応用統計学　Vol. 33, No. 2 (2004), 111–130

交互作用基準による再帰分割線形モデル

群馬大学工学部情報工学科　関　　庸一

(株) SRA　野島　　勇

要　旨　樹形回帰分析は古典的重回帰分析よりも交互作用効果を容易にモデル化できる．しかし，全ての共変量の効果を層別のみで説明するため，深く大きな回帰木を形成してしまう傾向がある．特に，全サンプルあるいは，あるサンプル群に共通の複数の効果が存在する場合，多くの類似した部分木を含む木を推定することとなる．そこで，本研究では，この冗長性を避けるため，決定節に線形回帰項を含みうるよう拡張したモデルを提案する．これは，線形回帰項により主効果を説明し，層別により効果の異質性を説明する樹形回帰モデルとなる．この場合，多くの候補モデルが存在し，モデル選択が大きな問題となる．そこで，MDL 基準を用いることにより，モデルの推定アルゴリズムを提案する．この基準は最大の交互作用効果をもつ分岐変量条件の選択とも解釈できる．最後に，数値例と高齢者介護時間データにより提案手法の有効性を示す．

1. はじめに

近年，デジタル化されたデータが低コストで大量に収集・蓄積される場面が増えてきている．そこで，このような多量データを分析する方法の確立が求められているが，データが広い範囲から収集されるため，従来の単一の統計モデルでは，その多様性に対応できない場合があるという問題点が存在する．本論文では，このようなデータのうち，多くの因子変量を含む共変量群を用いて連続応答変量を回帰するという問題を対象とし，樹形回帰モデルを拡張し，従来より簡潔な木の構造を与える再帰分割線形モデル(recursive partitioning linear model, RLIM)を提案する．

データに多様な構造がある場合，サンプルの適切な層別を発見し，層ごとにモデルを探索することが一つのアプローチとなる．このようなアプローチをとる分析方法として，樹形モデル(Breiman et al. 1984; Chambers & Hastie 1992; Murthy 1998 など参照)がある．この方法では，サンプルセットを再帰的に分割し，作成されたサンプルグループの応答変量の代表値を，そのグループの推定値として採用する．各分割では，共変量の値に基づく分割法の候補から，分割後のそれぞれのグループで応答変量の値が最も均質となる分割法が採用される．サンプルセットを分割しながら共変量の効果を探索するので，多様な構造を含むデータの一部にのみ存在する効果(高次の交互作用効果)を発見できる可能性が高く，データセット全体に一貫した線形モデルを推定しようとする重回帰分析などに比べ，柔軟な方法となっている．

しかし，層別のみで応答変量への共変量の効果を説明しようとするため，

1. 多くのサンプルに共通する主効果が複数ある場合，それらが部分木ごとに分割されてそれぞれで表現されるため，大きな回帰木が形成される．
2. サンプルセットが枝葉で細分化されることで，小さな効果が誤差に埋もれて発見されにくくなる．

という問題点が残されていると考える．データが，その多様性に比較して多量に存在すれば，後者の予測力の犠牲は少なくなるが，現象に簡潔な説明を与えるモデルとならないという問題点は残る．

この原因が，各決定節におけるサンプルの層別と，形成されたノードでの応答変量の説明とに，共通の共変量を用いることにあると考え，その両者を分離することを本研究で提案する．具体的には，各部分木に共通して存在する主効果については，その部分木の root ノードにおいて線形回帰項を用意することで説明し，分岐は主効果の大きさではなく，分岐条件と共変量の交互作用効果が大きいものを選ぼうとするものである．これにより，できるだけ簡潔な木構造のモデル推定をめざす．

なお，この木の分岐条件の探索においては，用意した線形回帰項が共変量の主効果を説明してしまうので，関・筒井・宮野 (1999) による分岐で説明される主効果の大きさを基準とした分岐条件の選択で，層別に効果的な分岐基準を選択できない．そこで，層別で説明される交互作用効果の大きさを基準とした分岐法を提案する．

以上の様な拡張を行うと，従来の解析におけるモデルクラスに比べ，選択対象のモデルクラスが非常に大きくなる．そこで，本研究では，簡潔なデータの記述を探索するためのモデル選択基準として Rissanen (1978, 1983, 1984) による MDL 基準を採用し，Quinlan & Rivest (1989) で行われたようなモデル選択を行いながらモデルを生成する Greedy な算法を実現する．

上述の方針のように，樹形回帰モデルに線形回帰を追加しようとするモデルの拡張としては，Quinlan (1992), Karalic (1992), Dobra & Gehrke (2002) などが既にある．これらの研究では，葉に線形回帰を追加する方法が提案されているが，本研究では，中間の決定節にも線形回帰項の採用を許すところが異なる．これにより，本提案法では，サンプルセットの広い範囲で共通して存在する効果は，できるだけ上位の決定節で説明するモデルを生成しようとする．また，交互作用に着目するものとしては，Loh (2002) がある．ここでは，分岐ごとに χ^2 検定統計量に基づく p 値を用い交互作用をもつ変数を見つけ出そうとしているが，本研究では，平行して行われる変数選択に用いる基準と共通の基準に基づき，線形モデルとして一貫した評価に基づくモデルを作成する方法を与える．

分岐変数としては，たとえば，Dobra & Gehrke (2002) のように，複数の変数を同時に考慮した複雑な分岐条件を考える場合もあるが，本研究では，意味のある変量組み合わせについては，事前に変量化して元データに追加すればよいと考え，複数変量を同時に用いた分岐条件については考慮しない．また，連続変量の取扱いについては，Friedman (1991) による多変量のスプラインを行う MARS などのように，閾値を探索して分岐に採用することで，区分的線形な効果を検討する方針も考えられるが，本研究では考慮しない．分岐変量としては，因子変量として与えられたもののみを候補とする．

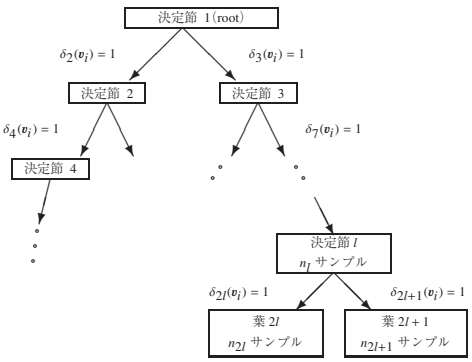
以下，第 2 節で従来の樹形回帰モデルの線形モデルとしての表現を与え，これを拡張し，第 3 節で再帰分割線形モデルを提案する．第 4 節で提案モデルの推定法を示し，第 5 節で数値実験，

第 6 節で高齢者介護時間データでの検証例を示す．そのうえで，第 7 節で提案法の残された課題について検討する．

2. 従来の樹形回帰モデル

以下では，n サンプルの応答変量 $\boldsymbol{y} = (y_1, \ldots, y_n)^t$ を共変量 $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)^t$ によって回帰する問題を考える．ここで，共変量は q 変量あり，$\boldsymbol{v}_i = (v_{i1}, \ldots, v_{iq})^t$ とする．また，$\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^t$ を $\boldsymbol{v}_i$ のうち連続変量はそのまま，因子変量はダミー変数化して並べた変数とし，$\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^t$ と表すものとする．たとえば，ダミー変数の作成に treatment 対比を用いる場合には，第 j 因子変量 v_{ij} が c 水準の因子変量であり，対応するダミー変数を $\{k_h, h = 1, 2, \ldots, c-1\}$ 番目の変数とすると，$v_{ij} = h$ なら $x_{ik_h} = 1$，それ以外のとき $x_{ik_h} = 0$ とすることになる．

図 1 に，樹形回帰モデルの概念図を示す．以下では，決定節と葉を合わせてノードと呼び，各ノードには幅優先で番号を付けてノード l と呼ぶことにする．l は root での 1 に始まり，ノード l における左の子ノードを $2l$，右の子ノードを $2l+1$ とする．ノード所属条件を，ダミー変数 $\delta_l(\boldsymbol{v}_i)$ を用いてあらわす．これは，ノード l に含まれるサンプル i に対しては $\delta_l(\boldsymbol{v}_i) = 1$，それ以外のサンプル i について $\delta_l(\boldsymbol{v}_i) = 0$ とするものである．つまり，ノード l が展開されていれば，ノード l に含まれるサンプル i に対しては，$\delta_{2l+1}(\boldsymbol{v}_i) = 1 - \delta_{2l}(\boldsymbol{v}_i)$ となる．また，サンプル i がノード l に含まれない，つまり，$\delta_l(\boldsymbol{v}_i) = 0$ なら，$\delta_{2l}(\boldsymbol{v}_i) = \delta_{2l+1}(\boldsymbol{v}_i) = 0$ となる．ノード l のサンプル数を $n_l = \sum_i \delta_l(\boldsymbol{v}_i)$ と表す．



$\delta_l(\boldsymbol{v}_i)$：第 l ノードでの所属条件を表わすダミー変数
共変量 $\boldsymbol{v}_i$ の値をもつサンプルがノード l に含まれるとき
1，その他のとき 0 なる変数とする．

