている（筆者らも含まれる）。たとえば Harvey

et al. [4] によれば、米国株式市場におけるクロスセクションの予測変数として、これまでトップジャーナルに「発見」が報告されたファクターは 316 個もあり、「発見」される頻度は年々高まっている。他の資産クラスや時系列の予測変数に関しても同様で、2011 年の米国ファイナンス学会会長講演 [1] では、学術界のこうした状況が“zoo of new factors”と称されている。

新たなファクター、あるいは改良を加えることによって、より予測精度が高まる方法を見つけることは、学術界だけでなく、実務の研究者によっても精力的に行われている。得られた結果を資産運用の投資戦略や商品開発に利用できるためである。特に最近、これまでの学術研究の成果を実務に直接応用するファクター投資の考え方が実践されている。

さらに近年では、機械学習やビッグデータに注目が集まっている。データスヌーピングとも呼ばれ、ネガティブな意味の用語として使われてきたデータマイニングが、ファイナンスの分野でもポジティブな意味として使われる場面が見受けられる<sup>1</sup>。予測対象のリターンデータは時間の経過分しか増えない一方で、情報入手コストの低減に伴い予測変数の候補として使用可能なデータは飛躍的に増大している。数値データのみならず、今日では Twitter 投稿のテキスト分析や、人工

うちやま ともり  
東京都立大学大学院経営学研究科  
〒100-0005 東京都千代田区丸の内 1-4-1  
uchiyama@tmu.ac.jp  
たきざわ ひであき, きくがわ ただし  
野村證券株式会社ストラクチャード・プロダクト部  
〒100-8130 東京都千代田区大手町 2-2-2  
hideaki.takizawa@nomura.com  
tadashi.kikugawa@nomura.com

<sup>1</sup> たとえば、MATLAB の開発元である MathWorks 社のパンフレット『機械学習のご紹介』(MathWorks [5]) には、「機械学習のさまざまなアルゴリズムは、データに潜む自然なパターンを見つけ出し、そこから洞察を導き、あなたがよりよい意思決定や未来予測をするのに助けてくれます。これらは、日々の医療診断や株取引、エネルギー需要予測など、さまざまな場面での意思決定に利用されています。」と記載されている。

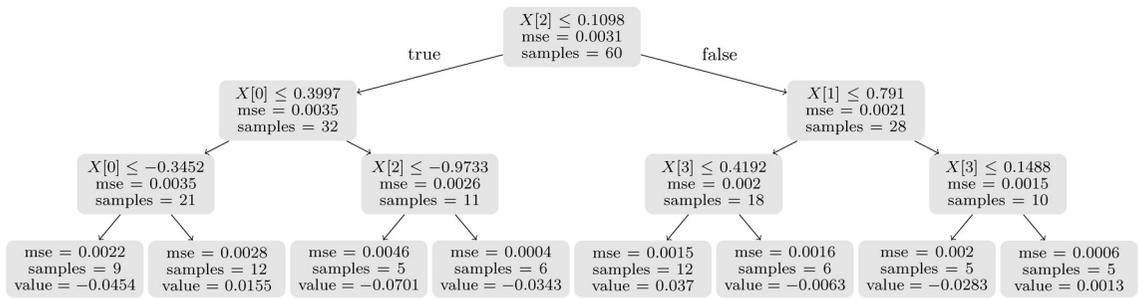


図 1 4 種類のマクロ変数を用いた回帰木による TOPIX リターンの予測の例 (モデル X)

(注) 回帰木による 2011 年 1 月分の TOPIX リターンの予測。X[0] は TOPIX 売買回転率 (前月差), X[1] は JGB10 年利回り (前月差), X[2] は UST2 年利回り (前月差), X[3] は鉱工業生産 (前月比) を表す。

衛星画像の交通量解析といった非構造化データも利用可能である。また、近年は工学的な側面から機械学習への注目度が増している。線形回帰だけでなく、回帰木やニューラルネットワークといった非線形モデルも利用される。計算処理能力の向上も相まって、データマイニングはこれからはますます容易になるだろう。しかし、データマイニングはオーバーフィッティング (過剰適合) というゆゆしき事態を招く。

