機械学習による有価証券保有額、確定拠出年金への加入の分析
～ 性別を処置変数としたCausal Treeによる解析結果 ～※

大　野　裕　之
林　田　　実ξ

目次

１．はじめに

「貯蓄から投資へ」が言われて久しい。しかし、1996年の金融ビッグバン以来、家計の証券投資を政策面で後押しするために制度改革が進められてきているが、その成果は十分表れているとは言い難い[1)]。我が国家計の金融資産の構成は、諸外国と比べて依然として預貯金偏重である。2016年９月末時点で、個人金融資産に占める「株式・債券・投信」の割合は、米国が35.8%、ドイツが19.4%であるのに対し、我が国は11.3%という水準である。反対に、「現金・預金」の割合は、米国が13.9%、ドイツが39.1%であるのに対し、我が国は52.3%もの高水準にある[2)]。

このように、家計の証券投資が進んでいないのは何故か。それを探るべく、これまで様々な政策・制度変更に焦点を当てた実証分析が行われてきている。その中でも、アンケート調査の個票データ

※ 本研究は、2020年度井上円了記念研究助成金を得て行われた。

ξ 北九州市立大学経済学部教授

1) 例えば、税制に焦点を当ててみると、1996年４月に有価証券取引税の軽減、1999年４月に廃止、2003年１月に株式譲渡益課税の簡素化・軽減税率の暫定適用、2003年４月の配当課税の簡素化・軽減税率の暫定適用、2004年１月の株式投資信託税制の簡素化などが行われている。

2) 日本証券業協会調べ。「試算の形成・円滑な世代間移転と税制の関係に関する研究会」（2017）p45参照。

を用いた研究では、個人属性などを説明変数にとり、政策効果を表す目的変数への影響を、質的変数に関する推定法[3]で追うことが多い。しかしながら、これらの手法では、属性ごとの政策効果はターゲットにしておらず、仮にターゲットにしたとしても、技術的な限界がある。例をあげて説明しよう。説明変数の候補としてカテゴリカル変数が二つ、X_1とX_2があるとしよう。この場合、それぞれが単独で推定式に入ると同時に、両者の交差項を考えることが多い。両者ともに0もしくは1をとるいわゆるダミー変数の場合には、（X_1・X_2）という交差項が1つ増えて説明変数が合計3つになるだけである[4]。しかし、X_1とX_2がそれぞれn_1とn_2個の選択肢からなっている場合、交差項の数は（n_1-1）×（n_2-1）個にまで増えてしまう。X_3、X_4・・・と説明変数が3以上に増える場合、追加の説明変数候補となる交差項の数はまさに指数的に増える。この場合、データ数を瞬く間に費消してしまうだけでなく、これを人力で行うのは極めて困難である。

ところが昨今、汎用性を増している機械学習の手法を使えば、このような分析も可能となる。詳細は次節に譲るが、そのような機械学習の中でも最近、Athey and Imbens（2016）らによるCausal Treeという斬新な分析手法が、さまざまな研究で急速に使用されるようになっている[5]。そこで本研究は、このCausal Treeを用いて個人の貯蓄投資行動を分析することとする。具体的には、その中でも有価証券保有額と確定拠出年金への加入・非加入を目的変数に据える。本稿の目的が、証券投資が進まない理由を探ることであることから、有価証券保有額を目的変数に据えることには何の異論もなかろう。ここで確定拠出年金の加入の有無を目的変数に据えるのは、以下の理由による。確定拠出年金は有価証券と同じく、老後の貯えとして用いられる金融資産で、制度も比較的新しく、公的年金の将来不安が増す中、おおきな注目を集めているからである。

Causal Treeは、処置効果分析を行う手法のひとつである。実験においては、処置変数（treatment variable）をコントロールした上で結果変数を観測するので、何を処置変数とするかは実験前にすでに決まっていることが多い。しかし、経済現象の分析に用いられるデータは事後的に観測されるものであるから、何を処置変数と考えるかは重要である。本稿では、この処置変数として性別を据える。我が国では一般に性別による投資行動の相違が存在すると考えられる。そこで、性別による投資行動の違いをCausal Treeで分析するとどのような結果が得られるかに関心をもった。もとよ

---

3) プロビットモデル、順序プロビットモデル、多項ロジットモデルなどが代表的な手法である。

4) 例えば、X_1とX_2が性別（男性0で女性1）と高齢者か否か（高齢者0で非高齢者1）とすると、両者の交差項を加えることで、男性高齢者と男性非高齢者で目的変数Yに与える影響が異なる可能性を追求できる。

5) Causal Treeをアンサンブル学習に拡張したCausal Forestによる解析例の方が優勢かもしれない。Causal ForestはCausal Treeよりもより良い予測結果をもたらすことが指摘されているが、どの変数がどのように影響をあたえているのかを調べるのには適していない。また、アンサンブル学習についての知識も必要となる。このような理由で、Causal Treeの紹介もかねて本稿を執筆した。

り、それ以外にも重要な変数は多数存在するが、本稿でそのすべてを取り上げるのは不可能である。そこで、手始めに性別を処置変数に採用し、他の重要な変数は後続の研究に譲ることとしたい。例えば、分析の結果、女性に顕著な特徴が示唆されれば、それを踏まえた政策提言も可能になる。用いるデータは、日本証券業協会が毎年実施している『個人投資家の証券投資に関する意識調査』（以下『調査』）の個票データである。

本稿の構成は以下のとおりである。次節では、機械学習とCausal Tree を、数値例を用いながら紹介する。第３節は、『調査』と用いた変数を説明する。第４節は2016年の『調査』を用いて、性別が有価証券保有額と確定拠出年金への加入・非加入に与える影響を解析する。最終の第５節は本稿のまとめであり、今回の研究の不足点をあげて、後続の研究を展望する。

２．機械学習とCausal Tree

２．１　処置効果

簡単な例を用いて、処置効果とは何かを解説しておこう。太郎君があるダイエット法を用いて一月生活したあとの体重をＹとする。我々は、このダイエット法が真に効果があるか否か、あるいは、どの程度効果があるかに関心を寄せる。この時、処置効果とは、以下のように定義される。

$$\text{処置効果} \equiv Y(1)\text{-}Y(0)$$

ここで、$Y(1)$は、このダイエット法を用いた時の体重であり、$Y(0)$ はダイエットをしないで普通に生活した時の体重である。したがって、今の例では、$Y(0)$ は観測されないことに注意されたい。さらに、太郎君には、様々な属性xがあるので、その属性を考慮したときの、そのダイエット法の条件付平均処置効果（Conditional Average Treatment Effect）を考える必要があろう。条件付平均処置効果は次のように定義される。

$$\tau(x) \equiv E\ [Y(1)\text{-}Y(0)|X_i{=}x]$$

Causal Treeはこの条件付平均処置効果を回帰木[6)]を用いて推定する手法であり、その発表後、その解析の切れ味が評判を呼び多数の実証例が報告されるようになった。