図 1. 樹形回帰モデル概念図

実際の算法でノード l への所属条件を探索する際には，ノード $\lfloor l/2 \rfloor$ への所属条件に追加する条件として何が適当かという形での探索を行なう．この追加論理式を $P_l(\boldsymbol{v}_i)$ とする．最も基本的な $P_l(\boldsymbol{v}_i)$ としては，$\boldsymbol{v}_i$ のうち一変量のみに依存した $x_{ik} = 1$ などがある．探索範囲が広がって計算量が増えることを厭わなければ，もっと複雑な複数変量を同時に考慮した分岐条件を含めて候補とすることもできる．

論理式 P について $I(P)$ をその特性関数とする．つまり，P が真のとき $I(P)=1$，P が偽のとき $I(P)=0$ とすると，δ_l は

$$\delta_l(\boldsymbol{v}_i) = \prod_{h \in Ancestor(l)} I(P_h(\boldsymbol{v}_i))$$

と表せることとなる．ただし，$Ancestor(l)$ を，ノード l の祖先ノード $Ancestor(l) = \{\lfloor l/2 \rfloor, \lfloor l/2^2 \rfloor, \lfloor l/2^3 \rfloor, \ldots, 1\}$，$\lfloor x \rfloor$ を x を越えない最大整数の記号とする．

以上の記法を用いると樹形回帰モデルは，root で分岐しない場合も一般に含め，

$$Y_i = g_1^*(\boldsymbol{v}_i) + \epsilon_i$$

と表現できる．ただし，誤差 $\epsilon_i \sim N(0, \sigma^2)$ であり，$g_l^*(\boldsymbol{v}_i)$ をノード l を root とする部分木における効果の再帰的表現

$$g_l^*(\boldsymbol{v}_i) = \begin{cases} \sum_{l'=2l,2l+1} \delta_{l'}^*(\boldsymbol{v}_i) g_{l'}^*(\boldsymbol{v}_i) & \text{分岐する場合} \\ \alpha_l^* & \text{分岐しない場合} \end{cases} \tag{1}$$

と定義する．ここで，α_l^* はノード（葉）l に所属するサンプルの平均値パラメータを表わすものとする．また，記号に * を付与した場合は母数を意味し，対応する推定量は * なしの記号とする．

2.1. 主効果による分岐基準

樹形回帰モデル g_1^* を推定するには，(1) 式の $\{g_l^*\}$ の推定結果 $\{g_l\}$ を求めることになるが，計算量の問題からモデル全体の最適性を保証することは難しく，各決定節ごとに最適化を行う Greedy な木の成長算法しか実用的でない．

具体的算法としては，個々のノードで分岐を行なうかや，分岐条件 $\delta_l(\boldsymbol{v}_i)$（または，$P_l(\boldsymbol{v}_i)$）として何を採用するかを決定する必要がある．これらを判定する基準を分岐基準と呼ぶことにする．分岐基準が，各々の決定節ごとに Greedy に評価できるためには，root からその決定節までの部分モデルのみを用いて算出できるものであることが必要である．つまり，ノード l での分岐を検討する際には，l の祖先ノード $Ancestor(l)$ までの分岐モデルとそのノードに含まれるデータ $\{(y_i, \boldsymbol{v}_i) \mid i : \delta_l(\boldsymbol{v}_i) = 1\}$ を利用して，評価できる基準であることが要求される．

このような分岐基準としては，Quinlan (1993) によるエントロピーを基準とする方法など各種の方法が提案されているが，ここでは分岐により説明できる主効果に関する最小二乗基準を従来法として比較対象とする．これは正規誤差の仮定の下で，最尤法を考えるのと同等となる．この場合，α_l^* の推定値は，ノード l が葉となる場合の残差平方和 $RSS_L(l) = \sum \delta_l(\boldsymbol{v}_i)(y_i - \alpha_l)^2$ を最小化することで，以下となる．

$$\hat{y}_{il} = \alpha_l = \sum_{i:\delta_l(\boldsymbol{v}_i)=1} y_i / n_l \tag{2}$$

また，分岐条件としては，子ノード（$l' = 2l, 2l+1$）への分岐 $\delta_{l'}(\boldsymbol{v}_i)$ を考えた残差

$$RSS_N(l, \{\delta_{l'}\}) = \sum_{i:\delta_l(\boldsymbol{v}_i)=1} \{y_i - (\delta_{2l}(\boldsymbol{v}_i)\alpha_{2l} + \delta_{2l+1}(\boldsymbol{v}_i)\alpha_{2l+1})\}^2 \tag{3}$$

が最も小さくなる $\delta_{l'}$ を選択することになる．

なお，$RSS_N(l, \{\delta_{l'}\})$ を分岐しない場合の $RSS_L(l)$ と比較すると，

$$RSS_L(l) - RSS_N(l, \{\delta_{l'}\}) = \sum_i \sum_{l'=2l,2l+1} \delta_{l'}(\boldsymbol{v}_i)(\alpha_l - \alpha_{l'})^2 \tag{4}$$

が非負となり，必ず適合度が改善するので，ノード l において分岐するかどうかを決定するには，クロスバリデーションなどのモデル評価方法を用いることになる．

3. 再帰分割線形回帰モデル

3.1. 線形回帰項の追加

提案する再帰分割線形回帰モデルでは，各ノードを root とする部分木に共通する主効果を説明するため，樹形モデル (1) 式に線形回帰項 $f_l^*(\boldsymbol{v}_i)$ を以下のように追加する．

$$g_l^*(\boldsymbol{v}_i) = f_l^*(\boldsymbol{v}_i) + \begin{cases} \sum_{l'=2l,2l+1} \delta_{l'}^*(\boldsymbol{v}_i) g_{l'}^*(\boldsymbol{v}_i) & \text{分岐する場合} \\ \alpha_l^* & \text{分岐しない場合} \end{cases} \tag{5}$$

ただし，線形回帰項 $f_l^*(\boldsymbol{v}_i)$ は，ノード l で採用する共変量の集合を $V_l \subset V$ として，以下のように表せるものとする．

$$f_l^*(\boldsymbol{v}_i) = \sum_{j \in J(V_l^*)} \beta_{lj}^* x_{ij} \tag{6}$$

ここで，ダミー変数化行列 $\boldsymbol{X}$ の列番のうち，変量群 V_l^* に対応するものを $J(V_l^*)$ で表すものとする．

以上より，ノード l における応答変量の推定値はそのノードまでの線形回帰項 $f_l(\boldsymbol{v}_i)$ の総和となる．

$$\hat{y}_{il} = \alpha_l + f_l(\boldsymbol{v}_i) + \sum_{l' \in Ancestor(l)} f_{l'}(\boldsymbol{v}_i) \tag{7}$$

線形回帰項 f_l に採用する変量については，以下の制約を置く．

1. モデルが不必要に複雑になることを避けるため，親ノード l と子ノード $(l' = 2l, 2l+2)$ で同時に同じ共変量を線形回帰項として採用することは避け，子ノードに採用される変量は，親ノードでは採用しないものとする．つまり，$V_l \cap \cup_{l'} V_{l'} \neq \phi$ ならば V_l を $V_l - \cup_{l'} V_{l'}$ とすることで，$V_l \cap \cup_{l'} V_{l'} = \phi$ となるようにする．これにより，ある共変量に主効果があった場合でも，その残差を説明する子ノードにおいて効果があった場合，つまり，分岐変数との交互作用効果もある場合には，交互作用効果のみを残すこととなる．
2. 一つの因子変量が複数のダミー変数で表現される場合，それらは一括して採用するか，採用しないかのいずれかとする．ただし，なんらかの事前の判断に基き，一つの因子変量を意味のある対比に分解することができるならば，それぞれを新しい因子変量として共変量群に追加し，効果の大きい対比のみを残せるようにすることも考えられる．

以上のモデルの下，$\{f_l(\boldsymbol{v})\}$ と $\{\delta_l(\boldsymbol{v})\}$ を選択する基準と推定算法について以降で議論する．

3.2. 提案モデルでの適合度評価

提案モデル (5) 式の回帰木の成長過程で，ノード l の先祖ノードのモデルを前提として，ノード l を構成するには，ノード l を葉と考える場合と，ノード l を決定節と考える場合の二通りの適合度を与える必要がある．

ノード l を葉と考える場合の線形回帰項選択のための適合度としては，従来法と同じく，$\hat{y}_{il}$ の残差平方和を (7) 式を用いて考えれば良く，以下となる．

$$RSS_L(l, V_l) = \sum_{i:\delta_l(\boldsymbol{v}_i)=1} (y_i - \hat{y}_{il})^2 \tag{8}$$

