経営課題にAIを!ビジネスインフォマティクス研究会 (第8回)
JSAI Special Interest Group on Business Informatics (SIG-BI #8)

危険なデータマイニング
— リターン予測とオーバーフィッティング —
Dangers of Data-Mining: Return Predictability and Overfitting

内山 朋規 [1*]　瀧澤 秀明 [2]　菊川 匡 [2]
Tomonori Uchiyama[1]　Hideaki Takizawa[2]　Tadashi Kikugawa[2]

[1] 首都大学東京
[1] Tokyo Metropolitan University
[2] 野村證券
[2] Nomura Securities Co., Ltd.

Abstract: Standard finance theory states that returns on assets are predictable. However finding empirical evidence of predictability is statistically difficult, and data-mining for detecting more significant evidence leads to overfitting, by which it looks like significant in appearance but is senseless in fact. Particularly in recent years, an enormous amount of information or big data becomes available at low cost, and machine learning attracts an increasing interest in engineering aspects. Big data and machine learning can contribute to improvement in prediction accuracy, while they increase possibility of overfitting. This study considers data-mining with both variable selection and model selection for return predictability in the time-series. Our results demonstrate that overfitting makes large degree of influence on backtests, giving rise to specious identifications.

1　はじめに

資産価格は将来キャッシュフローの期待値をプレミアム（言い換えれば期待リターン，資本コスト）で割り引いた現在価値として定まることから，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いる．プレミアムはクロスセクションで異なり，かつ時系列に変動し，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（例えば Cochrane (2011) 参照）．株式市場を例にとると，純資産株価倍率 (Fama and French, 1992) や配当利回り (Campbell and Shiller, 1988) 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．

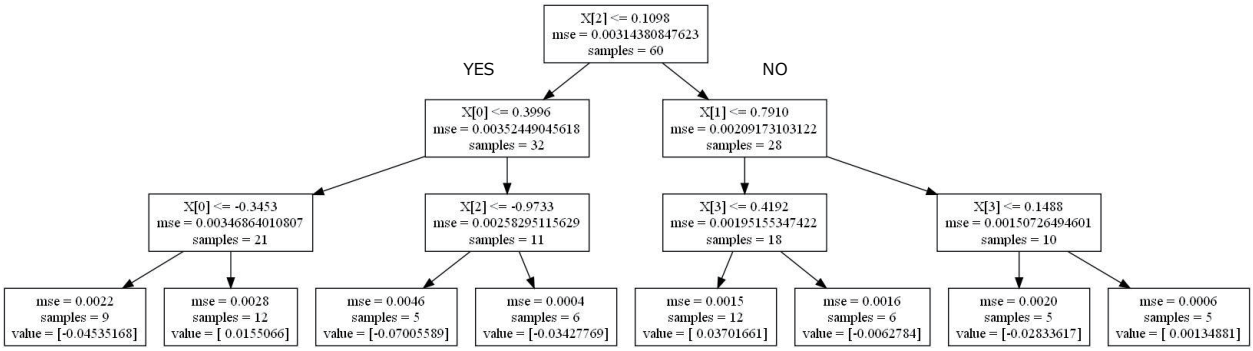
プレミアムの実証的な特徴を満足に記述できるモデルはいまだ存在せず，多くの研究者が様々な資産（株式，債券，クレジット，ボラティリティ，通貨，コモディティなど）を対象に，クロスセクションあるいは時系列における予測変数（ファクター）の研究を競い合い，新たな予測変数を「発見した」とする報告が続いている（筆者らも含まれる）．例えば Harvey, Liu and Zhu (2016) によれば，米国株式市場におけるクロスセクションの予測変数として，これまでトップジャーナルに「発見」が報告されたファクターは 316 個もあり，「発見」される頻度は年々高まっている．他の資産や時系列の予測変数に関しても同様で，2011 年の米国ファイナンス学会会長講演 (Cochrane, 2011) 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ファクターを見つけることは，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始めている．

さらに近年では，機会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

* 連絡先：首都大学東京大学院社会科学研究科経営学専攻
〒100-0005 東京都千代田区丸の内 1-4-1
E-mail: tomonori-uchiyama@tmu.ac.jp

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] は鉱工業生産を表す．

われる場面が見受けられる[1]．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，例えば今日では Twitter 投稿のテキスト分析や，人工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，最近傍法やサポートベクターマシン，ニューラルネットワーク，決定木，回帰木，ランダムフォレストといったモデルも利用されている．情報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も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）という忌忌しき事態を招く．