クロスセクションや時系列におけるリターンの予測可能性に関して、合理的説明と行動論的説明の双方からさまざまな理論的根拠が示されているものの、統計的にこれを検出することは容易ではないことが知られる (たとえば Cochrane [6] 参照)。そこで新たな実証的証拠を発見するには、データを念入りに調べることになる。一方で、念入りに調べるほど、本来は意味がないにもかかわらず、興味深いパターンが偶然に出現してしまうというオーバーフィッティングの可能性が高まる [7]。偽発見であればこうした成果に価値はなく、むしろ成果が応用されることによって弊害をもたらす。

オーバーフィッティングは、予測対象の標本数が有限であるにもかかわらず、変数選択に自由度があること、すなわち、変数の選択や値の処理方法、変数の組み合わせ方などに自由度 (あるいは恣意性) があることから生じる。加えて、モデル選択に自由度があること、言い換えれば、複数のモデルから「ベストフィット」するモデルを選択することからも生じる。このため、検定の際に多重性を考慮する必要がある。本稿では、時系列におけるリターン予測を対象に、オーバーフィッティングの影響を分析する。変数選択によるマイニングだけでなく、機械学習に伴うモデル選択によるマイニングの影響も扱う。

## 2. 時系列のリターン予測

### 2.1 例：予測モデル X

時系列において将来のリターンを予測する例から始めよう。月次データを用いて、アウトオブサンプルで毎月末に翌月の TOPIX (東証株価指数) リターンの予測を行う。具体的には、

1. TOPIX 売買回転率 (前月差)
2. JGB (日本国債) 10 年利回り (前月差)
3. UST (米国債) 2 年利回り (前月差)
4. 鉱工業生産 (前月比)

の 4 種類のマクロ変数を予測変数として、回帰木により翌月の TOPIX リターンの期待値を推定する。回帰木のパラメータは毎月末に過去 60 ヶ月間のデータから再推定する。たとえば、2011 年 1 月分の TOPIX リターンの予測値は、2010 年 12 月までの過去 60 ヶ月間の TOPIX リターンと前月のマクロ変数の値からモデルのパラメータを推定し、2010 年 12 月のマクロ変数の値を用いて予測する。

使用データの期間は 2006 年 1 月から 2015 年 12 月であるが、過去 60 ヶ月間のデータを推定に用いるため、予測の分析期間は 2011 年 1 月から 2015 年 12 月までの 60 ヶ月間である。図 1 は、例として 2011 年 1 月分の予測分を表す。以後、この予測方法をモデル X と呼ぶことにする。

### 2.2 検定

このモデルに予測力はあるだろうか。これを検定するため、ナイーブに過去 60 ヶ月間の TOPIX リターンの平均を翌月の期待値とするものをベンチマークとして比較する。 $t$  月における TOPIX の実現リターンを  $r_t$  とし、前月末におけるベンチマークによる予測値を  $\bar{r}_t := E_{t-1}^{\text{benchmark}}[r_t]$ 、上記の 4 種類のマクロ変数を

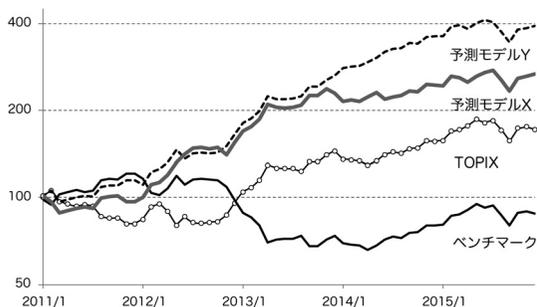


図2 TOPIX リターン予測

(注) 予測モデル X, 予測モデル Y, ベンチマークは、それぞれのモデルの TOPIX リターンの予測値が正ならばロング、負ならばショートした戦略の累積リターンを表す。2011 年 1 月から 2015 年 12 月まで。月次リバランス。

用いた回帰木による予測値を  $\hat{r}_t := E_{t-1}^{\text{model } X}[r_t]$  とおく。ベンチマークの予測誤差を  $\bar{u}_t := \bar{r}_t - r_t$ 、モデル X の予測誤差を  $\hat{u}_t := \hat{r}_t - r_t$  とすると、もしモデル X の方がベンチマークよりも予測力が高ければ、予測誤差の平方  $\hat{u}_t^2$  の期待値は、 $\bar{u}_t^2$  の期待値よりも小さいはずである。