一方，ノード l を決定節と考える場合には，$\delta_l(\boldsymbol{v})$ の選択が必要となる．線形回帰項が決定され $\hat{y}_{il}$ が求められたとした際，(3) 式の RSS_N を単純に拡張しようとすると，

$$RSS'_N(l, \delta_{l'}) = \sum_{i:\delta_l(\boldsymbol{v}_i)=1} \{y_i - (\hat{y}_{il} - \alpha_l + \delta_{2l}(\boldsymbol{v}_i)\alpha_{2l} + \delta_{2l+1}(\boldsymbol{v}_i)\alpha_{2l+1})\}^2 \tag{9}$$

となる．しかし，分岐による効果 $\delta_{l'}(\boldsymbol{v}_i)\alpha_{l'}$ は，線形回帰項 $f_l(\boldsymbol{v}_i)$ により説明される変数群 $\boldsymbol{v}_i$ の主効果の一つであって，線形回帰項の選択と分岐の選択が干渉しあって妥当な分岐の評価ができなくなる．特に分岐条件 $P_{l'}(\boldsymbol{v}_i)$ が $x_{ik} = 1$ のような単純な場合で，ダミー変数 x_{ik} が線形回帰項に採用されている場合には，その分岐による残差の改善がないことになる．つまり，親ノードで線形回帰項を採用することにより，子ノードへの分岐で説明できる主効果を基準とした適合度評価は利用できなくなる．

そこで，親ノードでの分岐条件を探索する際には，両子ノードでの線形回帰項 f_{2l}, f_{2l+1} までの説明能力を評価したものを分岐の適合度とする．つまり，$\hat{y}_{i2l}, \hat{y}_{i2l+1}$ を用いた際の残差平方和

$$RSS_N(l, \{\delta_{l'}\}, V_l, \{V_{l'}\}) = \sum_{i:\delta_l(\boldsymbol{v}_i)=1} (y_i - \{\delta_{2l}(\boldsymbol{v}_i)\hat{y}_{i2l} + \delta_{2l+1}(\boldsymbol{v}_i)\hat{y}_{i2l+1}\})^2 \tag{10}$$

を適合度とする．

両子ノードでの線形回帰項を考えることにより，前節の条件 1 により，親ノードでの線形回帰項 f_l を修正する必要が生ずることがある．子ノードの作成に伴い f_l を f'_l に変更しているとすれば，分岐することによって生ずる適合度の改善は以下となる．これは，共変量 V_{2l}, V_{2l+1} と分岐条件との交互作用項を回帰に投入することによる適合度の改善量となる．

$$\begin{aligned}
&RSS_L(l, V_l) - RSS_N(l, \{\delta_{l'}\}, V_l, \{V_{l'}\}) \\
&= \sum_{i:\delta_l(\boldsymbol{v}_i)=1} (\hat{y}_{il} - \{\delta_{2l}(\boldsymbol{v}_i)\hat{y}_{i2l} + \delta_{2l+1}(\boldsymbol{v}_i)\hat{y}_{i2l+1}\})^2 \\
&= \sum_{i:\delta_l(\boldsymbol{v}_i)=1} \Big(\alpha_l + f_l(\boldsymbol{v}_i) - f'_l(\boldsymbol{v}_i) - \delta_{2l}(\boldsymbol{v}_i)\{\alpha_{2l} + f_{2l}(\boldsymbol{v}_i)\} - \delta_{2l+1}(\boldsymbol{v}_i)\{\alpha_{2l+1} + f_{2l+1}(\boldsymbol{v}_i)\}\Big)^2
\end{aligned} \tag{11}$$

つまり，ノード l において，ノード l での主な主効果を除いた上で，その分岐と共変量との交互作用効果の大きさを基準としているという見方もできる．線形回帰項を採用する場合，分岐はその両サンプル集団の主効果の異質性を説明するためにのみ利用されるべきという考え方を実現したものとなる．

また，子ノードでの線形回帰項は，子ノードで用意される分岐の主効果と解釈できるので，この基準は，従来の主効果基準より 1 つ深い分岐まで先読みした分岐条件の探索基準ともいえる．

3.3. MDL 基準

前節の交互作用基準により，適合度の改善程度は評価できる．しかし，分岐候補を検討する際に必要な線形回帰項の選択においては，任意の変量の採用は何がしかの適合度の改善を生むので，改善レベルに閾値を設けて変量の採用を適切に限定して，暫定モデルの大きさを適当に抑える必要がある．そこで，MDL 基準に基づく制限を与えることとする．

Rissanen (1978, 1983, 1984) による MDL 原理はデータ圧縮理論を基礎としたモデル選択原理で，“ある確率モデルのもとでのデータの記述長” と “そのモデル自身の記述長” を合せて符号とする二段階符号化を考えたとき，最も短く符号化できるような確率モデルが，最良のモデルであると考えるものである．その主要項は Schwarz (1978) による BIC 基準と等価となるが，モデル選択バイアスの調整項に工夫ができる点が異なる（関，1996 などを参照）．ここでは，前節の考え方より与えられる交互作用残差を利用して MDL 基準を構成し，線形回帰項の変数選択基準および分岐基準を考える．

線形回帰項 f_l の評価に用いる葉 l に対する MDL を以下とする．

$$MDL_L(l, V_l) = \frac{RSS_L(l, V_l)}{2\sigma^2} + \frac{n_l}{2}\ln 2\pi\sigma^2 + \frac{|V_l|}{2}\ln n_l + \ln 2 \tag{12}$$

ただし，$RSS(l, V_l)$ はその葉の線形回帰項までを考慮した推定値 $\hat{y}_{il}$（(7) 式）を用いて計算される残差平方和 (8) 式であり，$V_l \subset \{1, \ldots, q\}$ は f_l に選択された変量の集合，$|V|$ は変量集合 V のランクを表わすものとする．

(12) 式では，第 2 項までがノードが葉であるときの負の対数尤度，第 3 項がそのノードでのパラメータ数に依存した，パラメータ最適化バイアスの罰金項，第 4 項がそのノードでの分岐の存在フラッグの記述長で，木の構造が大きくなることで生ずるモデル選択バイアスを補正するための罰金項となる．第 4 項については，Quinlan & Rivest (1989) の結果を簡略化して定めたもので，木の構造の記述方法をどう定めるかによる恣意性が残されている．なお，誤差分散 σ^2 は通常未知であるが，既知として MDL を与えている．

次に，分岐条件の判定に用いる MDL は，ノード l と子ノード l' (= $2l, 2l+1$) から構成される部分木モデルに基づき以下とする．第 3 項のパラメータ数の罰金項は，ノードごとに評価するものとした．

$$\begin{aligned} &MDL_N(l, \{\delta_{l'}\}, V_l, \{V_{l'}\}) \\ &= \frac{RSS_N(l, \{\delta_{l'}\}, V_l, \{V_{l'}\})}{2\sigma^2} + \frac{n_l}{2}\ln 2\pi\sigma^2 + \sum_{h=l,2l,2l+1}\frac{|V_h|}{2}\ln n_h + 3\ln 2 \end{aligned} \tag{13}$$

なお，(12), (13) 式の第 2 項は線形回帰項選択や分岐選択を行う際には定数として無視できる．よって，第 3 項以降のモデルの大きさへの罰金項と適合度との重み付けを定数 σ^2 が定めることとなる．以下では，この定数 σ^2 を $\dot{\sigma}^2$ と表す．$\dot{\sigma}^2$ を定める根拠がない場合には，後述のようにいくつかの $\dot{\sigma}^2$ を設定して分析を行い，得られたモデルを評価選択するものとする．

また，モデル全体の MDL は以下の再帰式により $MDL(1)$ として定義する．

$$MDL(l) = \begin{cases} \displaystyle\sum_{l'=2l,2l+1} MDL(l') + \frac{|V_l|}{2}\ln n_l + \ln 2 & \text{ノード } l \text{ が決定節の場合} \\ \text{(12) 式} & \text{ノード } l \text{ が葉の場合} \end{cases} \tag{14}$$

4. モデル推定算法

4.1. 木の構成アルゴリズム

提案モデルを推定するため，通常の樹形回帰分析と同様に再帰的にノードを構成していく以下の算法を提案する．基本的には，親ノードでの定数項を除く残差 $\tilde{y}_{il} = y_i - \hat{y}_{i\lfloor l/2 \rfloor} + \alpha_{i\lfloor l/2 \rfloor}$ に対し

$$\tilde{y}_{il} = f_l(\boldsymbol{v}_i) + \alpha_l + e_i \tag{15}$$

または

$$= f_l(\boldsymbol{v}_i) + \delta_{2l}(\boldsymbol{v}_i)\{\alpha_{2l} + f_{2l}(\boldsymbol{v}_i)\} + \delta_{2l+1}(\boldsymbol{v}_i)\{\alpha_{2l+1} + f_{2l+1}(\boldsymbol{v}_i)\} + e_i \tag{16}$$

$$= \sum_{j \in J(V_l)} \beta_{lj} x_{ij} + \sum_{l'=2l,2l+1} \delta_{l'}(\boldsymbol{v}_i) \left(\alpha_{l'} + \sum_{j \in J(V_{l'})} \beta_{l'j} x_{ij} \right) + e_i$$