２．２　Causal Tree [7)]

結果変数Y、処置変数W、説明変数X（Ｋ次元）とし、N個の観測値があるとする。すなわち、次のようなデータを想定する。

---

6)　回帰木については巻末の補論を参照せよ。

7)　本節は、Ashley and Imbens（2016）および次のサイトに大きく依存している。
https://github.com/susanathey/causalTree

$$(Y_i^{obs}, W_i, \mathrm{X}_i), i=1, \cdots, N.$$

W_iはi番目の観測値が処置されていれば 1 を、処置されていなければ 0 をとる変数である。また、Y_i^{obs}は、

$$Y_i^{obs}=Y(W_i)=\begin{cases}Y_i\ (0)\ if\ W_i=0,\\ Y_i\ (1)\ if\ W_i=1.\end{cases}$$

である。ここで、p=pr（W_i=1）は周辺処置確率（marginal treatment probability）、$e(x)$=pr(W_i=1|X_i=x)はプロペンシティスコアとして、広く知られている。

条件付平均処置効果が推定できるための条件として、以下のUnconfoundednessが成立しているとする。

$$W_i \perp\ (Y_i(1)-Y_i\ (0))|X_i$$

これは、X_iが条件付けられれば，処置W_iと処置効果（$Y_i(1)$-$Y_i(0)$）とが直交していることを意味している。

以上のようなデータと条件の下で、Causal Treeは回帰木における分類基準である誤差平方和を以下のものに置き換えたものである。

$$\frac{1}{N^{tr}}\sum_{i\in S^{tr}}\hat{\tau}^2(X_i; S^{tr}, \Pi)-\left(\frac{1}{N^{tr}}+\frac{1}{N^{est}}\right)\sum_{l\in\Pi}\left(\frac{S^{2}_{S^{tr}_{treat}}(l)}{p}+\frac{S^{2}_{S^{tr}_{control}}(l)}{1-p}\right)$$

ただし、符号を逆転しているので、この量が大きいほど、その分類が良いと判断される。ここで、それぞれの記号の意味は以下のとおりである。

Π：回帰木。$\Pi=\{l_1, \cdots, l_{\#(\Pi)}\}$。$l_j$は$j$番目のリーフ。$\#(\Pi)$は回帰木のリーフ数。

$l(x; \Pi)$：$x\in l$であるような、回帰木Πの要素。

$\tau(l)$：リーフlにおける、処置効果。これは、リーフ内の処置されたデータの結果変数の平均からリーフ内の処置されていない結果変数の平均を引いたもので推定される。

$S^{2}_{S^{tr}_{treat}}(l)$：リーフ内の処置されたデータの結果変数の標本分散

$S^{2}_{S^{tr}_{control}}(l)$：リーフ内の処置されていないデータの結果変数の標本分散

N^{tr}：トレーニングデータの観測数

N^{est}：デフォルトではN^{tr}と同じ。完成した回帰木を別途estimation dataで再計算する場合には、このestimation dataのデータ数。

S^{tr}：トレーニングデータ

3．データと変数

3．1　データの概要

本研究で用いる『調査』は2006年より、日本証券業協会が毎年６～７月に実施し、９～11月に結果を公表しているアンケート調査である。対象者は証券投資を行っている全国の、満20歳から89歳までの投資家である[8)]。毎年の調査対象者数は2015年調査までは2,100～2,300人で、回答率は約50％、回答者数はおおよそ1,000人、2016年調査では調査対象者数5,000人、回答者数は2,024人である。ここまでは郵送調査であったが、2017年以降はインターネット調査に切り替えて、5,000程度の回答者数を確保している[9)]。質問項目は、年齢、性別などのフェース項目の他、株式、株式投信、債券など証券投資に関して多岐にわたっている。ただし、毎年、質問項目が多少変わっていることには注意を要する[10)]。

本研究では、2016年のデータを用いる。先述したように、有価証券保有額を目的変数にしたCausal Treeと、確定拠出年金への加入・非加入を目的変数としたCausal Treeとを作成した。また、説明変数としては、世帯所得、個人所得、年齢をとり、処置変数としては性別を採用した。機械学習の特性としては、多くの説明変数が存在する状況が有利であるが、従来の研究成果を踏まえ、かつ、Causal Treeの挙動を解析するということも本研究の目的であることからそのようにした。

3．2　目的変数

①有価証券保有額

目的変数のひとつである有価証券保有額は、投資家対象の調査では以下の２問を用いている。2016年『調査』問１で指定した１～５の５つの金融商品の合計額を、問３で10段階のカテゴリーで尋ねている。

> **問１　次の金融商品（又は取引）のうち、あなたが現在保有（又は取引）しているものをお答えください。ご回答に当たってはページ上部の【主な金融商品・取引一覧表】をご参照ください。（いくつでも）【n=2,024】**
>
> **１　預貯金（普通預金、当座預金や定期預金など）　２　信託　３　株式　４　投資信託　５　公社債　６　有価証券関連デリバティブ取引　７　有価証券関連デリバティブ取引以外のデリバティブ取引　８　その他　９　いずれも持っていない（及び、行っていない）**

---

8)　ただし、2014年については投資未経験者を対象とした調査も同時に行っている

9)　調査方法の切り替えは、回答者の属性に影響を与える可能性があり、切り替え年を跨いで分析を行う際には注意が必要である。典型的なものが年齢である。2015年と2016年の調査で、20～30代はそれぞれ９％、８％であるが、2017年は10.1％、2018年以降は12.6％～14％へと上昇している。

10)　詳細は日本証券業協会のホームページを参照のこと。

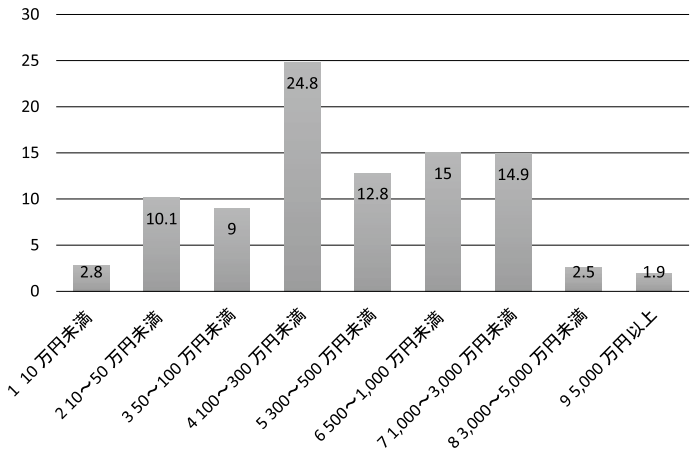
【問１で「１」～「５」のいずれかに○をつけた方へお伺いします。】

問３ そのうち、有価証券（問１の「３」～「５」）の保有額はいくら位ですか。時価で計算してください。（１つだけ）【n=2,024】

１　10万円未満　２　10～50万円未満　３　50～100万円未満　４　100～300万円未満　５　300～500万円未満　６　500～1,000万円未満　７　1,000～3,000万円未満　８　3,000～5,000万円未満　９　5,000万円以上　10　有価証券は保有していない