実際には分析期間のデータを用いて、 $\bar{u}_t^2$  と  $\hat{u}_t^2$  の差を評価するが、仮にモデル X にベンチマークを上回る予測力がないとしても、標本が有限個であるために  $\bar{u}_t^2 - \hat{u}_t^2$  の標本平均はゼロではなく、負になるバイアスをもつ。そこで、Clark and West [8] の方法により、このバイアスを補正した以下の  $f_t$  を定義する。

$$f_t := \bar{u}_t^2 - \hat{u}_t^2 + (\bar{u}_t - \hat{u}_t)^2 \quad (1)$$

第 3 項がバイアスの補正項である。もし  $f$  の平均が有意に正であれば、モデル X の予測精度はベンチマークよりも高いと判断される。したがって、片側検定により診断すればよい。

この結果、モデル X の  $f$  の平均値の  $t$  値は 3.96 となった。5% 有意水準に対応する漸近的な  $t$  値の臨界点は 1.64 なので、モデル X の  $t$  値はこれを超えている。したがって、4 種類のマクロ変数を用いた回帰木による予測の精度は、過去平均によるナイーブな予測を有意に上回る。

図 2 は、モデル X やベンチマークの予測リターンが正の場合に TOPIX をロング、負の場合にはショートする、月次リバランスによる戦略の累積リターンを表したものである（モデル Y については 4.1 節で述べる）。モデル X はあくまで翌月のリターンの水準を予測するものであり、その方向、すなわち符号が正か負を予測するものではないが、予測値の符号に応じてポ

ジションを切り替えると、ベンチマークの予測によるロングショート戦略や、TOPIX そのもの（これは常にロングする戦略に相当する）よりもパフォーマンスは良好であることがわかる。

はたして、モデル X は本当に有効なのであろうか。データマイニングの結果でなければ（すなわち単一検定ならば）、有意といえる。しかし、データマイニングを行い、予測精度が高いモデルを探し出した結果であるならば、どう考えるべきであろうか。

### 3. 既存研究

データマイニングに伴う危険性は古くから認識されてきたが、その対処は、経済学的原理を裏付けとした実証を行う、実証方法やモデルや変数を簡潔にする、純粋な実証分析であれば広範にロバスト性を検証する、といった定性的な規範にもっぱら頼ってきた。ファイナンスでは従来、オーバーフィッティングの問題はさほど定量的に扱われてこなかった。ところが近年、このテーマに注目が集まりつつある。代表的な既存研究はいずれも米国株式市場の個別銘柄を対象に、クロスセクションにおけるリターン予測について論じている。

McLean and Pontiff [9] は、リターンの銘柄間の格差を予測する既知の 97 個のファクター（特性）を対象に、論文の公表後に超過リターンが約半分の大きさに低下することを報告し、この原因として、投資家にファクターが知れ渡ったことによる影響か、あるいは単にデータマイニングの結果に過ぎなかったこと（またはその双方であること）を論じている。

また、Harvey et al. [4] は、過去に行われてきた実証分析すべてを多重検定として認識すべきであることを論じ<sup>2</sup>、2012 年までに代表的なジャーナルに掲載された 316 個のファクターを分析の対象にしている。論文に記載されている  $t$  値 ( $p$  値) を抽出し、多重検定であることを考慮して Bonferonni 法などで調整を行うと、掲載されたファクターのうち半分程度は有意とはいえず、偽発見として判定されることを述べている。さらに、両側検定<sup>3</sup>における 5% 水準の  $t$  値の臨界点について、単一検定では 2.0 であるが、多重検定である

<sup>2</sup> 多重性の問題は、ある一人の研究者が異なる予測変数やモデルを用いて分析を繰り返すことほかに、複数の研究者がそれぞれ分析を行っても生じる。たとえば有意水準を 5% とし、100 人の研究者が独立に分析を行う場合、そのすべてが本当は有意でないとしても、確率的に 5 人は有意な結果を得る。この 5 人が結果を公表すると、真実とは異なる成果が世の中に出回ることになる。