なる回帰式の変数選択を実施し，$V_l, \delta_{2l}, V_{2l}, \delta_{2l+1}, V_{2l+1}$ を選ぶ手順となる．変数選択には，前述の MDL 基準で σ^2 を指定値 $\dot{\sigma}^2$ とした評価基準を用いて，変数増加減少法（奥野 他，1981 などを参照）を行う．

具体的には，ノード l において以下のステップで処理を行う．

1. $MDL_L(l, f_l)$ を基準とし変数増加法により (15) 式の f_l を暫定的に定める．
2. 全分岐条件候補 $\delta_{l'}$ ごとに以下を行い，(13) 式を基準とし，探索範囲で最適な (16) 式を決定する．
   (a) 初期状態として $V_{l'} = \phi$ とする（$l' = 2l, 2l+1$）．
   (b) $v \in V_l$ について，回帰式 f_l に変数群 $\{\delta_{l'}(\boldsymbol{v}_i)x_{ij} | j \in J(v), l' = 2l, 2l+1\}$ を追加して改善が見られるかを調べる．改善があれば，$V_{l'} \leftarrow V_{l'} \cup \{v\}, V_l \leftarrow V_l - \{v\}$ とする．
   (c) $V_l, V_{l'}$ に含まれない全変量 $v \in V$ について，変数群 $\{\delta_{l'}(\boldsymbol{v}_i)x_{ij} | j \in J(\boldsymbol{v})\}$ を追加して改善が見られるかを調べる．改善があれば，$V_{l'}$ に逐次追加する（$l' = 2l, 2l+1$）．
   (d) 以上で得られた回帰式について変数減少法を適用し，改善がある変量があれば削除する．
3. $MDL_L(l, f_l)$ が Step 2. の最良の $MDL_N(l, \{\delta_{l'}\}, V_l, \{V_{l'}\})$ より小さければ分岐をせず，Step 1. の f_l を確定し，ノード l での処理を終了する．
4. Step.3 で終了しなければ，$MDL_N(l, \{\delta_{l'}\}, V_l, \{V_{l'}\})$ が最小となる $\delta_{l'}$ を分岐条件に決定し，このときの f_l をノード l の線形回帰項と確定する．
5. 分岐条件によりサンプルを分割し，両子ノードそれぞれについて，このアルゴリズムを再帰的に適用した後，ノード l での処理を終了する．

なお，上記のアルゴリズム中で回帰式を推定する場合には，葉に相当する定数項 α_l はモデルに含めて推定を行う．たとえば，ノード l での分岐を考慮するときには，分岐条件のダミー変数 $\delta_{l'}$ は必ず採用するようにする．

4.2. 線形回帰項と分岐条件の選択における付加的条件

含まれるサンプル数が少ない葉を生成することは，回帰木生成に用いるサンプルセットへの過度な適合を生み，汎化誤差を大きくする原因となる．また，線形回帰項に関しても，ある葉で因子変量を線形回帰項として採用することは，その葉にその因子変量で分岐を用意することに等しいから，同様の制限を設けることが考えられる．そこで，ノード l で分岐条件や線形回帰項を選

択する際に，以下の制限を設ける．なお，以下の 2, 3 の条件は S-plus の tree 関数（Chambers & Hastie, 1992 を参照）で採用されている．

1. 分岐条件 $P_l(\boldsymbol{v}_i)$ としては，最も基本的な，$\boldsymbol{v}_i$ のうち一因子変量のみに依存したものを，本論文での対象とする．v_{ij} が名義因子変量の場合には，水準を 2 群に分ける全ての分岐を候補とし，v_{ij} が順序変量の場合には，水準の順序が連続する分割のみを候補とする．l 水準の場合，それぞれ 2^{l-1}，$l-1$ 通りの分岐条件が候補となる．
2. サンプル数 n_l が `minsize` 以下のノードは決定節としない．つまり，分岐を行なわず葉にする．
3. サンプル数 `mincut` 以下のノードを生ずる分岐条件は利用しない．つまり，分岐条件 $\delta_{2l}, \delta_{2l+1}$ において分岐を行なった際に，$(n_{2l} < \texttt{mincut}) \vee (n_{2l+1} < \texttt{mincut})$ が真ならば，この分岐条件はノード l における分岐条件の候補から外す．
4. 因子変量については，因子の水準ごとのサンプル数の中に `minreg`= M 以下のものがあるなら，その因子変量 $\boldsymbol{v}$ を線形回帰項として採用しない．例えば，変量 v のダミー変数の作成に treatment 対比を用いているとすれば，ノード l の所属サンプルで $x_{ik} = 1$ となるサンプル数を n_{lj} として，$\bigwedge_{j \in J(v)} (n_{lj} \geq M) \bigwedge (n_l - \sum_{j \in J(v)} n_{lj} \geq M)$ が真である説明変量のみを選択対象とする．

5. 数 値 実 験

5.1. 実験モデル

提案法について，C 言語によりプログラムを作成し，S-plus 環境上で利用できるようにした．提案算法がどのような性能をもつかを確かめるために次の 4 つのモデルを設定し，数値実験を行う．

まず，共変量として，2 水準 $\{-1, 1\}$ の因子変量 $\{x_1, x_3, x_5\}$，3 水準 $\{-1, 0, 1\}$ の因子変量 $\{x_2, x_4, x_6\}$ の 6 つを考え，それぞれは等確率の多項分布に独立に従うものとする．応答変量は，この共変量から以下のモデルに従い発生するものとする．ただし，$\epsilon_i \sim N(0, \sigma^2)$ であり，$\sigma^2 = 1.0, 0.09, 0.01$ とする．データセットのサンプル数 n を 1000 とした．

モデル 1 ：線形回帰モデルで説明できる主効果だけのモデルで分岐が不要な場合．分岐を行わないモデルが選択できるかが問題となる．

$$y_i = \epsilon_i + 2x_{i6} + x_{i4} + 0.7x_{i2} \tag{17}$$

モデル 2 ：線形回帰項を含む樹形回帰モデルで，最も簡潔なモデルの一つ．単一の決定節の分岐が交互作用効果の大きさに基づき正しく選択できるかが問題となる．

$$y_i = \epsilon_i + 2x_{i6} + \begin{cases} I(x_{i1} = -1)(-0.7)x_{i2} \\ I(x_{i1} = +1)(+1.0)x_{i2} \end{cases} \tag{18}$$

モデル 3 ：モデル 2 に分岐と線形回帰項を追加し，複雑さを増したモデル．逐次分岐を進める方法が成功するかが問題となる．

$$y_i = \epsilon_i + 2x_{i6} + \left\{ \begin{array}{l} I(x_{i1} = -1)\left(+1.0x_{i2} + \left\{ \begin{array}{l} I(x_{i3} = -1)(-0.7)x_{i4} \\ I(x_{i3} = +1)(+0.5)x_{i4} \end{array} \right. \right) \\ I(x_{i1} = +1)\left(-0.7x_{i2} + \left\{ \begin{array}{l} I(x_{i3} = -1)(+0.3)x_{i4} \\ I(x_{i3} = +1)(-0.7)x_{i4} \end{array} \right. \right) \end{array} \right. \tag{19}$$

モデル 4 ：線形回帰項が葉にしか存在せず，しかも，大きな効果が一番深い分岐にあり，木の形もアンバランスなモデル．回帰木の深いところに大きな交互作用がある場合に木の構造を正しく推定できるかが問題となる．

$$y_i = \epsilon_i + \left\{ \begin{array}{l} I(x_{i1} = -1)\left(\left\{ \begin{array}{l} I(x_{i3} = -1)\left(\left\{ \begin{array}{l} I(x_{i5} = -1)(+2.0)x_{i2} \\ I(x_{i5} = +1)(-0.7)x_{i2} \end{array} \right. \right) \\ I(x_{i3} = +1)(+1.0)x_{i2} \end{array} \right. \right) \\ I(x_{i1} = +1)\left(\left\{ \begin{array}{l} I(x_{i5} = -1)(+0.7)x_{i2} \\ I(x_{i5} = +1)(0.0) \end{array} \right. \right) \end{array} \right. \tag{20}$$

表 1 にそれぞれのモデルの特徴をまとめる．なお，Δ^* をモデル中の最小の効果として，以下の記号を用いる．

$$\Delta^* = \min_{l,v} \sum_i \delta_l(\boldsymbol{v}_i) \sum_{j \in J(v)} (\beta_{lj} x_{ij})^2$$

表 1.　実験モデルの性質

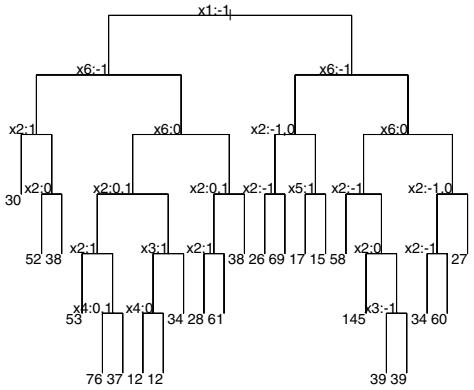
				σ^2 ごと理論決定係数		
	n^*_{leaf}	n^*_{par}	Δ^*	1.0	0.09	0.01
モデル 1	1	7	326.67	0.7850	0.9759	0.9973
モデル 2	2	8	163.33	0.8583	0.9854	0.9984
モデル 3	4	18	15.00	0.8865	0.9886	0.9987
モデル 4	5	13	40.83	0.7468	0.9704	0.9966