この変数の分布は以下の図１のとおりである。100～300万円未満のクラスが全体の24.8％と最も多く、次いで「500～1,000万円未満」が15.0％、「1,000～3,000万円未満」が14.9％で続く。「5,000万円以上」の高額投資家も1.9％おり、100万円未満の少額投資家は2.8％となっている。

図１：有価証券保有額（単位：％）



注）『個人投資家の証券投資に関する意識調査』2016年版問３の集計結果より筆者作成。N=2,204。合計が100％にならないのは、無回答者（6.4%）がいるため。尚、「10　有価証券は保有していない」を選択した回答者は０人であった。

②確定拠出年金への加入・非加入

いまひとつの目的変数である確定拠出年金への加入・非加入には、以下の問を用いた。2016年『調査』では問41である。

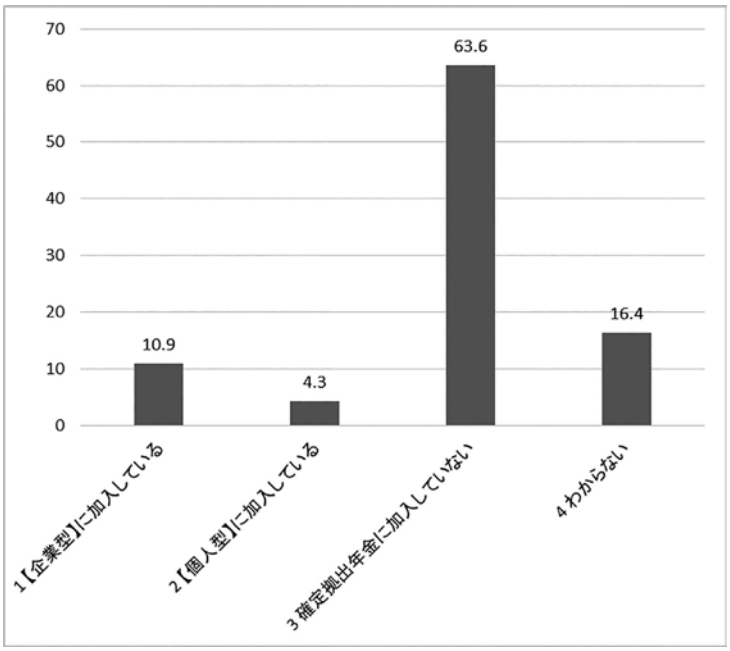
問41　あなたは現在、確定拠出年金に加入していますか。（１つだけ）

１　【企業型】に加入している　２　【個人型】に加入している　３　確定拠出年金に加入していない　４ わからない

この問いに対する回答から、選択肢１もしくは２を選んだ人には１、選択肢３を選んだ人には０を充てる二値変数を構築した。なお、選択肢４を選んだ人は無回答者とともに、分析から除外した。

この変数の分布は以下の図２のとおりである。最も多いのが「加入していない」で全体の63％もある。投資家でも２/３近い人が加入していない。加入している人は企業型で10.9％、個人型では4.3％にとどまる。これは、企業型の制度を持っている企業が少なかったり、企業型、個人型ともに制度への理解が進んでなかったりするためであろう。投資家を対象とした『調査』でも、「わからない」が16.4％もいることからも、そのことが窺える[11)]。

図２：確定拠出年金への加入・非加入



注）『個人投資家の証券投資に関する意識調査』2016年版問41の集計結果より筆者作成。N=2,024。合計が100％にならないのは、無回答者（4.8％）がいるため。

３．３　処置変数

本研究では、処置変数に性別をとる。2016年『調査』ではＦ２の問いである。無論、性別以外にも興味深い、処置変数の候補はいくつも考えられるが、その全てを本稿で取り上げることはできないので、今回は手始めに性別を処置変数とした。尚、性別は外生性が最も明確なので、処置変数と

11）他に無回答者が4.8％もいるので、両者を合計すると21.2％にも上る。無回答者の多くはわからないので、面倒なため回答そのものをしなかった人も多いと推測される。

して適当な候補と言える[12)]。

性別の分布は以下の表１のとおりである。男性が57.4%、女性が42.6%と、やや男性に偏った分布になっているが、これは『調査』が投資家を対象としており、投資家の中では男性の方が女性より多いと考えられるので、こうした分布になっているのであろう。

表１：処置変数（性別）の分布

	割合（単位：%）
男性	57.4
女性	42.6
合計	100

注）N=2,024。

３．４　説明変数

説明変数は、これまでの研究の蓄積を参考に、(a) 年齢、(b) 世帯収入、(c) 個人収入を採用した。それぞれ、2016年『調査』では、Ｆ２、Ｆ５、Ｆ７の問いである。このうち、(a) 年齢は以下の11段階の選択肢からなる変数である。これはageと表記する。

1　20～24歳　2　25～29歳　3　30～34歳　4　35～39歳　5　40～44歳　6　45～49歳　7　50～54歳　8　55～59歳　9　60～64歳　10　65～69歳　11　70歳以上

(b) 世帯収入と (d) 個人収入は、以下の８段階の選択肢からなる変数である。それぞれ、revenue_h、revenueと表記する。

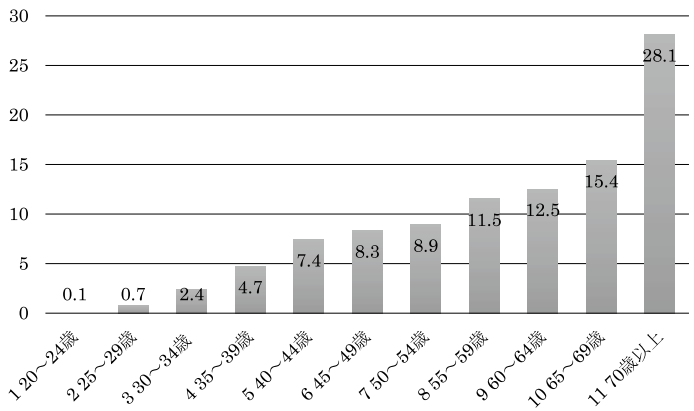
1　300万円未満　2　300万円～500万円未満　3　500万円～700万円未満　4　700万円～1,000万円未満　5　1,000万円～1,200万円未満　6　1,200万円～1,500万円未満　7　1,500万円～2,000万円未満　8　2,000万円以上