<sup>3</sup> 超過リターンの平均が有意にゼロとは異なるかどうかを検定されることから、両側検定による。

表1 予測変数の候補

TOPIX	UST2 年利回り*	バルチックドライ指数	米製造業 ISM・新規受注*	ユーロ景況感指数*
TOPIX 配当利回り*	米 LIBOR*	景気ウォッチャー現状*	米製造業 ISM・雇用*	ユーロ消費者信頼感指数*
TOPIX PBR*	S&P 商品価格インデックス	現状・家計*	米製造業 ISM・価格*	中国消費者信頼感指数*
TOPIX PSR*	S&P レバレッジドローン指数	現状・企業*	米非製造業 ISM*	豪州消費者信頼感指数*
TOPIX 売買回転率*	NY ダウ	現状・雇用*	米非製造業 ISM・新規受注*	最終需要財在庫率指数
TOPIX ボラティリティ*	USDJPY	景気ウォッチャー先行き*	米非製造業 ISM・雇用*	機械受注
JGB10 年利回り*	USDEUR	先行き・家計*	米非製造業 ISM・価格*	新設住宅着工床面積
JGB2-10 年スプレッド*	金スポット	先行き・企業*	米消費者信頼感指数*	耐久消費財出荷指数
UST10 年利回り*	VIX 指数	先行き・雇用*	米消費者信頼感・現状*	日経商品指数
UST2-10 年スプレッド*	ユーロストック	米製造業 ISM*	米消費者信頼感・先行き*	鉱工業生産

(注) \*は前月差, 他は前月比を表す。

ことを考慮すると 2012 年時点において 3 から 4 程度に設定すべきであることを論じている。

Novy-Marx [10] は, ファクターを探索することによるデータマイニングの影響を調べている。ある一つのファクターの値が大きくなる (もしくは小さくなる) ようにポートフォリオを組むと, 平均超過リターン (アルファ) の実現値は正か負の値になる。事後的にアルファが正になる方向でファクターを採用する余地が分析者にはある。実証分析における標本数は必然的に有限なので, 多くのファクターを試せば, 本来は予測力がなくても, 有意に見えるものが出現する。さらに, こうした選択バイアスだけでなく合成バイアスもあり, 分析者には複数のファクターを事後的に定めたウェイトで合成する余地もある。Novy-Marx [10] は, 正規確率変数 (これらに予測力はない) をファクターに用いて, 選択バイアスや合成バイアスを考慮した  $t$  値の分布を算出している。特に,  $n$  個のファクターを試してこの中から事後的に  $t$  値の高い  $k$  個を合成して使用することは,  $n^k$  個という非常に多くのファクター候補から  $t$  値が最大のもを一つ選ぶのに匹敵することを示している。

一方, Yan and Zheng [11] は, 実際のファクターに着目している。たとえば 1 節で述べた純資産株価倍率は代表的な財務上の予測変数で, この値で銘柄をソートしてロングショート戦略のポートフォリオを構築すると, アルファは有意にゼロとは異なることが知られている。これに対する自然な疑念として, 多くの財務変数を試した結果, 偶然に純資産株価倍率を用いた場合の  $t$  値が高かっただけかもしれない。これを考えるには, 利用可能なすべての財務変数を対象にした多重検定を行えばよい。Yan and Zheng [11] は, 18,113 個の財務変数を対象にそれぞれのアルファを算出し, これらの  $t$  値のクロスセクションにおける分布と, アルファをゼロとおいた人工的な超過リターンからブートストラップ・シミュレーションによって得たアルファ

の  $t$  値の分布を比較している。もし, 実際に観測される極端な  $t$  値が偶然であるならば, 高い  $t$  値は真のアルファがゼロでもある程度の頻度で出現するはずである。しかし, Yan and Zheng [11] は, 財務変数もたらず実際のアルファの  $t$  値のうち極端に高いものが, アルファをゼロとしたシミュレーションでは 5% 以下でしか出現しないことを示し, データマイニングでは説明できないことを述べている。

これらの既存研究はいずれもクロスセクションにおけるリターン予測のオーバーフィッティングを扱っている。一方, 時系列についてはこれまでほとんど探究されていない。時系列はクロスセクションに比べて予測対象の標本数が少ないことから, オーバーフィッティングの影響はより深刻なはずである。本稿では, 時系列におけるリターン予測のオーバーフィッティングに関して, 変数選択によるマイニングだけでなく, 機械学習に伴うモデル選択によるマイニングの影響も扱う。