注　n^*_{leaf}：RLIM での理論葉数
n^*_{par}：RLIM での理論パラメータ数（葉での定数項 0 を含めている．）
理論決定係数：生成された 100 データでの平均値

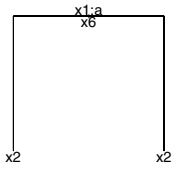
5.2.　従来法での推定結果

従来法として，S-plus（Version 6.0.4 Release 1 for Sun SPARC, SunOS 5.7: 2002）の関数 tree を用いて前節のモデルを推定した．`minsize` を 5，`mincut` を 10，分岐する最小の deviance `mindev` を 0.0 として最大の回帰木を構成し，クロスバリデーションで最良とされた葉数を選んでいる．形成された回帰木のうち $\sigma^2 = 1.0$ の場合の例を図 2 の (a), (c) に示す．図よりまず，tree 関数が全ての効果を分岐で説明しようとして複雑な木を推定していることがわかる．特に，通常の線形回帰が適切なモデル 1 では，単純な真のモデルに対し複雑な構造を与えることで，推定上問題が多いことが予想される．

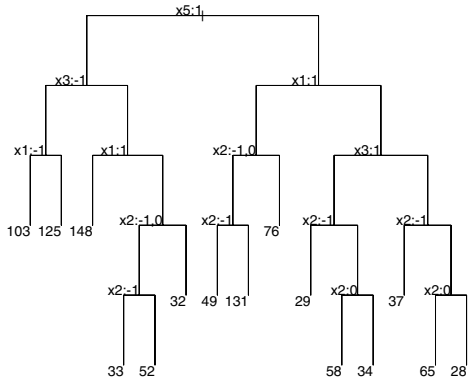
結果の数値的な評価結果を表 2 に示す．誤差分散が $\sigma^2 = 0.01$ と小さい場合には，標本決定係数は表 1 に示す理論値とほぼ等しく，予測能力はあることがわかる．ただし，線形回帰項が多い



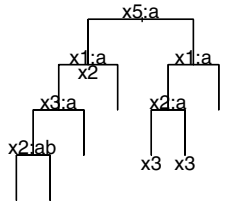
(a) モデル 2 従来法



(b) モデル 2 提案法($\dot{\sigma}^2 = \sigma_A^2$)



(c) モデル 4 従来法



(d) モデル 4 提案法($\dot{\sigma}^2 = \sigma_A^2$)

図 (a), (c) については，葉に所属サンプル数を示した．

図 2.　解析結果例($\sigma^2 = 1.0$)

表 2.　従来法数値実験結果例

モデル	σ^2	n_{leaf}	$\dot{n}^*_{leaf}$	$\hat{\sigma}^2$	標本決定係数	MSE
1	0.01	27	27	0.0090	0.9967	0.0003
2	0.01	19	18	0.0090	0.9982	0.0003
3	0.01	85	108	0.0261	0.9965	0.0179
4	0.01	18	20	0.0091	0.9965	0.0003
1	1.0	47	27	0.9357	0.7456	0.1103
2	1.0	23	18	0.9906	0.8401	0.0487
3	1.0	43	108	1.0130	0.8775	0.1935
4	1.0	15	20	1.0311	0.7084	0.0353

注　n_{leaf}: 推定された樹形モデルでの葉の数
$\dot{n}^*_{leaf}$: 従来の樹形モデルで表現した場合の理論上の葉の数
MSE: 誤差平方和 $\frac{1}{n}\sum_i (y_i^* - \hat{y}_i)^2$

ためパラメータ数の多いモデル 3 で，真値との誤差が多く若干効果の見落としがあることがわかる．これは，分岐が深くなり小さな葉を作らないという分岐の制約条件のために表現できていない効果があるためと考えられる．

一方，誤差分散が $\sigma^2 = 1.0$ の場合には，単純なモデルでは真のモデルより多くの枝が生成され，汎化誤差が生じている．また，複雑なモデルでは真のモデルより少ない枝が生成され，小さな効果が誤差に埋もれて発見できていないことがわかる．標本決定係数は理論値より少し劣るレ

ベルであり，特に理論葉数の多いモデルで真値との誤差 MSE が大きくなっている．

5.3. 提案法実験結果

提案法により，真のモデルが推定できるかを推定実験を行い評価した．提案算法では，線形回帰項の選択に (12) 式の基準，分岐条件の選択に (13) 式の基準を用いるが，この際指定する必要のある誤差分散 $\dot{\sigma}^2$ としては，以下の三通りを用いた．

1. 応答変量の全変動 $\hat{\sigma}_A^2 = \frac{1}{n-1}\sum_i\{y_i - \bar{y}\}^2$．ただし，$\bar{y}$ は y_i の平均．
2. 全変量を用いた線形回帰モデルでの残差 $\hat{\sigma}_L^2 = \frac{1}{n-n_{par}}\sum_i\{y_i - \sum_{j=0}^{p}\beta_j x_{ij}\}^2$．ただし，ここで n_{par} は線形回帰モデルでのパラメータ数．
3. 真の誤差分散 σ^2．

また，分岐に関する付加条件としては，`minsize`= 50，`mincut`= 25，`minreg`= 10 を用いた．モデル，誤差分散のそれぞれの組み合わせについて，100 セットのデータを生成し，$\dot{\sigma}^2$ 三種類について推定をおこなった．

実験結果を表 3 に示す．また，形成された回帰木のうち平均値に近い決定係数となる回帰木の例を図 2 の (b), (d) に示す．回帰木では，各ノードの上側に分岐条件が示され，下側に線形項に選択された説明変量が示されている．

モデル 1, 2 の簡潔なモデルでは，$\dot{\sigma}^2$ として何を与えても，モデル 2 の $\sigma^2 = 1.0, \dot{\sigma}^2 = \sigma_A^2$ を除けば，決定係数の有効数字 3 桁程度は理論値に一致し，効果推定率，同形率ともに高く，真のモデルを推定できている．しかし，モデル 3, 4 の複雑なモデルでは，かなり推定に失敗している場合がある．

まず，$\dot{\sigma}^2 = \sigma_A^2$ の場合は，$\dot{\sigma}^2$ を過大評価している場合に相当するが，この場合，決定係数が理論値を下回り，効果推定率も低くなっている．これは，n_{par} の値から判るように得られた回帰木が真のモデルより小さいためであると考えられる．つまり，基準に用いる誤差分散が大きいと，罰金項の比重が重くなり，必要以上に簡潔なモデルを推定してしまうことがわかる．$\dot{\sigma}^2 = \sigma_L^2$ の場合も，モデル 3, 4 では交互作用を考えない線形モデル残差を用いることにより，$\dot{\sigma}^2$ が過大評価され，同様の傾向が若干ある．一方，$\dot{\sigma}^2 = \sigma^2$ とした場合には，若干，大き目のモデルを推定する傾向があるようで，分岐数やパラメータ数が理論値より多くなっている．提案算法における MDL 基準では，木の構造の大きさの罰金項を導入しているが，それが不十分であるのかもしれない．全体として，$\dot{\sigma}^2$ に何を用いるかで，推定されたモデルの大きさにはかなりの違いがあることとなる．

モデル 4 のように，深い分岐の下に線形回帰項がある回帰木モデルの場合，分岐変数の採用順番が元モデルと異なってくることにより，同形率が低くなる．また，元モデルの葉が分割されて表現される推定モデルが得られる場合がある．この場合でも，サンプル数が十分あれば，全ての平均値効果を見つけてはいる．

以上の実験から提案法について，今回用意した程度のモデルについては，次の点がわかった．

1. 実験した規模のモデルでは，誤差分散の見積もりが正しければ，主な要因効果を発見でき，良好な予測を与える．特に，モデル 3 のように共通効果の多いモデルでは分岐が少なくて済むため，樹形モデルにくらべ，的確な推定が可能となる．