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（例えば Cochrane (2008) 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 (Lo and MacKinlay, 1990)．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．通常の単一検定の基準ではなく，多重検定であることを考慮して有意性を評価する必要がある．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．例えば，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 ヵ月間の 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

[1] 例えば，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, 2016) 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を助けてくれます．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．

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 $\overline{r}_t := E_{t-1}^{\text{benchmark}}[r_t]$，上記の 4 種類のマクロ変数を用いた回帰木による予測値を $\widehat{r}_t := E_{t-1}^{\text{model } X}[r_t]$ とおく．ベンチマークの予測誤差を $\overline{u}_t := \overline{r}_t - r_t$，モデル X の予測誤差を $\widehat{u}_t := \widehat{r}_t - r_t$ 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 $\widehat{u}_t^2$ の期待値は，$\overline{u}_t^2$ 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，$\overline{u}_t^2$ と $\widehat{u}_t^2$ 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 $\overline{u}_t^2 - \widehat{u}_t^2$ の標本平均はゼロではなく，負になるバイアスを持つ．そこで，Clark and West (2007) 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．

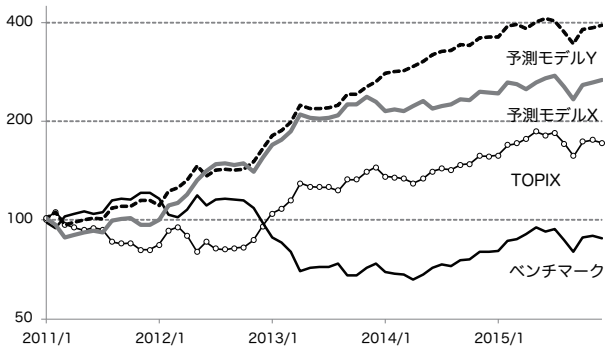
$$f_t := \overline{u}_t^2 - \widehat{u}_t^2 + (\overline{u}_t - \widehat{u}_t)^2 \tag{1}$$

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 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分かる．

図 2: TOPIX リターンの予測



(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前月末における 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．

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して，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 (2016) 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していることを報告し，この原因が投資家にファクターが知れ渡ったことによる影響か，あるいは単なるデーマイニングの結果に過ぎなかなったことを論じている．

また，Harvey, Liu and Zhu (2016) 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，2012 年までに代表的なジャーナルに掲載された 316 個のファクターを対象にしている．論文に

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*	ユーロストックス	米製造業 ISM*	米消費者信頼感・先行き*	鉱工業生産

(注) *は前月比，他は前月差を表す．

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[2]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 (2016) 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にティルトした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．多くのファクターを試せば，本来は予測力がなくても，実証分析における標本数は必然的に有限なので，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で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 (2016) は，正規乱数（乱数に予測力はない）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 値の分布をシミュレーションによって算出している．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 値が最大のものを 1 つ選ぶのに匹敵することを示している．

一方，Yan and Zheng (2017) は，実際のファクターに着目している．例えば 1 章で述べた純資産株価倍率は代表的な財務変数の予測変数で，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 (2017) 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 (2017) は，財務変数がもたらす実際のアルファの 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 章で述べたモデル X に話しを戻そう．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

[2] クロスセクションにおいては，超過リターンの平均が有意にゼロとは異なるかどうかが検定されることから，両側検定による．

指標など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[3]の 3 種類の予測モデルの中から，(1) 式で定義された 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ても一つを用いると，予測変数の組み合わせは $2^{50}-1$ 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，予測モデル 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 3 つのモデルから「ベストフット」するものを採用するのは，近年の機械学習への注目に対応している．

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）の的中率を「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[4]．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．

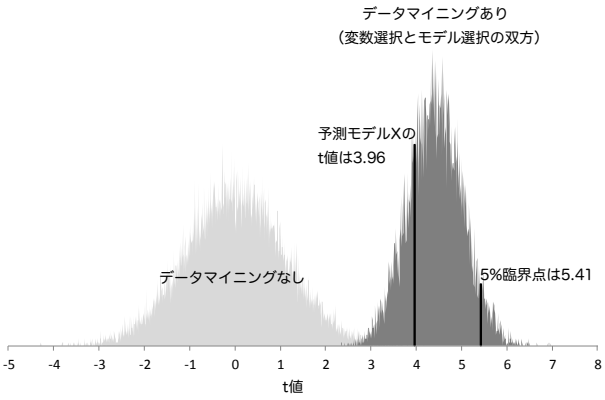
4.2 データマイニングの影響

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での 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．

