JMP による対話的パーティショニング

SAS Institute Japan 株式会社 JMP ジャパン事業部

2009 年 5 月

１．はじめに

JMP では、メニュー「パーティション」により、決定木の分析を行うことができます。本文書は、このパーティションのメニューに関する技術的事項を述べます。

２．パーティションに関する Q&A

この章では、JMP のパーティションについての疑問を、Q&A 形式で回答します。

Q1. パーティションという名前の由来はなんですか。通常は決定木と呼ばれていますが。

A1. Chi-squared Automatic Interaction Detector(CHAID)の先駆的な文献である Kass and Hawkins[1]の一部分には、決定木分析のことを、"Recursive Partitioning"と呼んでいます。これを短くして、JMP では、Partition(パーティション)というメニュー名にしています。"パーティション"という言葉は、探偵が何かを発見するために、手がかりを使って捜索範囲を狭くするような感覚を受けますので。[2]

Q2. JMP のパーティションの特徴は?

A2. 次のような点が特徴として挙げられます。

- 目的変数（応答）は、連続でもカテゴリカルでも構いません。また、説明変数は、連続、カテゴリカルの変数を混在させることができます。
- グラフ機能が充実しており、対話的なパーティションをビジュアル的にサポートします。
- 分岐したくない列をロックする（分岐をさせないようにする）ことができます。
- 通常の多変量解析とは違い、欠測値があるデータでも、ランダムに分岐することにより、データの情報を生かすことができます。

Q3. パーティションは、CART や CHAID、C4.5、C5.0 のような分類アルゴリズムを用いているのでしょうか。

A3. 類似する点はいくつかありますが、JMP のパーティションでは、CART や CHAID のような分類アルゴリズムそのものを用いているわけではありません。

Q4. 停止基準などにより、自動的に分岐を行う方法はありますか。

A4. JMP 8 では、レポートの赤い三角ボタンより、**[K 分割交差検証]** を選択する、または、あらかじめ（検証データ用に）行を除外したときに、パーティションのレポート画面に**[実行]** ボタンが追加されます。**[実行]** ボタンを押すと、交差検証の R2 乗または除外した行に対する R2 乗が改善されなくなるまで分岐を自動的に行います。（図 1）

JMP 8 以前のバージョンでは、上記のような機能はありません。

---

[1] Hawkins, D.M. and Kass, G.V.(1982), "Automatic InteractionDetection," in Hawkins, D.M., ed., Topics in Applied Multivariate Analysis,267-302, Cambridge Univ Press:Cambridge.

[2] JMPer Cable Spring 2005 Issue 17 http://www.jmp.com/about/newsletters/jmpercable/backissues.shtml

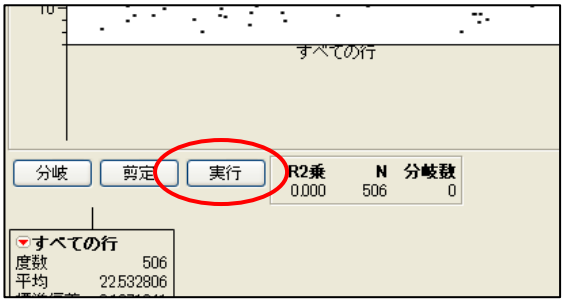


図 1

(JMP 全体の特徴でもありますが、)パーティションは、対話的に分析を行えることが特徴になります。そのため、**[実行]** ボタンを用いない限り、分岐は対話的に行われます。昨今のデータマイニングでは、決定木分析というと、モデルを作成し、予測をすることに重点をおくことが多いですが、JMP のパーティションでは、予測だけではなく、目的変数に影響する要因を見つける"要因分析"として用いることにも重点を置いています。要因をあれこれ見つける際、この対話的な分析が効力を発揮します。

３．パーティションの分岐基準

JMP のパーティションメニューでは、次の 2 種類の分岐方法があります。

- **[分岐統計量を最大化]**
  連続変数の場合は平方和の値、カテゴリの変数の場合は G^2 の値に基づいて分岐を行います。
- **[有意度を最大化]** (デフォルトの分岐方法)
  各分岐候補の有意度を計算して最適な分岐を決定します。

分岐基準の変更は、パーティションのレポートの左上にある赤い三角ボタンをクリックし、**[基準]** から変更できます。以下、それぞれの分岐基準の詳細を示します。

■分岐統計量を最大化

2 つの応答の差が最大になるように分岐候補を探します。

○応答が連続変数のとき

平方和(SS)が基準となります。レポートの「候補」の欄には、"候補 SS"と表示されます。

候補 SS は、JMP で分散分析を行った際に表示される、分散分析表の要因の平方和に相当します。

○応答がカテゴリ変数のとき

尤度比カイ 2 乗(G^2)が基準となります。レポートの「候補」の欄には、"候補 G^2"と表示されます。

候補 G^2 は、JMP でモザイク図(二変量の関係で X,Y にカテゴリ変数を選択した場合)を描いたとき、「検定」の欄に表示される尤度比のカイ 2 乗が該当します。