表 3.　数値実験結果（誤差分散既知）

モデル	σ^2	$\hat{\sigma}^2$	n_{leaf}	n_{par}	MDL	決定係数	$\hat{\sigma}^2$	MSE	同形率	効果推定率
1	1.0	$\hat{\sigma}^2_A$	1.00	7.00	1817.352	0.7856	0.9985	0.0071	1.00	1.00
		$\hat{\sigma}^2_L$	1.01	7.04	1439.017	0.7857	0.9983	0.0074	0.99	1.00
		σ^2	1.01	7.04	1439.562	0.7857	0.9983	0.0074	0.99	1.00
	0.09	$\hat{\sigma}^2_A$	1.00	7.00	1615.466	0.9760	0.0902	0.0006	1.00	1.00
		$\hat{\sigma}^2_L$	1.00	7.02	236.761	0.9760	0.0902	0.0006	1.00	1.00
		σ^2	1.00	7.02	237.260	0.9760	0.0902	0.0006	1.00	1.00
	0.01	$\hat{\sigma}^2_A$	1.00	7.00	1594.266	0.9973	0.0100	0.0001	1.00	1.00
		$\hat{\sigma}^2_L$	1.00	7.03	−861.178	0.9973	0.0100	0.0001	1.00	1.00
		σ^2	1.00	7.03	−860.745	0.9973	0.0100	0.0001	1.00	1.00
2	1.0	$\hat{\sigma}^2_A$	2.50	7.38	1993.026	0.8565	1.0174	0.0344	0.44	0.99
		$\hat{\sigma}^2_L$	2.02	8.02	1476.302	0.8593	0.9985	0.0082	0.98	1.00
		σ^2	2.12	8.19	1441.810	0.8594	0.9974	0.0094	0.91	1.00
	0.09	$\hat{\sigma}^2_A$	2.01	7.99	1862.260	0.9854	0.0904	0.0011	0.99	1.00
		$\hat{\sigma}^2_L$	2.00	8.00	741.537	0.9854	0.0902	0.0007	1.00	1.00
		σ^2	2.04	8.51	240.152	0.9855	0.0901	0.0008	0.96	1.00
	0.01	$\hat{\sigma}^2_A$	2.00	8.00	1849.070	0.9984	0.0100	0.0001	1.00	1.00
		$\hat{\sigma}^2_L$	2.00	8.00	595.823	0.9984	0.0100	0.0001	1.00	1.00
		σ^2	2.08	9.71	−855.247	0.9984	0.0100	0.0001	0.93	1.00
3	1.0	$\hat{\sigma}^2_A$	3.44	9.97	2108.421	0.8661	1.1909	0.2251	0.00	0.00
		$\hat{\sigma}^2_L$	4.29	14.21	1668.906	0.8847	1.0296	0.0620	0.01	0.00
		σ^2	4.79	18.05	1468.212	0.8885	0.9999	0.0285	0.24	0.17
	0.09	$\hat{\sigma}^2_A$	3.13	10.04	2005.482	0.9629	0.2960	0.2081	0.00	0.00
		$\hat{\sigma}^2_L$	4.55	15.87	1334.892	0.9868	0.1056	0.0184	0.45	0.18
		σ^2	4.24	20.54	269.713	0.9888	0.0901	0.0021	0.80	1.00
	0.01	$\hat{\sigma}^2_A$	3.01	10.01	1996.223	0.9722	0.2198	0.2081	0.00	0.00
		$\hat{\sigma}^2_L$	4.12	16.12	1295.973	0.9969	0.0244	0.0143	0.88	0.12
		σ^2	4.08	26.28	−817.157	0.9988	0.0100	0.0003	0.94	1.00
4	1.0	$\hat{\sigma}^2_A$	6.76	11.67	1776.273	0.7372	1.0517	0.0835	0.00	0.00
		$\hat{\sigma}^2_L$	7.42	14.70	1611.205	0.7469	1.0150	0.0439	0.00	0.50
		σ^2	7.92	18.19	1460.593	0.7533	0.9928	0.0291	0.00	0.78
	0.09	$\hat{\sigma}^2_A$	6.53	13.16	1537.645	0.9581	0.1298	0.0440	0.00	0.25
		$\hat{\sigma}^2_L$	7.01	16.97	1200.514	0.9707	0.0903	0.0017	0.00	1.00
		σ^2	8.16	19.45	261.679	0.9710	0.0897	0.0028	0.00	1.00
	0.01	$\hat{\sigma}^2_A$	6.43	13.54	1512.845	0.9837	0.0497	0.0401	0.00	0.31
		$\hat{\sigma}^2_L$	7.01	16.99	1148.303	0.9966	0.0101	0.0002	0.00	1.00
		σ^2	8.12	19.86	−835.474	0.9967	0.0100	0.0003	0.00	1.00

注　表中の数字はそれぞれ以下の 100 推定実験での平均値.

1. n_{leaf}: 推定された木の葉の数.
2. n_{par}: 推定されたモデルのパラメータ数.
3. MDL: 生成された回帰木における MDL (14) 式.
4. 決定係数：応答変量の観測値と推定値の相関係数の二乗.
5. $\hat{\sigma}^2$: 残差分散 $\hat{\sigma}^2 = \frac{1}{n-n_{par}} \sum_i (y_i - \hat{y}_i)^2$.
6. MSE: 誤差平方和 $\frac{1}{n} \sum_i (y_i^* - \hat{y}_i)^2$.
7. 同形率：真のモデルを表す回帰木と一致した回帰木の生成率．葉の数および線形項に選択された説明変量の数を比較し，どちらかの数が真のモデルと異なっているとき，推定されたモデルは回帰木の形が誤っていると判定した．そのため，推定された効果が誤っているとは限らないが，最も簡潔な木構造モデルを与えていないといえる.
8. 効果推定率：誤差平方和について $MSE < \sigma^2/(n/n_{leaf}) + 0.8\Delta^*/n$ となった率．存在する効果が推定できたと考えられる推定結果の割合.

2. 通常の線形モデル(モデル 1)，一分岐のモデル(モデル 2)に対し，要因効果を分岐と線形回帰項の何れで説明すると簡潔な表現が与えられるかをほぼ的確に判断して，推定を行う．
3. $\dot{\sigma}^2$ の指定値が推定結果のモデルの簡潔さに大きな影響がある．実際に用いる場合には，適切な値を探索する必要があると思われる．
4. 分岐レベルが二レベル以上のモデル(モデル 3, 4)に対しては，変量の効果は発見できても，最も簡潔な木構造を発見するとは限らない．

6. 高齢者介護時間データへの適用

介護保険制度における要介護認定一次判定方式には，施設における介護時間を樹形回帰モデルを用いて推定したモデルが用いられている．そこで用いられている高齢者介護時間データ(関・筒井・宮野，2000 を参照)について，提案する再帰分割線形モデルを適用し，実データでの提案手法の検証を行う．データセットのうち，一日当りの食事の介護時間を応答変量とし，73 のカテゴリカルな順序変量と，それらを双対尺度法で集約した 7 つの中間評価項目の合計 80 変量を共変量とした．サンプル数は 2896 高齢者であるが，そこから 2000 サンプルをランダムに抽出してモデル推定用のデータとし，残り 896 サンプルを汎化誤差の評価に用いた．

線形回帰分析による結果と，通常の樹形回帰分析結果(S-plus tree 関数)，さらに提案法による評価結果を表 4 に示す．提案法では，$\dot{\sigma}^2$ の指定をいくつか試し，汎化誤差分散の少ないものを選んだ．また，提案法によるモデルを図 3 と (21) 式 [$\dot{\sigma}^2 = 3.5\sigma_L^2$], (22) 式 [$\dot{\sigma}^2 = 3\sigma_L^2$] に示す．

表 4.　数値実験結果(誤差分散既知)

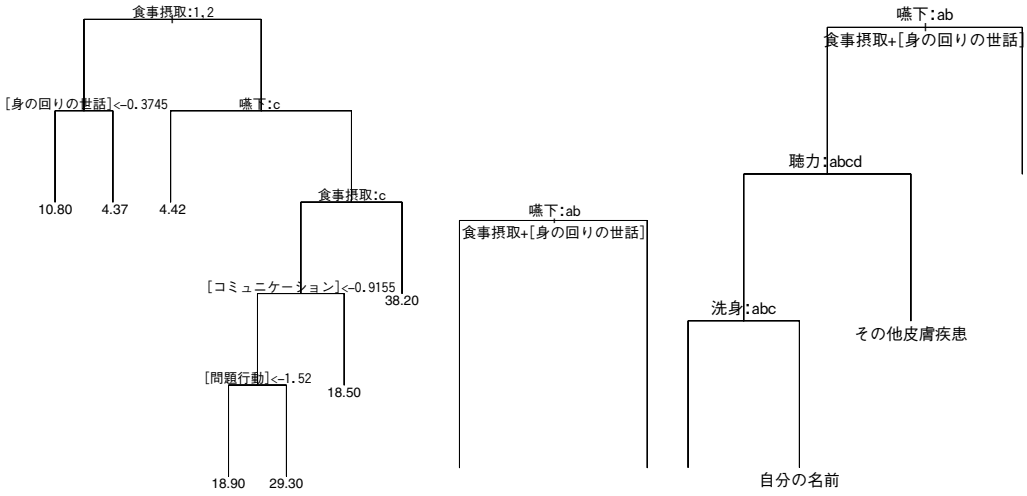
モデル	方法	$\dot{\sigma}^2$	n_{leaf}	n_{par}	残差分散	汎化誤差分散
1	線形回帰	–	1	150	87.125	100.074
2	tree 関数	–	7	7	93.892	100.702
3	提案法	$3.5\sigma_L^2$	2	6	97.037	97.272
4	提案法	$3\sigma_L^2$	4	10	94.285	99.232
5	提案法(再計算)	$3\sigma_L^2$	4	10	94.281	97.350
6	提案法(σ^2 未知)	–	41	183	69.330	128.176