説明変数の分布を下の図３～５に示した。まず、図３はageすなわち年齢の分布である。調査対象者は高齢者に大きく偏っていることがわかる。40歳未満は7.9%しかいないのに対し、70歳以上はその３倍以上の28.1%に及ぶ。これはそもそも、証券投資家は若年者よりも高齢者が圧倒的に多いことに由来していると思われる。

---

12) このCausal Treeが典型的に適用される用例は、投薬の効果の解析である。投薬を受ける患者群、受けない患者群は外生的に決まる。その意味で性別は処置変数として適切な選択である。一方で、投薬の効果を解析する目的は、薬の効能を調べるという政策的なものであるから、その点では、例えば制度変更を処置変数とすることがよいかもしれない。但し、その場合は制度変更前後で回答者が同じである、パネルデータの利用が必要になる。

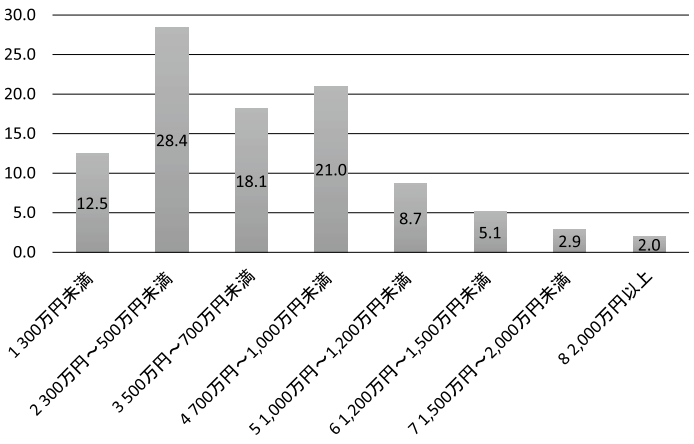
図３：説明変数ageの分布



注）『個人投資家の証券投資に関する意識調査』2016年版Ｆ２の集計結果より筆者作成。N=2,024。

図４はrevenue_hすなわち世帯収入の分布である。「300万円から500万円未満」が28.4%と最も多いことは、やや意外である。次に、「700万円～1,000万円未満」が最も多く、「300万円から500万円未満」を除くと、非対称ながら、概ねそこから左右になだらかに数が少なくなる形状を示す。

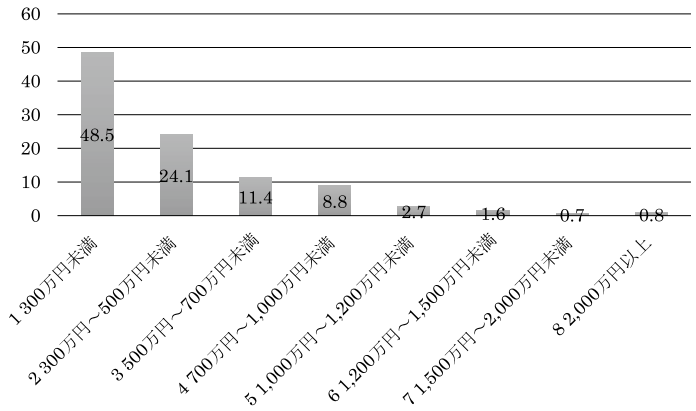
図４：説明変数revenue_h（世帯収入）の分布



注）『個人投資家の証券投資に関する意識調査』2016年版Ｆ７の集計結果より筆者作成。N=2,024。合計が100％にならないのは、無回答者（1.2%）がいるため。

最後に図５は、revenueすなわち個人収入の分布である。「300万未満」がほぼ半数の48.5％を占め、収入額が上がるにしたがって数が単調に少なくなる形状をとっていて、図４のような形状にはなっていない。

図５：説明変数revenue（個人収入）の分布



注）『個人投資家の証券投資に関する意識調査』2016年版F５の集計結果より筆者作成。N=2,024。合計が100％にならないのは、無回答者（2.4％）がいるため。

４．性別を処置変数にとった解析結果[13)]

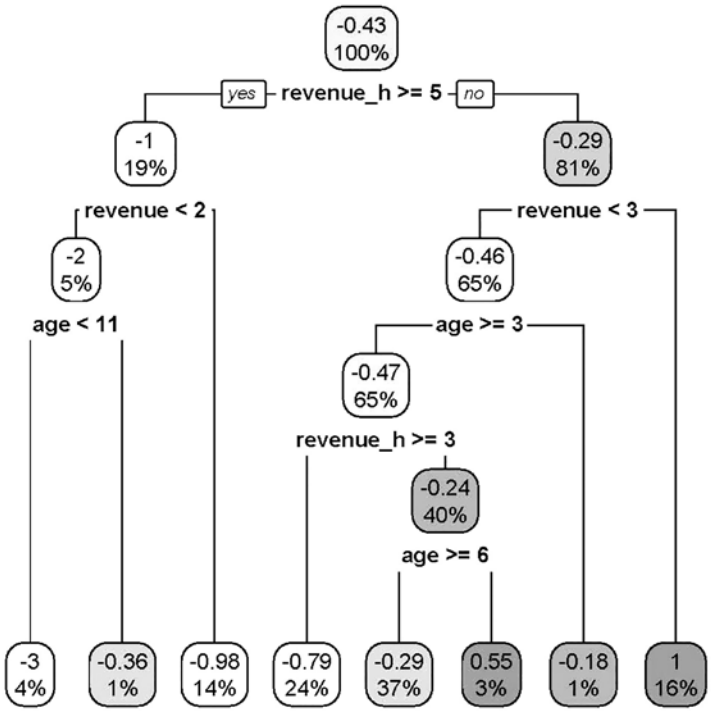
本節では、『調査』2016年版のデータを用い、処置変数を性別とした場合の結果を報告する。図６および７は、統計解析向けのプログラミング言語であるＲのライブラリ“Causal Tree”によって描かれている。

①有価証券保有額を目的変数にとった結果

図６が、有価証券保有額に関する分析結果である。ルートと称する最上段の数値は－0.43と負の値であるから、女性は男性に比して有価証券保有総額の回答が0.43カテゴリー低いことを意味している。しかし、この結果は粗すぎることが、ルートから下にいくつも枝分かれをしていることからわかる。ルートのすぐ下に、「revenue_h>=5」とあり、その左にyes、右にnoの文字がある。「revenue_h>=５」に該当すれば左、該当しなければ右の枝に進むことを意味している。左に進んだ場合、すなわち世帯所得が選択肢５以上を選んだ女性、つまり世帯年収が1,000万円以上の女性の場合、「revenue<2」という枝分かれに突き当たる。yesつまり、個人収入が300万円未満の女性は、その下の「age<11」という枝分かれに突き当たり、yesを選ぶ69歳以下の女性は、最下段の「－３　４％」という囲みに至る。この意味は、世帯年収が1,000万円以上で個人年収が300万円未満の69歳以下のデータは全体の４％を占め、同じ分類の男性に比して、有価証券保有総額が３カテゴリー低いことを意味している。また、最下段の最も右の「１　16％」は、世帯収入が５未満、つまり1,000万円