なお，50 個のマクロ変数の相関を考慮した正規乱数を用いているが，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 (2012) 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容易に検証できるという利点がある．

この結果を表したのが図 3 である．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．

図 3: t 値の分布



（表）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) 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．

[3] 回帰木では，木の深さが 3，各枝先端の最小サンプル数が 5 になるようにパラメータを設定した．

[4] 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．

当然のことながら，3 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,「データマイニングあり」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，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 の t 値は 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はずである．

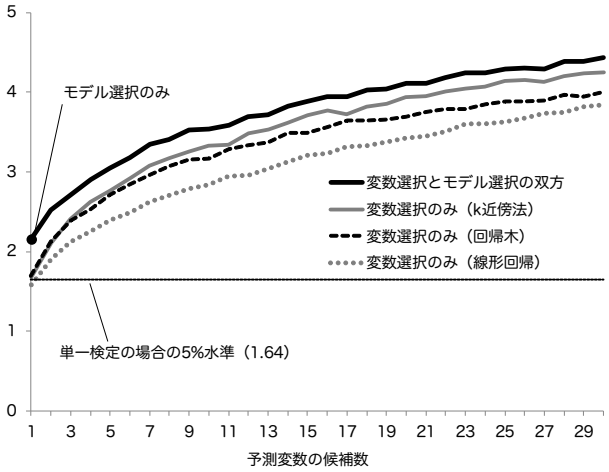
図 3 の表には，モデル選択のみを行った場合や，変数選択のみを行った場合の結果についても記載した．モデル選択のみは，50 組の予測変数（乱数）をそのまま用いて，3 通りのモデルから t 値が最も高いものを選択したモデルマイニングの結果を表す．t 値の 5% 点は 2.17 で，Bonferroni 法による有意水準の補正 0.05/3 に対応する t 値の 2.13 にほぼ等しい．また，モデル選択を行わずに，変数のみをマイニングしても シミュレーションの t 値は相当に上昇することが分かる．

これまでは 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 1 つずつ増や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分かる．上昇の程度はモデルによって異なり，例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うと，当然のことながら，5% 点の t 値はさらに高まる．例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，変数選択のその影響は大きく，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．加えて併せてモデル選択も行うと，さらに影響は大きくなる．

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．

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」として認識することがしばしばある．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あることを本稿の結果は示唆している．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，乱数シュミレーションによって容易に検証することができる．

5 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ゆくためには，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唆している．

参考文献

Campbell, J. Y. and R. J. Shiller (1988) "The Dividend-Price Ratio and Expectations of Future Dividends and Discount Factors," *Review of Financial Studies*, 1(3), 195–228.

Clark, T. E. and M. W. McCracken (2012) "Reality Checks and Comparisons of Nested Predictive Models," *Journal of Business & Economic Statistics*, 30(1), 53–66.

Clark, T. E. and K. D. West (2007) "Approximately Normal Tests for Equal Predictive Accuracy in Nested Models," *Journal of Econometrics*, 138(1), 291–311.

Cochrane, J. H. (2008) "The Dog That Did Not Bark: A Defense of Return Predictability," *Review of Financial Studies*, 21(4), 1533–1575.

——— (2011) "Presidential Address: Discount Rates," *Journal of Finance*, 66(4), 1047–1108.

Fama, E. F. and K. R. French (1992) "The Cross-Section of Expected Stock Returns," *Journal of Finance*, 47(2), 427–465.

Harvey, C. R., Y. Liu, and H. Zhu (2016) "... and the Cross-Section of Expected Returns," *Review of Financial Studies*, 29(1), 5–68.

Lo, A. W. and A. C. MacKinlay (1990) "Data-Snooping Biases in Tests of Financial Asset Pricing Models," *Review of Financial Studies*, 3(3), 431–467.

MathWorks (2016) 『機械学習のご紹介』, The MathWorks, Inc.

McLean, R. D. and J. Pontiff (2016) "Does Academic Research Destroy Stock Return Predictability?," *Journal of Finance*, 71(1), 5–32.

Novy-Marx, R. (2016) "Testing Strategies Based on Multiple Signals," Working Paper.

Yan, X. S. and L. Zheng (2017) "Fundamental Analysis and the Cross-Section of Stock Returns: A Data-Mining Approach," *Review of Financial Studies*, 30(4), 1382–1423.