注　モデル 6 については，次節で言及．

提案法の $\dot{\sigma}^2 = 3.5\sigma_L^2$ の場合のモデルを，(22) 式の形式で分岐条件変数を含む線形回帰として再計算した結果を表 4 のモデル 5 に示す．(23) 式のような回帰係数の調整結果が得られ，若干，残差分散，汎化誤差分散ともに改善している．

結局，汎化誤差分散から考えてモデル 3 または 5 がほぼ等価なモデルとして，推奨できると考えられる．なお，モデル 3 は分岐が一つなので，木を推定する算法の中で実施される線形回帰が，モデル全体となっており，再計算する必要がない．

$$g_1(\boldsymbol{v}_i) = \text{食事摂取 (自立: 0, 見守り: 2.621, 一部介助: 13.28, 全介助: 27.21)} \\ -2.688 \times [\text{身の回りの世話}] + \begin{cases} I(\text{嚥下} = \text{不能})(-25.56) \\ I(\text{嚥下} \neq \text{不能})(7.179) \end{cases} \tag{21}$$



(a) 従来法　(b) 提案法($\dot{\sigma}^2 = 3\sigma_L^2$)　(c) 提案法($\dot{\sigma}^2 = 3.5\sigma_L^2$)

図 3.　介護時間モデル推定結果

$$
\begin{aligned}
g_1(\boldsymbol{v}_i) = {} & \text{食事摂取 (自立: 0, 見守り: 2.621, 一部介助: 13.28, 全介助: 27.21)} \\
& - 2.688 \times [\text{身の回りの世話}] \\
& + \begin{cases} I(\text{嚥下} = \text{不能})(-25.56) \\ I(\text{嚥下} \neq \text{不能}) \begin{cases} I(\text{聴力} = \text{判断不能})(16.17 + \text{その他皮膚疾患 (あり: 0, なし: } -20.79)) \\ I(\text{聴力} \neq \text{判断不能}) \\ \quad \times \begin{cases} I(\text{洗身} = \text{行わない}) \\ \quad \times(9.399 + \text{自分の名前 (いえる: 0, いえない: } -21.47)) \\ I(\text{洗身} \neq \text{行わない})(7.178) \end{cases} \end{cases} \end{cases}
\end{aligned}
\tag{22}
$$

$$
\begin{aligned}
g_1(\boldsymbol{v}_i) = {} & \text{食事摂取 (自立: 0, 見守り: 2.674, 一部介助: 13.412, 全介助: 27.424)} \\
& - 2.676 \times [\text{身の回りの世話}] \\
& + \begin{cases} I(\text{嚥下} = \text{不能})(-32.890) \\ I(\text{嚥下} \neq \text{不能}) \begin{cases} I(\text{聴力} = \text{判断不能})(8.868 + \text{その他皮膚疾患 (あり: 0, なし: } -20.814)) \\ I(\text{聴力} \neq \text{判断不能}) \\ \quad \times \begin{cases} I(\text{洗身} = \text{行わない}) \\ \quad \times(2.212 + \text{自分の名前 (いえる : 0, いえない : } -21.595)) \\ I(\text{洗身} \neq \text{行わない})(7.138) \end{cases} \end{cases} \end{cases}
\end{aligned}
\tag{23}
$$

7. アルゴリズムの評価と今後の課題

7.1. 変数選択の方法と計算量

提案法は，分岐条件の評価として分割による交互作用を利用し，できるだけ簡潔な木構造を与

えるサンプルセットの分割を探索する点が特徴となる．ただし，提案法が機能するためには，各決定節での線形回帰項の的確な変数選択が自動的に出来ることが必要となる．

この際，分岐により対象サンプル数が減少すると，必然的あるいは偶然に別名関係が発生することがある．そこで，作成したプロトタイププログラムでは，線形回帰の計算時にランク落ちによる問題が発生するおそれを特異値分解法を用いることで避けている．`minreg` 以外にも変数のモデルへの投入方法を再検討し，変数選択の効率的な算法を組込む必要がある．

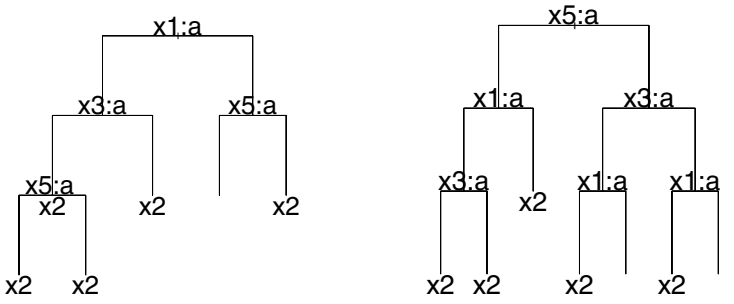
なお，上記の理由を除いても，提案算法は，かなりの計算時間が必要なアルゴリズムとなっている．この算法の一決定節当りの計算量を評価してみると，共変量の数が q であり，各共変量が等しく c 水準であるとすると，分岐条件の評価回数を単位として $O(q \cdot 2^{(c-1)})$ となる．いま，共変量を 2 水準に限定すると，これは $O(q)$ となるが，回帰木を生成するための計算量は，生成された回帰木のノード数を r として，$O(qr)$ となる．これに加え，線形回帰項の変数選択が各分岐条件において必ず行われる．仮に変数選択が $O(p^3)$ で行われるならば，回帰木の生成の計算量は $O(qp^3r)$ となる．共変量や因子変量の水準数が多くなると，計算量が急速に増加することがわかる．

7.2. 推定結果の評価方法

提案モデルは前節で例を示したように分岐条件の特性変数を用いて，標準的な線形モデルへの変換が可能である．分岐が 1 つだけの木の構造なら推定結果は変らないが，それ以上深い木の結果については，分岐の特性関数を利用した線形モデルとして推定し直すことによって，若干，推定精度の向上がある．さらに，標準的な回帰診断の手法と固有技術的検討を合せて変数選択を実施することでモデルの改善を図ることが可能であり，また，それが必要と考えられる．

さらに，5.3 節での実験モデル 4 で，効果の推定には成功しても同形率が一貫して 0.0 であったように，提案法で得られた結果により最も簡潔な木の構造が得られるとは限らない．図 4 に示すように，構造が異なり，葉の数の異なるモデルが同じ効果をもつ場合がある．これは，効果が与えられても，提案モデル (5) 式の表現が一意にならないためである．このような場合に対処するためには，得られた回帰木で類似した部分木が現れる部分などに注目し，モデルの整理を検討する必要がある．

なお，今回の適用例データに関しては，評価基準で指定した分散 $\dot{\sigma}^2$ より残差分散 $\hat{\sigma}^2$ が小さめのモデルが汎化誤差分散からみて適当とされた．$\dot{\sigma}^2$ を大きく設定することで，従来モデルに比べて汎化誤差の少ないモデルを選択することができたが，MDL 基準のモデル構造の罰金項の設定



(a) 真のモデル　(b) 解析結果例（$\sigma^2 = 0.01, \dot{\sigma}^2 = \sigma^2$）

図 4. モデル 4

方法や重みの調整方法が今後の課題として残されている．

7.3. 誤差分散の仮定

本論文では，誤差分散が木全体で一定既知と仮定してモデル選択基準を定め，推定を行う際には，いくつかの誤差分散値 $\hat{\sigma}^2$ を指定し，得られたモデルを比較選択する方法を提案した．これに対し，誤差分散未知の仮定の下でのモデル選択基準を利用することも考えられる．ただし，誤差分散の推定量が各分岐でのモデル選択を行う過程で必要であり，Greedy な算法の過程でこの推定量を木全体で一貫して推定しながら，木構造モデルを推定することは難しい．

そのため，誤差が葉ごとに異なる分散をもつ正規分布 $N(0, \sigma_l^2)$ に従い，その分散が未知と仮定することが考えられる．このときモデルは，

$$Y_i = g_1^*(\boldsymbol{v}_i)$$

で，ノード l を root とする部分木における効果の再帰的表現を以下の $g_l^*(\boldsymbol{v}_i)$ とすることになる．

$$g_l^*(\boldsymbol{v}_i) = f_l^*(\boldsymbol{v}_i) + \begin{cases} \sum_{l'=2l,2l+1} \delta_{l'}^*(\boldsymbol{v}_i) g_{l'}^*(\boldsymbol{v}_i) & \text{分岐する場合} \\ \alpha_l^* + \epsilon_i & \text{分岐しない場合}(\epsilon_i \sim N(0, \sigma_l^2)) \end{cases} \tag{24}$$