13) 本稿の解析では乱数のシードとして、set. seed（1）を用いている。

図６：有価証券保有額を目的変数にとった結果（Causal Tree）



未満でかつ個人収入が３以上、つまり500万円以上のデータは全体の16％を占め、同じ分類の男性に比して、有価証券保有総額が１カテゴリー高い、ということになる。

このようにひとつひとつ解釈を施していくと、Causal Treeの結果は以下の(ⅰ)～㈧のようにまとめることができる。尚、ここで報告している統計的有意度の求め方は、以下のようなものである。最下段には８つの分類が描かれている。この分類のことをCausal Treeのterminologyでリーフ（leaf）とよぶことが通例なので、以下でもこれに従う。最下段の各リーフについて、そこに至るまでの枝分かれの条件を全て満たす場合に１、それ以外は０をとるダミー変数を、それぞれのリーフについて作成する。次いで、その「リーフ・ダミー変数」と、処置変数の条件を満たしている場合には１、それ以外には０をとる「処置変数ダミー変数」との交差項を、それぞれのリーフについて作成する。最後に、目的変数を、すべての「リーフ・ダミー変数」と「処置変数ダミー変数」との交差項で回

帰し、そのＰ値で統計的有意度を判定する[14)15)]。

(i)［最下段の最も左］世帯収入が1,000万円以上で個人収入が300万円未満、69歳以下のデータは全体の４％を占め、同じリーフの男性に比して有価証券保有総額が３カテゴリー低い。P値は0.001で、統計的には１％水準で有意である。

(ii)［最下段の左から２つめ］世帯収入が1,000万円以上で個人収入が300万円未満、70歳以上のデータは全体の１％を占め、同じリーフの男性に比して有価証券保有総額が0.36カテゴリー低い。P値は0.672で、統計的には10％水準でも非有意である。

(iii)［最下段の左から３つめ］世帯収入が1,000万円以上で個人収入が300万円以上のデータは全体の14％を占め、同じリーフの男性より有価証券保有総額が0.98カテゴリー低い。P値は0.000で、統計的には１％水準で有意である。

(iv)［最下段の左から４つめ］世帯収入が1,000万円未満で500万円以上（revenue_hが３または４）、個人収入が300万円未満で、年齢が30歳以上のデータは全体の24％を占め、同じリーフの男性に比して、有価証券保有総額が0.79カテゴリー低い。P値は0.000で、統計的には１％水準で有意である。

(v)［最下段の左から５つめ］世帯収入が500万円未満で（revenue_hが１または２）、個人収入が300万円未満で、年齢が45歳以上のデータは全体の37％を占め、同じリーフの男性に比して、有価証券保有総額が0.29カテゴリー低い。P値は0.032で、統計的には５％水準で有意である。

(vi)［最下段の左から６つめ］世帯収入が500万円未満で（revenue_hが１または２）、個人収入が300万円未満で、年齢が45歳未満のデータは全体の３％を占め、同じリーフの男性に比して、有価証券保有総額が0.55カテゴリー高い。P値は0.222で、統計的には10％水準でも非有意であ

---

14) この手法は、GitHubで展開されているSusan Ashleyの提案による。例えば、ここの有価証券保有額の例では、最終段に７リーフできているため、目的変数を14（＝７×２）個の説明変数で回帰することになる。最終段にnリーフできれば、２n個の説明変数で回帰することになる。他にも有意性に関しては、http://pypi.orgというサイトに上げているTran and Zheleva のソフトウェア（causal-tree-lean）などで、各リーフ内で平均値の差の検定を行うという手法も提案されている。いずれにしても、こうしたリーフの作り方は、Causal Treeの方法によらねば、極めて多数にのぼり、その結果、回帰分析を実効あらしめなくする。Causal Treeは、こうした極めて多数の候補から、分析が実行可能な数にまで絞り込むことを可能にしているという面でも、メリットがあると考えることができる。

15) 以下では最終段のリーフの中にも統計的に非有意なものも含まれている。これは、第２節で展開したように、枝分かれの作り方が統計的な有意性を基準とせず、それとは異なる基準によっていることから、起こりうることである。Causal Treeによる枝分かれの作成と、伝統的な計量経済学の有意度は根本的に別の基準である。それにも拘わらず、有意度を示したのは、伝統的な計量経済学の検定に慣れ親しんだ読者の関心に応えんがためである。

る。

㈦ ［最下段の右から２つめ］世帯収入が1,000万円未満で、個人収入が500万円未満、年齢が30歳未満のデータは全体の１％を占め、同じリーフの男性に比して有価証券保有総額が0.18カテゴリー低い。P値は0.872で、統計的には10％水準でも非有意である。

㈧ ［最下段の最も右］世帯収入が1,000万円未満でかつ個人収入が500万円以上のデータは全体の16％を占め、同じ分類の男性に比して、有価証券保有総額が１カテゴリー高い。P値は0.026で、統計的には５％水準で有意である。

ここで、上記の結果を概観してみたい。(ⅰ)～㈧の８つのリーフのうち、－の符号がついているものは６つ、割合にして81％にあたる。このリーフでは女性は男性に比して、有価証券保有総額は小さい。有価証券投資は、全体として、男性の方が女性よりも盛んに行っていることが確認できる。