## 4. 時系列予測におけるオーバーフィッティング

### 4.1 予測モデル $X$ の構築方法

2 節で述べたモデル  $X$  に話を戻そう。実は, 予測モデル  $X$  は多くの候補の中から「アウトオブサンプル」での予測精度が最も高いものを選択した結果である。具体的には, 表 1 に記載した 50 種類の市場データや経済指標<sup>4</sup>のマクロ変数を予測変数の候補とし, さらに線形回帰,  $k$  近傍法, 回帰木<sup>5</sup>の 3 種類の予測モデルの中から, 式 (1) で定義された  $f$  の平均値の  $t$  値が最大になったものである。

具体的な計算方法は以下による。50 個のマクロ変数のうち少なくとも一つを用いると, 予測変数の組み合わせ

<sup>4</sup> 経済指標は公表までの間にラグがあるが, これを考慮して各自時点で入手可能な直近値を使用した。

<sup>5</sup> 回帰木のハイパーパラメータは, 木の深さを 3, 各枝先端の最小サンプル数を 5 として設定した。また,  $k$  近傍法のハイパーパラメータは  $k = 5$  とした。

わせは  $2^{50} - 1$  通りあり、さらに予測モデルが 3 通りあるため、計  $(2^{50} - 1) \times 3 \approx 3,300$  兆通りの組み合わせがある。これらのすべてについて  $t$  値を算出するには莫大な時間を要することから、3 種類の予測モデルのそれぞれについて、50 個のすべての予測変数から始めて、 $t$  値が最も高くなるように、1 個ずつ予測変数を減らした。この方法では  $2^{50} - 1$  通りの中から  $t$  値が最大となる予測変数の組を選択できるとは限らないが、計算時間は大幅に短縮される。最後に、3 種類の予測モデルのうち、 $t$  値が最大のものを選択した。

したがって、モデル  $X$  の予測力を評価するには、多重検定の枠組みが必要になる。予測対象の TOPIX リターンの標本数は 60 個しかないのに対して、予測変数の候補は 50 個もあって極端なように思えるかもしれない。しかし、予測対象の標本数に対して、予測変数の候補が多いことが問題になるため、ビッグデータを扱う最近の風潮では現実的な設定である。また、線形回帰、 $k$  近傍法、回帰木の三つのモデルから「ベストフィット」するものを採用するのは、近年の機械学習への注目に对应している。

なお、図 2 のモデル  $Y$  は、50 個の予測変数と 3 通りのモデルを用いている点ではモデル  $X$  と同様であるが、予測精度が最大になるもの、すなわち  $f_t$  の平均値の  $t$  値が最大になるものを探索したのではなく、翌月の TOPIX リターンの方向（正か負か）的中率を「アウトオブサンプル」で最大になるように探索した結果である<sup>6</sup>。モデル  $Y$  の方が図 2 の投資戦略に合致しているが、本稿ではリターンの予測精度を議論の対象にするため、以後ではモデル  $X$  のみについて論じる。

#### 4.2 データマイニングの影響

モデル  $X$  の予測精度は有意なのであろうか。多重検定であることを考慮するために、50 種類の予測変数（マクロ変数）を 50 個の多変量正規乱数に置き換えたシミュレーションを行う。具体的には、50 個の予測変数（乱数）を用いて、4.1 節で述べた方法により、 $t$  値が最大になるものを探索する。このシミュレーションを 1 万回繰り返す。したがって、シミュレーションごとに、 $t$  値が最大になるものとして選択された予測変数（乱数）の組やモデルは異なる。乱数には予測力がないため、シミュレーションから得られた  $t$  値の分布は、予測力がないという帰無仮説もとの  $t$  値の分布に対応する。もし、モデル  $X$  の  $t$  値がシミュレーションによる 5% 臨界値よりも大きければ、5% 水準で有意といえる。

<sup>6</sup> モデル  $Y$  は、31 個のマクロ変数（記載を省略）を予測変数とし、回帰木により予測するものである。

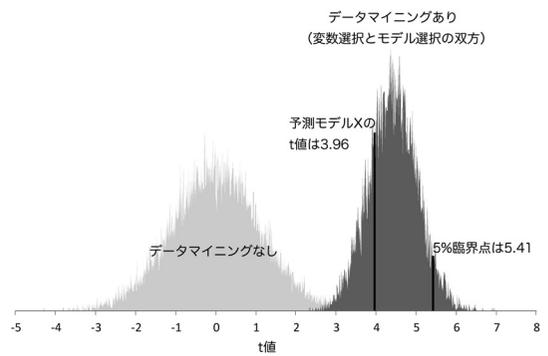


図 3  $t$  値の分布

(注) 乱数を予測変数に用いたシミュレーションによる  $t$  値の分布を表す。予測対象は 2011 年 1 月から 2015 年 12 月までの TOPIX リターン。月次データによる。