この場合には線形回帰項の選択に (25) 式の MDL 基準，分岐条件の選択に (26) 式の MDL 基準を用いることとなる．

$$MDL_L(l, V_l) = \frac{n_l}{2}\left(1 + \ln 2\pi \frac{RSS_L(l, V_l)}{n_l}\right) + \frac{|V_l| + 1}{2} \ln n_l + \ln 2 \tag{25}$$

$$\begin{aligned} MDL_N(l, \{\delta_{l'}\}, V_l, \{V_{l'}\}) = & \frac{n_l}{2}\left(1 + \ln 2\pi \frac{RSS_N(l, \{\delta_{l'}\}, V_l, \{V_{l'}\})}{n_l}\right) \\ & + \frac{|V_l| + 1}{2} \ln n_l + \sum_{h=2l,2l+1} \frac{|V_h|}{2} \ln n_h + 3 \ln 2 \end{aligned} \tag{26}$$

これを用い 5 節と同じモデルに対し，数値実験を行った結果を表 5 に示す．この数値実験のよ

表 5. 数値実験結果（誤差分散未知）

モデル	σ^2	n_{leaf}	n_{par}	MDL	決定係数	$\hat{\sigma}^2$	MSE	同形率	効果推定率
	1.0	1.00	7.01	1442.476	0.7857	0.9985	0.0072	1.00	1.00
1	0.09	1.00	7.02	240.213	0.9760	0.0902	0.0006	1.00	1.00
	0.01	1.00	7.03	−857.726	0.9973	0.0100	0.0001	1.00	1.00
	1.0	2.02	8.09	1446.980	0.8593	0.9981	0.0086	0.98	1.00
2	0.09	2.00	8.03	244.559	0.9854	0.0901	0.0007	1.00	1.00
	0.01	2.01	8.05	−853.625	0.9984	0.0100	0.0001	0.99	1.00
	1.0	4.08	17.66	1475.567	0.8884	1.0001	0.0244	0.79	0.23
3	0.09	4.02	18.46	276.211	0.9887	0.0905	0.0020	0.98	1.00
	0.01	4.34	21.25	−813.569	0.9987	0.0101	0.0003	0.74	1.00
	1.0	6.99	17.42	1474.954	0.7517	0.9983	0.0230	0.00	0.86
4	0.09	7.16	17.48	271.662	0.9709	0.0899	0.0018	0.00	1.00
	0.01	7.11	17.49	−826.768	0.9967	0.0100	0.0002	0.00	1.00

注　数値の定義については，表 3 参照

うなバランスされたデータで誤差分布の仮定が正しいデータに対しては，5 節で示した分散既知とした場合と比べ，ほぼ同様か，より優れた予測力がこの分散未知の基準で得られることが判った．しかし，6 節の介護時間データに用いると，多くの細かい分岐を生じ，適切な結果が得られなかった(表 4, モデル 6)．これは，応答変数の値が正値に限定され，正規分布でなく歪んだ分布をもつことなどが原因と考えられる．葉ごとに分散を推定する場合，応答変量値がほとんど同じサンプルから構成される葉ができると (25) 式の適合度が劇的に改善する．そこで，介護時間のように 0 に近い値にかなりのサンプルがある場合，0 に近いデータを中心に細かな枝を生成することになったと考えられる．このような現象を考えると，5 節で示した分散既知とした場合が，誤差分布の仮定などに頑健な結果を与え，現状では有用であると考える．

8. おわりに

本研究では，従来の樹形回帰モデルに線形回帰項を追加することで，部分木に共通する主効果を発見しようとするモデルを提案した．数値実験や適用例より，提案法がかなり的確に効果を発見できることが確認できた．数千サンプルで因子変量を多く含む回帰問題において，主効果の異質性に基づく層別をした上での回帰モデルを探索する方法として，提案法は有望であると考える．

回帰木を生成するにあたって必要となる分岐基準としては，交互作用効果に注目した MDL 基準を用いたが，モデル選択バイアスのための罰金項をどうすべきかなどについては，線形回帰項として採用する変量の変数選択算法の改良とともに，今後の課題となる．また，今回の数値実験では，共変量が独立な場合についてしか行なっていないが，説明変量相互に相関のある場合などについての有効性を検証することが必要と考えられる．

謝　辞　本研究は高齢者介護時間の樹形モデルによる解析を行なった際に着想を得て行なわれた．この解析の中で有意義な議論を戴いた，国立保健医療科学院の筒井孝子氏，立命館大学の宮野尚哉氏に感謝する．また，本研究は科学研究費(課題番号 13680507)の援助を受けた．

参考文献

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984): Classification and Regression Trees, Wadsworth International Group, Belmont, C.A.

Chambers, J.M. and Hastie, T.J. eds. (1992): *Statistical Models in S*, Wadsworth & Brooks/Cole Advanced Books & Software, A Division of Wadsworth, Inc., Pacific Grove, C.A.

Dobra, A. and Gehrke, J. (2002): SECRET: A Scalable Linear Regression Tree Algorithm, Proceedings of SIGKDD'92. ACM.

Friedman, J.H. (1991): Multivariate adaptive regression splines, *Ann. Statist*., **19**, 1–67.

Murthy, S.K. (1998): Automatic construction of decision trees from data: A multi-disciplinary survey, *Data Mining and Knowledge Discovery* **2**, 345–389.

Karalic, A. (1992): *Employing Linear Regression in Regression Tree Leaves*, Proceedings of ECAI'92, 440–441, Wiley & Sons.

Loh, W.-Y (2002): Regression trees with unbiased variable selection and interaction detection, *Statistica Sinica* **12**, 361–386.

奥野 忠一，久米 均，芳賀 敏郎，吉澤 正 (1981): 多変量解析法，日科技連．

Quinlan, J.R. and Rivest, R.L. (1989): Inferring decision trees using the minimum description length principle, *Information and Computation* **80**, 227–248.

Quinlan, J.R. (1992): Learning with Continuous Classes, Proceedings of 5th Australian Joint Conference on Artificial Intel-

ligence, 323–348, World Scientific.
Quinlan, J.R. (1993): *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, California.
Rissanen, J. (1978): Modeling by shortest data description, *Automatica* **14**, 465–471.
Rissanen, J. (1983): A universal prior for integers and estimation by minimum description length, *The Annals of Statistics* **11**(2) 416–431.
Rissanen, J. (1984): Universal coding, information, prediction, and estimation, *IEEE Trans. on IT* **IT-30**, 629–636.
Schwarz, G. (1978): Estimating the dimension of a model, *Ann. Statist*. **6**, 461–464.
関 庸一 (1996): 経営工学におけるモデル選択，オペレーションズ・リサーチ 41(7), 387–391.
関 庸一，筒井 孝子，宮野 尚哉 (1999): 線形回帰項を含む樹形回帰モデル推定におけるモデル選択，修善寺，情報理論的学習理論ワークショップ予稿集，123–128.
関 庸一，筒井 孝子，宮野 尚哉 (2000): 要介護認定一次判定方式の基礎となった統計モデルの妥当性，応用統計学 29(2), 101–110.

（2003 年 6 月 26 日受付　2004 年 3 月 7 日最終修正　4 月 9 日採択）

著者連絡先：〒 376–8515 桐生市天神町 1–5–1
群馬大学情報工学科 関 庸一
E-mail: seki@cs.gunma-u.ac.jp
〒 171–8513 東京都豊島区南池袋 2–32–8
（株）SRA 野島 勇
E-mail: nojima@sra.co.jp

Japanese J. Appl. Statist. 33 (2) (2004), 111–130

Recursive Partitioning Linear Model Using Interaction Effect Criterion

Yoichi Seki[1,*] **and Isamu Nojima**[2]

[1] Department of Computer Science Gunma University
1–5–1 Tenjin-cho, Kiryu, Gunma 376–8515, Japan
[2] Software Research Associates, Inc.
2–32–8, Minami-Ikebukuro, Toshimaku, Tokyo 171–8513, Japan

Abstract

In this paper, we propose a method to construct tree regression models including linear regression terms. Ordinary tree regression models reliably detect high order interactions compared with multiple regression models, but they tend to make large, deep trees because they explain all covariate effects using only stratification of samples. In particular, if there are many effects common to all samples or some sample group, they estimate a tree with many almost identical subtrees. In order to avoid this redundancy, we propose a tree regression model that explains main effects by linear regression terms in each node, and heterogeneity of the effects by stratification. When we take this strategy, there may be a huge number of candidate models; hence model selection is one of the main tasks. Thus we propose an algorithm to estimate the models using a criterion based on the MDL principle. This criterion can be interpreted as a criterion to select split variables which have the maximum interaction effects. Finally, we demonstrate the efficiency of our method using numerical examples as well as real data on the amount of time caregivers provided individually to elderly persons.

Key words: data mining, linear regression term, MDL, tree regression analysis

* Corresponding author
E-mail address: seki@cs.gunma-u.ac.jp (Yoichi Seki)

Received June 26, 2003; Received in final form March 7, 2004; Accepted April 9, 2004.