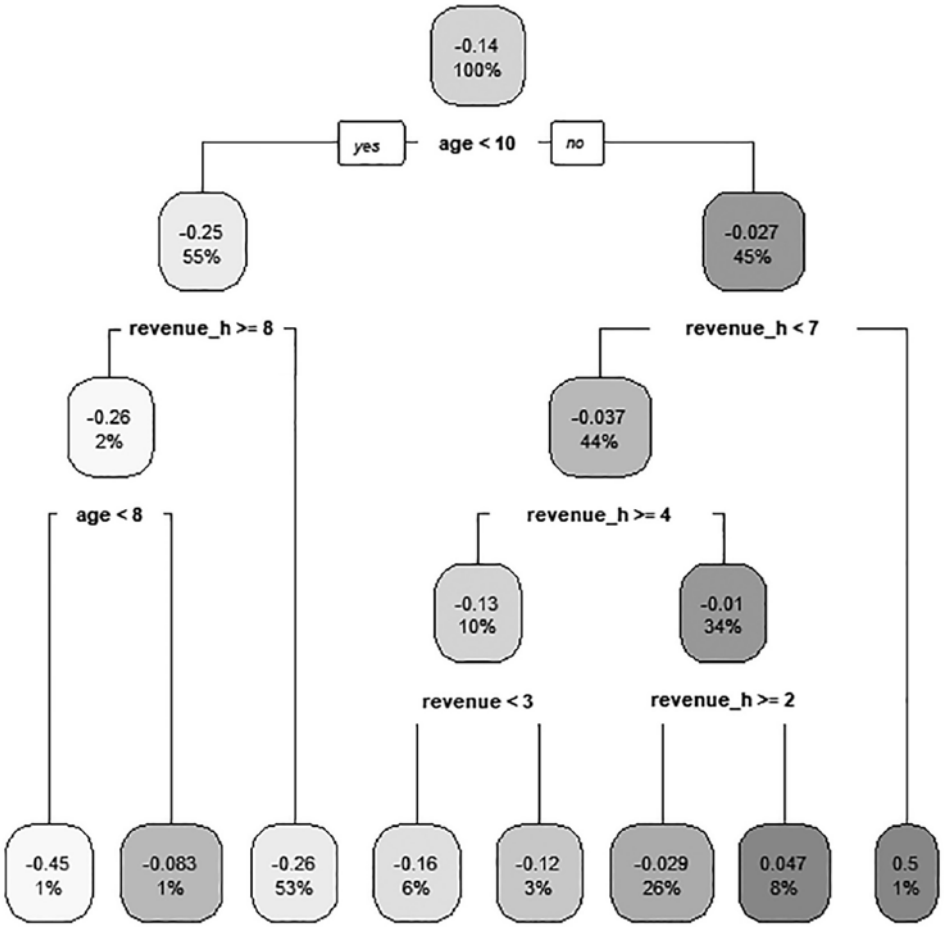
残りの２つのリーフ、㈥と㈧では逆に女性の方が男性よりも、有価証券保有総額が大きい。㈥の女性は、世帯収入が500万円未満の、比較的所得の低い世帯で、個人収入が300万円未満、年齢が45歳未満と比較的若い女性である。低所得世帯で有業の女性像が浮かび上がる。このリーフにおいては、男性の資力が乏しいため、有価証券資産を多く持てない結果、結果的に女性の方が多くなっているということかもしれない。一方、㈧の女性は、世帯収入が1,000万円未満でかつ個人収入が500万円以上の女性である。世帯年収は中庸でありながら、個人年収は高い部類の女性である。これは、家庭内で経済的に自立している度合いが高い女性のようである。高所得世帯とは言えない世帯のなかで比較的稼いでいる女性は、男性よりも証券投資に励んでいると解釈できる。

ところで、㈤と㈥のリーフを比較すると、分岐「age>= 6 」でyesかnoかで符号が逆転している。世帯収入が500万円未満で（revenue_hが１または２）、個人収入が300万円未満という条件は、両者ともに共通である。違いは45歳以上か45歳未満かのみである。45歳未満の場合、男性よりも証券保有額が多く、45歳以上では少なくなっていることは注目すべき、興味深い結果と言えよう。45歳未満であれば有業である確率が高く、かつ様々な職種を選ぶことが可能であろうから、同じ「revenue< 3 」のカテゴリーであっても、45歳未満の女性の方が稼いでおり、資産の蓄積により励んでいると解釈できるかもしれない。

②　確定拠出型年金への加入・非加入を目的変数にとった結果

図７が、確定拠出型年金への加入・非加入に関する分析結果である。まずルート直下の枝分かれは、ageが10未満かいなか、つまり64歳以下（左へ向かう）か65歳以上（右へ向かう）の高齢者かとなっている。それ以降については、先と同様にひとつひとつ解釈を施していくと、Causal Treeの結果は以下にまとめることができる。尚、確定拠出型年金への加入・非加入は０または１の値をとる二値変数であるから、Tree内の数値は加入率の男女差と解釈することが可能である。

図７：確定拠出型年金への加入・非加入を目的変数にとった結果（Causal Tree）



⑴［最下段の最も左］年齢が54歳以下で、世帯収入が2,000万円以上のデータは全体の１％を占め、同じ部類の男性に比して加入率が45％ポイント低い。P値は0.037で、統計的には５％水準で有意である。

⑵［最下段の左から２つめ］年齢が55歳以上64歳以下で、世帯収入が2,000万円以上の女性は、同じ部類の男性に比して加入率が8.3％ポイント低い。P値は0.674で、統計的には10％水準でも非有意である。

⑶［最下段の左から３つめ］年齢が64歳以下で、世帯収入が2,000万円未満のデータは全体の53％を占め、同じ部類の男性に比して加入率が26％ポイント低い。P値は0.000で、統計的には１％水準で有意である。

⑷［最下段の左から４つめ］年齢が65歳以上の高齢者で、世帯収入が700万円以上1,500万円未満（選択肢４～６）で、個人収入が500万円未満のデータは全体の６％を占め、同じ部類の男

性に比して加入率が16％ポイント低い。P値は0.032で、統計的には５％水準で有意である。

(v) ［最下段の左から５つめ］年齢が65歳以上の高齢者で、世帯収入が700万円以上1,500万円未満（選択肢４～６）で、個人収入が500万円以上のデータは全体の３％を占め、同じ部類の男性に比して加入率が12％ポイント低い。P値は0.464で、統計的には10％水準でも非有意である。

(vi) ［最下段の左から６つめ］年齢が65歳以上の高齢者で、世帯収入が300万円以上700万円未満（選択肢１～２）のデータは全体の26％を占め、同じ部類の男性に比して加入率が2.9％ポイント低い。P値は0.463で、統計的には10％水準でも非有意である。

(vii) ［最下段の左から７つめ］年齢が65歳以上の高齢者で、世帯収入が300万円未満（選択肢１）のデータは全体の８％を占め、同じ部類の男性に比して加入率が4.7％ポイント高い。P値は0.495で、統計的には10％水準でも非有意である。

(viii) ［最下段の最も右］年齢が65歳以上の高齢者で、世帯収入が1,500万円以上の高所得世帯のデータは全体の１％を占め、同じ部類の男性に比して加入率が50％ポイント高い。P値は0.014で、統計的に５％水準で有意である。

上記の結果を俯瞰しよう。符号が正な、つまり女性の加入率が男性より高いのは、(vii)と(viii)のみで、併せて９％に過ぎない。それ以外の、91％を占めるリーフでは符号が負であるから、男性の加入率の方が高いことになる。そこで、まず、確定拠出型年金は大多数の場合で、男性の方が女性より加入率が高いことがわかる。確定拠出型年金は、税制上有利であるとはいえ、制度が複雑であるため、それに関する知識が必要である。また、確定拠出型年金のひとつである企業型は、企業に勤務していなければ加入ができないが、確定拠出型年金制度を持っている企業での就業者は男性の方が女性より多いであろう。こうしたことが、この男女差の背景かも知れない。