表 2  $t$  値のパーセンタイル

	中央値	5%
データマイニングなし		
線形回帰	0.00	1.74
$k$ 近傍法	0.00	1.80
回帰木	-0.01	1.69
データマイニングあり		
モデル選択のみ	0.83	2.17
変数選択のみ（線形回帰）	3.60	4.97
変数選択のみ（ $k$ 近傍法）	4.14	5.28
変数選択のみ（回帰木）	3.80	4.90
変数選択とモデル選択の双方	4.41	5.41

(注) 図 3 の  $t$  値のパーセンタイルを表す。

ここでは 50 個のマクロ変数の相関を考慮した正規乱数を用いるが、マクロ変数が正規分布に従っているとは限らないため、本来は Clark and McCracken [12] で述べられているようなブートストラップなどのノンパラメトリックな方法が望ましいだろう。一方で、本稿の乱数の方法は実装が容易で、モデルの予測力の有意性を簡単に検証できるという利点がある。

この結果を図 3 と表 2 に表した。グラフの「データマイニングなし」の  $t$  値の分布は、3 通りのモデル（線形回帰、 $k$  近傍法、回帰木）のそれぞれについて、50 個の予測変数（乱数）を選択せずにすべて用いた結果を表す。当然のことながら、三つのモデルの  $t$  値の分布は中央値がゼロである。これらは同様な形状をしているため、同色で重ねて表示した。

一方、表 2 の「データマイニングあり」の最下行は予測変数（乱数）と 3 通りのモデルの双方で選択を行った場合を表す。本来、乱数には予測力がないにもかかわらず

ず、 $t$  値は随分と高い水準にシフトすることがわかる。その5%点は5.41と相当に高い。モデル  $X$  は  $t$  値が3.96だったので有意とはいえないばかりか、中央値の4.41にも満たない。すなわち、予測力がない乱数を予測変数に用いても、変数選択とモデル選択を行えば、 $t$  値が4.41以上となる予測精度のものが50%の確率で出現する。このことは、モデル  $X$  には本来予測力がなく、分析期間中では偶然に予測力があるように見えるに過ぎないことを意味している。モデル  $X$  を分析期間よりも将来の予測のために用いたとしても何の価値もない。

表2には、モデル選択のみを行った場合や、変数選択のみを行った場合の結果についても記載した。モデル選択のみは、50組の予測変数(乱数)をそのまま用いて、3通りのモデルから  $t$  値が最も高いものを選択したモデルマイニングの結果を表す。 $t$  値の5%点は2.17で、Bonferroni法による有意水準の補正  $0.05/3$  に対応する  $t$  値の2.13にはほぼ等しい。また、モデル選択を行わずに、変数のみをマイニングしてもシミュレーションの  $t$  値は相当に上昇することがわかる。

これまで50個の予測変数の候補を考えてきたが、次に予測変数の候補の数を変えた場合の影響について考察する。予測変数(乱数)の候補数を1個から30個まで一つずつ増やしたシミュレーションを行う。ここでの正規乱数は互いに独立とするが、他の条件はこれまでと同様である。図4にシミュレーションによる  $t$  値の5%点をプロットした。「変数選択のみ」で、予測変数の候補数が1個の場合は、マイニングをしていないことを意味するため、単一検定に相当する。この場合の  $t$  値は漸近的に1.64で、シミュレーションの結果もほぼこれに等しい。

5%点の  $t$  値は、予測変数の候補数に関して単調に増加し、当初の上昇は急激なことがわかる。上昇の程度はモデルによって異なり、たとえば候補となる予測変数が5個のとき、線形回帰では2.39であるが、回帰木では2.71、 $k$ 近傍法では2.76とより高い。さらに、変数選択だけでなくモデル選択も併せて行くと、当然のことながら、5%点の  $t$  値はさらに高まる。たとえば予測変数の候補が5個のときの  $t$  値は3.05になり、これは単一検定であれば漸近的な  $p$  値の0.11%に相当する。