(vii)と(viii)はともに、年齢が65歳以上である。違いは、世帯年収が300万円未満と低いか、逆に1,500万円以上と高額収入世帯かである。65歳以上の高齢女性では、年収の低い層と高い層で同じ傾向が出ていることは興味深い。その理由について若干の推測をしてみると、収入の低い層でこうした傾向が出たのは、夫の遺族年金も低いため、老後の生活保障としては不十分であり、女性は積極的に自分の老後を考えた結果なのではないかと考えられる。一方、高い層では、男性は公的年金受給額も多く、また貯蓄も相当額あるため、男性が確定拠出型年金への加入の誘因を持っていなかったため、つまり女性が高いというよりも、男性が低いということではないかと考えることは可能である。いずれにしても、この理由については、より精緻な調査とそれに基づく検証が必要である。

５． おわりに

1996年の金融ビッグバン以来、家計の証券投資を政策面で後押しするために制度改革が進められてきているが、我が国家計の金融資産の構成は、諸外国と比べて依然として預貯金偏重である。こ

の理由を探るべく、これまで様々な政策・制度変更に焦点を当てて実証分析が行われてきたが、その中で、アンケート調査の個票データを用いた研究では、個人属性などを説明変数にとり、政策効果を表す目的変数への影響を、質的変数に関する推定法で分析することが多い。しかし、この手法はダミー変数として取り入れる説明変数の選択に恣意性が介在するという技術的な限界がある。そこで、本論文はそうした難点を克服した、機械学習の手法、その中でも昨今、急速に利用され始めている、Athey and Imbens［2016］らによるCausal Treeという分析手法を用いて、『個人投資家の証券投資に関する意識調査』の個票データで解析を試みた。目的変数には有価証券保有額に加え、確定拠出年金への加入・未加入を採用して、性別を処置変数、世帯収入、個人収入、年齢を説明変数に用いた。

有価証券保有額の分析では、全８つのリーフ中、全体の81％を占める６つのリーフで、女性の有価証券保有額は男性より少ないことが確認された。一方、世帯収入が500万円未満の、比較的所得の低い世帯で、個人収入が300万円未満、年齢が45歳未満の比較的若い女性と、世帯収入が1,000万円未満でかつ個人収入が500万円以上の女性では、女性の有価証券保有額が男性よりも多くなった。低所得世帯で有業の女性、世帯年収は中庸でありながら、個人年収は高い部類の女性は、男性よりも証券投資に励んでいることが示された。

確定拠出年金の分析でも、全７つのリーフ中、全体の91％を占める６つのリーフで、男性の方が女性より加入率が高いことが確認された。制度の複雑さや、確定拠出年金のひとつである企業型は、企業に勤務していなければ加入ができないことが、この男女差の背景と考えられる。逆に女性の方が、加入率が高いリーフは、どちらも年齢が65歳以上で、世帯年収が300万円未満と低いか、逆に1,500万円以上と高額収入世帯かの両極端である。収入の低い層では、夫の遺族年金も低いため、老後の生活保障が不十分であり、女性は積極的に自分の老後を考えたためだと考えられる。一方、所得の高い層では、男性は公的年金受給額も多く、また貯蓄も相当額あるため、男性が確定拠出年金への加入の誘因を持っていなかったため、男性が低いということではないかと考えることが可能である。但し、より確かなことを言うためには、さらに精緻な調査とそれに基づく検証が必要である。

本稿を終えるにあたり、その限界に言及し、後続の研究を展望したい。いうまでもなく、処置変数は性別以外にも様々なものが考えられる。目的変数についても、『調査』には株式、公社債、投資信託それぞれの保有額を問う問いもあるが、今回はその合計額を用いた。説明変数についても、今回用いたもの以外に多くの候補が考えられる。そうした中、今回の研究は処置変数、目的変数、説明変数のいずれにおいても限定的に取り上げたのは、紙幅の都合もさることながら、Causal Treeという、未だ多くの人になじみのない手法を紹介することを貢献の一つに据えたためである。故に、今後の研究においては、まず、さまざまな目的変数、処置変数、説明変数の組み合わせで分析を拡

張して行くことが考えられる。併せて、Causal Treeを発展させたCausal Forestも昨今、盛んに研究され、また諸問題に活用されつつある。今後は、本論文が追及するテーマに関しても、このCausal Forestの活用が期待されるところである。

参考文献

Ashley, S and G. Imbens, "Recursive Partitioning for Heterogeneous Causal Effects," *Proceedings of the National Academy of Science of the United States*（*PNAS*）, Vol.113（27）, pp.7353-7360, 2016.

荒木雅弘『フリーソフトによる機械学習入門』、森北出版、2014.

Muller, A. C. and S. Guido著、中田秀基訳『Pythonで始める機械学習』、オライリー・ジャパン、2017.

補論：回帰木とは何か[16)]

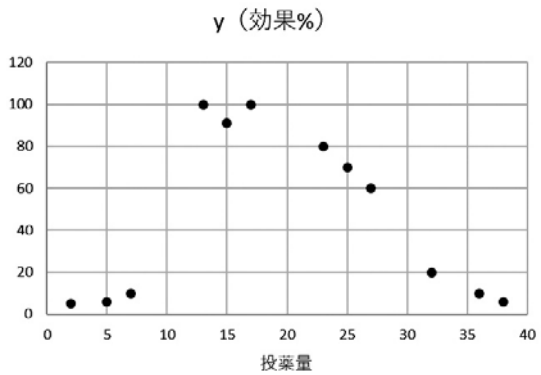
回帰木とは機械学習の一手法であって、分析の対象となる連続型目的変数yと説明変数xとの間の関係を、伝統的な計量経済学の手法によって分析するのではなく、以下のような考え方を用いて分析する手法のことである。以下に、具体例で解説しておこう。

ある薬の投薬量xとその効果yとの関係が表Ａ１と図Ａ１に示されている。

表A１　薬の摂取量とその効果

x（摂取量）	2	5	7	13	15	17	23	25	27	32	36	38
y（効果％）	5	6	10	100	91	100	80	70	60	20	10	6

図A１　薬の摂取量とその効果（１）



図A１によれば、投薬量が少ない場合と多い場合に効果がほとんど見られず、中間的な投薬量の場合にかなりの効果があることが観察される。したがって、図A１のデータに対して線形回帰モデルを適用しても不適切であることは明白である。このような例に対して回帰木は適切な解を与えてくれる。

１．予測値とノード・リーフ

まず、投薬の効果がどれほどあるかの推定値としては、最もシンプルには、xを分割せず、全てのyの平均値を用いることが妥当であろう。このデータでは、その平均値は46.5となる。これを図