要約すると、たとえ候補とする予測変数が少数であっても、変数選択の影響は大きく、線形回帰よりも回帰木や  $k$  近傍法の方が変数マイニングの影響が深刻である。加えて、モデル選択も併せて行くと、さらに影響は大きくなる。

なお、本分析では  $k$  近傍法と回帰木のハイパーパラメータを固定して用いたが、本来はハイパーパラメー

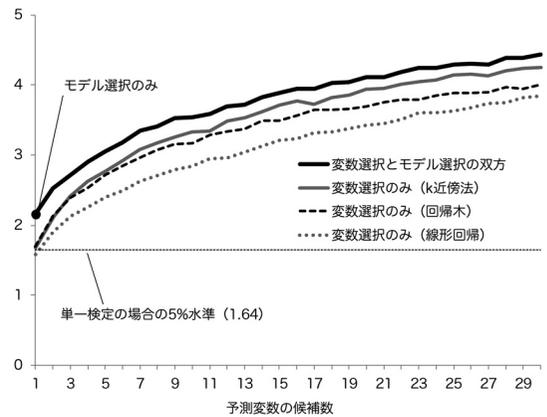


図4 予測変数の候補数と5%水準の  $t$  値  
(注) 乱数を予測変数に用いた  $t$  値の5%臨界点を表す。予測対象は2011年1月から2015年12月までのTOPIXリターン。月次データによる。

タの選択にも自由度がある。これも併せてマイニングを行うとオーバーフィッティングの影響はさらに高まる。また、機械学習の文脈では、標本の期間を「トレーニング期間」「バリデーション期間」「テスト期間」に分けて分析することがあるが、これらをすべて含めて繰り返し分析を行い、ベストフィットするモデルを探索すれば本稿の設定と同様の問題が発生する。

ヒストリカルデータを用いて構築した予測モデルがモデル構築後に機能しなくなることがあり、その理由を「市場のパターンや環境が変わったため」として認識してしまうことがしばしば見受けられる。しかし、それが複数の予測変数やモデルの候補から選択されたものであるならば、そもそも本質的に予測力がなかった可能性が高いことを本稿の結果は示している。オーバーフィッティングは、候補の数が少数でも生じ、候補数の増加とともに、その程度は大きくなる。複数の候補の中から選択する際には、多重検定を考慮する必要があり、乱数シミュレーションによって容易に検証することができる。

## 5. リターン予測の困難性

なぜ、実証的にリターンの予測可能性を検出することは困難なのだろうか。最後にこの点を簡単に考察する。確率過程であるリターン  $r_t$  の実現値とその事前の条件付き期待値  $\mu_t := E[r_t | I_{t-1}]$  の間には

$$r_t = \mu_t + \varepsilon_t$$

の関係がある。 $I_{t-1}$  は時点  $t-1$  において利用可能な情報を表す。 $\varepsilon_t$  は時点  $t-1$  では予測することができない成分で、時点  $t-1$  から  $t$  までの間に到達する情

報がもたらすショックを表す。

資産の期待リターン  $\mu_t$  は、合理的な要因（たとえば投資家のリスク回避度の変化）や非合理的な要因（たとえば投資家のセンチメントの変化）、その他の摩擦（たとえば情報が価格に織り込まれるまでのタイムラグ）から、時間を通じて変動しているものと考えられている。しかし、決定係数  $R^2 = \text{Var}(\mu_t)/\text{Var}(r_t) = 1 - \text{Var}(\varepsilon_t)/\text{Var}(r_t)$  はかなり低いことが想定される。真の  $R^2$  を見積もることは難しいが、たとえば Ross [13](p. 56) は、理論的な考察にもとづき、 $R^2$  の上限が 10% 程度（株式市場を対象にした月次データの場合）であろうことを論じている。これは上限であって、実際には 10% よりもずっと低いかもしれない。すなわち、リターンの変動の大部分は予測不可能な成分であると考えられる。データマイニングによって、高い決定係数が標本から得られたとしても、それはかなり怪しい。