16) 本節はhttps://statquest.org/に依拠している。

示すると以下のようになる。

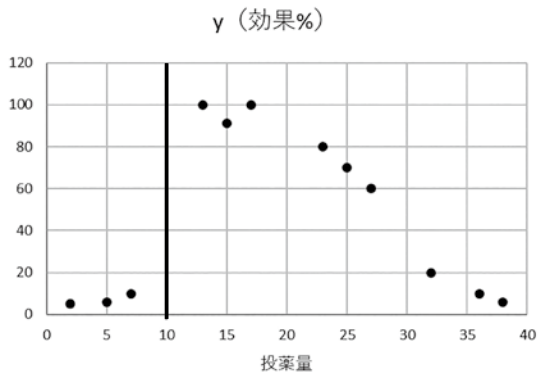
図A２　回帰木（１）（ルート）

予測値＝46.5

図A２はxを全く分割しない状態でyを予測しており、これを「ルート」と呼ぶ。

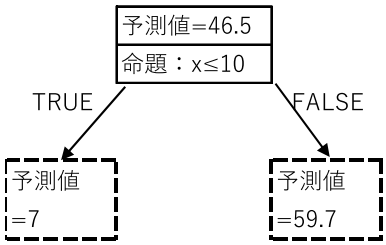
次に、ｘを図Ａ３のように分割して（閾値を設けて）考えてみよう。

図A３　薬の摂取量とその効果（２）



図A３によれば、xが10以下の場合のyの予測値は、（５＋６+10）/３＝７、同様にして、xが10より大きい場合には、残りのデータの平均59.7で予測値とすることが自然であろう。ここで、図A４を得る。二つの矢印は、「命題：x≤10」が真のときは左下へ、偽のときには右下へ進むことを示す。

図A４　回帰木（２）

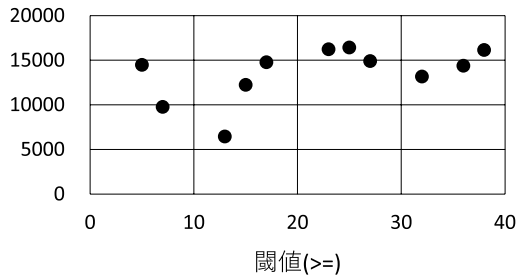


ここで、┈┈┈┈┈はリーフと呼ばれるが、下にこれ以上進まない末端でなければノードという。回帰木⑵は回帰木⑴よりも、詳細なｙの予測値を与えていることが自明であろう。

ここで、回帰木⑴から回帰木⑵に移行する際に、閾値をx=10としたのは、まったくアドホックに

決められている。そこで、この分割点を合理的に判断する基準として、予測値の残差平方和を考えることにする。回帰木⑴に戻って、xの閾値を考えると、データ数が12なので、11の分割点を考えることができる。この11の分割点のそれぞれに対して残差平方和を計算し、図示したものが、図A５になる。

図A５　全ての閾値に対する残差平方和



図A５によれば、３番目の閾値の時に残差平方和が最小になっていることが分かる。これは、小さい値からxを並べたときに、３番目と４番目のデータの間の値を閾値にしていることになる。すなわち、回帰木⑵が、実は、ルートから第１層のノードを作る際の最適な閾値を採用した結果となっていることが分かる。

この様にして、回帰木⑵のリーフの配下に、次の閾値をそれぞれ考えて、次々とリーフを下方向に展開していくことは容易に想像できよう。そして、そのようなアルゴリズムによって最適化された回帰木を最終的な予測のための回帰木として採用することになる。

２．過学習の回避法

しかしながら、ここで、次のような疑問が出てくるであろう。すなわち、回帰木を次々と展開していけば、一つのリーフに一つのデータがある回帰木が、残差平方和が０になることによって、常に採用されることになるのではないかと。しかしながら、そのような回帰木は、予測モデルとしては妥当ではないことは直感的に理解できる。これを「過学習」（over fitting）と呼ぶ。というのは、この回帰木の性能を確かめるために、全く別のデータ（テストデータ）をもってきて、すでに作成された回帰木で予測してみると、大きな残差平方和になる可能性があるからである（たとえば、新しいデータでは、x= 2 に対して、yが５ではなく７が観測されるようなケースを想起せよ）。したがって、次の問題はこの過学習をどのように防いで最適な回帰木を構築するかということになる。

この問題は、どの回帰木が最適かを判断する基準として、残差平方和だけを採用していることに

起因している。これを解決する方法は複数存在するが、最もポピュラーなツリースコアを紹介しよう。ツリースコアとは、以下のように定義される。

$$\text{ツリースコア} = \text{残差平方和} + \alpha \times \text{リーフ数}$$

αは正のチューニングパラメータであり、リーフの数が不必要に多い回帰木にペナルティを課す目的で導入される。たとえば、αが十分に大きいと、ルートだけからなる回帰木が採用されることになる。また、$\alpha = 0$であれば、ツリースコア＝残差平方和となるので、一つのリーフに一つのデータが対応する回帰木が最良とみなされることになる。

αの決定に際しては、クロスバリデイション（cross validation）という手法がとられる。まず、全てのデータを用いて、フルサイズの回帰木[17]とルートだけからなる回帰木、およびその中間の回帰木を用意する。これを、$\Pi_1, \cdots, \Pi_m$としよう。そして、$\Pi_1, \cdots, \Pi_m$のそれぞれについて最小のツリースコアを与えるαをあらかじめ用意しておく。以下、それらを$\alpha_1, \cdots, \alpha_m$とよぶ。そのうえで、全てのデータを、「訓練データ」（回帰木を構築するためのデータ）9割と「テストデータ」1割に分割し、訓練データでさらに別の回帰木n個を構築し（$\pi_1, \cdots, \pi_n$）、それらn個の回帰木をテストデータに適用して、残差平方和を計算する。こうして、最小の残差平方和を与える回帰木を特定する（π^*）。ついで、こうして特定された回帰木のツリースコアを最小とするαを$\alpha_1, \cdots, \alpha_m$なかから選び（$\alpha^*$）、記録しておく。

次に、先ほどとは異なる1割のデータをテストデータ、残りを訓練データとし、先ほどと同じように、最小の残差平方和を与えるαを記録する（α^{**}）。これを10回繰り返して得る、そうしたα（α^*, α^{**}など）の中から最も平均的に良いものを最終的なαとして採用し、このαを用いたツリースコアが最小になる回帰木Πを最終的な回帰木として用いることにする。このような手法を10-fold cross validationと呼ぶ。

さらに詳しい説明は、荒木（2014）、Muller and Guido（2017）などを参照されたい。

17) 例えば、リーフには最低ℓ個のデータが存在しなければならないという条件のもとで、残差平方和が最小になるようなxの分割を進めた結果、得られる回帰木のこと。