ビッグデータや機械学習を用いてリターンを予測することは、選択された予測変数  $Z_{t-1}$  と、選択されたモデル  $g(\cdot)$  によって、 $\mu_t \approx g(Z_{t-1})$  として期待リターンを近似できることを本来は意図している。しかし、リターンの  $R^2$  は本質的に小さく、標本は有限個しかないため（金融の時系列データはさほどの標本数ではない）、変数選択とモデル選択を繰り返し行えば、標本にもとづく  $g(Z_{t-1})$  の値が  $\varepsilon_t$  に偶然に適合してしまうものを見つけてしまうことは十分に起こりうる。

さらには、分析者の動機もまた事態を悪化させる。金融市場に対するわれわれの理解を深めるには、データを念入りに調べる必要があるのは事実である。データマイニングは本質的に不可避なうえ、都合の良い有意な証拠を見つけたという動機が研究者にはある（p ハッキングと呼ばれる）。

## 6. 結論

資産価格のプレミアム（期待リターン）は時系列に変動（時間変動）し、リターンが予測可能であることは現在のファイナンスにおける標準的な考え方である。しかし、統計的にこれを検出することは容易でないため、より有意な実証的証拠を得ようとデータマイニングを行うと、実際には無意味であるにもかかわらず有意に見えてしまうというオーバーフィッティング（過剰適合）を引き起こす。特に近年では、ビッグデータとして多様なデータを低コストで扱えるようになり、また、工学的な側面から機械学習への注目度が増している。これらは予測精度の向上に貢献する可能性がある一方で、オーバーフィッティングの可能性を高めてし

まう。

本稿では、変数選択のマイニングだけでなく、モデル選択のマイニングについても考察した。この結果、時系列におけるリターンの予測可能性を対象に、オーバーフィッティングの影響が大きいことを実証的に示した。変数選択やモデル選択に伴う多重検定を考慮すると、 $t$  値の分布は大幅に上方にシフトし、有意水準の臨界値は極めて高くなる。本稿の結果は、ファイナンス研究がデータサイエンスの技術も活用し、人々の意思決定の改善に役立つことを通じて社会に貢献していくためには、オーバーフィッティングの問題を考慮する必要があることを示している。

謝辞 内山による本研究の一部は、JSPS 科研費 JP18K01691 の助成を受けている。

## 参考文献

- [1] J. H. Cochrane, “Presidential address: Discount rates,” *Journal of Finance*, **66**, pp. 1047–1108, 2011.
- [2] E. F. Fama and K. R. French, “The cross-section of expected stock returns,” *Journal of Finance*, **47**, pp. 427–465, 1992.
- [3] J. Y. Campbell and R. J. Shiller, “The dividend-price ratio and expectations of future dividends and discount factors,” *Review of Financial Studies*, **1**, pp. 195–228, 1988.
- [4] C. R. Harvey, Y. Liu and H. Zhu, “... and the cross-section of expected returns,” *Review of Financial Studies*, **29**, pp. 5–68, 2016.
- [5] MathWorks, 『機械学習のご紹介』, The MathWorks, Inc., 2016.
- [6] J. H. Cochrane, “The dog that did not bark: A defense of return predictability,” *Review of Financial Studies*, **21**, pp. 1533–1575, 2008.
- [7] A. W. Lo and A. C. MacKinlay, “Data-snooping biases in tests of financial asset pricing models,” *Review of Financial Studies*, **3**, pp. 431–467, 1990.
- [8] T. E. Clark and K. D. West, “Approximately normal tests for equal predictive accuracy in nested models,” *Journal of Econometrics*, **138**, pp. 291–311, 2007.
- [9] R. D. McLean and J. Pontiff, “Does academic research destroy stock return predictability?,” *Journal of Finance*, **71**, pp. 5–32, 2016.
- [10] R. Novy-Marx, “Testing strategies based on multiple signals,” Working Paper, University of Rochester, 2016.
- [11] X. S. Yan and L. Zheng, “Fundamental analysis and the cross-section of stock returns: A data-mining approach,” *Review of Financial Studies*, **30**, pp. 1382–1423, 2017.
- [12] T. E. Clark and M. W. McCracken, “Reality checks and comparisons of nested predictive models,” *Journal of Business & Economic Statistics*, **30**, pp. 53–66, 2012.
- [13] S. A. Ross, *Neoclassical Finance*, Princeton University Press, 2005.