応答が連続変数の場合は"候補 SS"が、応答がカテゴリ変数の場合は"候補 G^2"が一番大きい項目で分岐されます。

■有意度を最大化

[分岐統計量を最大化]を基準とした場合に、水準数の多い変数が分岐候補になる傾向があり、これらを調整したのが、**[有意度を最大化]**という基準です。各分岐候補の有意度を計算して最適な分岐を決定します。候補のレポートには、「対数価値」という列があり、この列の値が一番大きい項目で分岐されます。

対数価値は、調整済み p 値を用いて、次のように計算されます。

$$対数価値 = -\log_{10}(調整済み\ p\ 値) \quad \cdots \quad (1)$$

調整済み p 値は、考えられる分岐候補の組み合わせ数を考慮した複雑な方法で算出され、水準数の多い X に有利になってしまう未調整の p 値に比べ、公正な分析になります。この手法については、下記のホワイトペーパーで検証されています。(英語)

「Monte Carlo Calibration of Distributions of Partition Statistics」

http://www.jmp.com/software/whitepapers/pdfs/montecarlocal.pdf

４．パーティションの分岐基準に対する具体例

この章では、3 章で説明したパーティションの分岐基準について、サンプルデータを用いた具体例を示します。

■応答がカテゴリ変数のとき

●使用する JMP のサンプルデータ　:**「車の調査.jmp」**

●パーティションでの列の指定:

[Y, 目的変数]:**「生産国」**

[X, 説明変数]:**「性別」、「年齢」、「タイプ」**

基準は、**「分岐統計量を最大化」**を選択します。

「すべての行」に対する「候補」を表示させたときのパーティションのレポートは、図 2 のようになります。

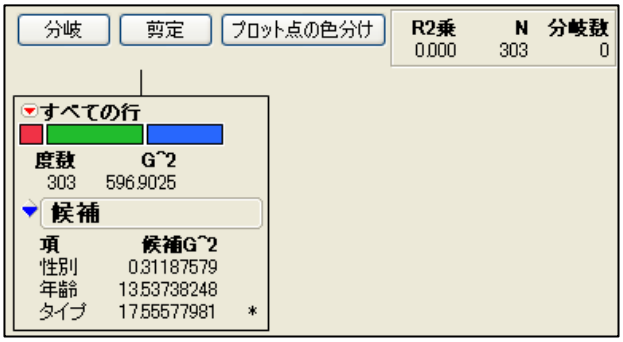


図 2

図 2 に表示されているそれぞれの項目に対する「候補 G^2」は次の要領で算出されます。

★「性別」に対して

「性別」は、”男性”、“女性”の 2 つのカテゴリを持ちます。そのため、性別を 2 つのグループで分岐するのであれば、”男性”のグループと”女性”のグループに分かれます。メニュー**[二変量の関係]** を用いて、[Y,目的変数] に**「生産国」**、[X,説明変数]に**「性別」**を選択して分析を行うと、図 3 の検定表が表示されます。ここに表示される尤度比カイ 2 乗（または、「(−1)*対数尤度」(=0.1559)の 2 倍）が、候補に表示されている性別の候補 G^2(=0.312)になります。

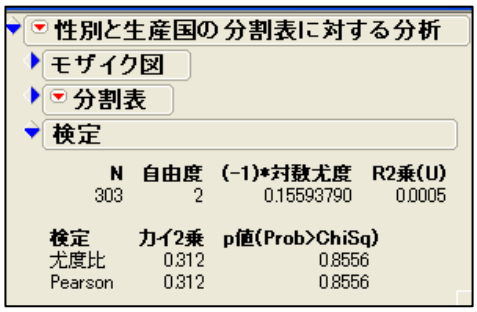


図 3

★「タイプ」に対して

「タイプ」は、”スポーツ”,”ファミリー”,”ワーク” の 3 つのカテゴリを持ちます。3 つのカテゴリを 2 つのグループに分岐する方法は次の 3 通りが考えられます。

a. （スポーツ、ファミリー） と （ワーク）
b. （スポーツ） と （ファミリー、ワーク）
c. （スポーツ、ワーク） と （ファミリー）

そのため、**「タイプ」**を上記 a,b,c のようにデータを 2 つのグループに分け、**「性別」**と同じ要領で二変量の関係の分析を実行します。

図 4 は、左から右へ、a,b,c の分析を行ったときの検定結果になります。この 3 つの中で、最も尤度比カイ 2 乗が大きいのは、b のときです。そのため、**「タイプ」**を(スポーツ)と(ファミリー、ワーク)の 2 つのグループに分けたときの尤度比カイ 2 乗が、候補に表示されているタイプの候補 G^2(=17.556)になります。

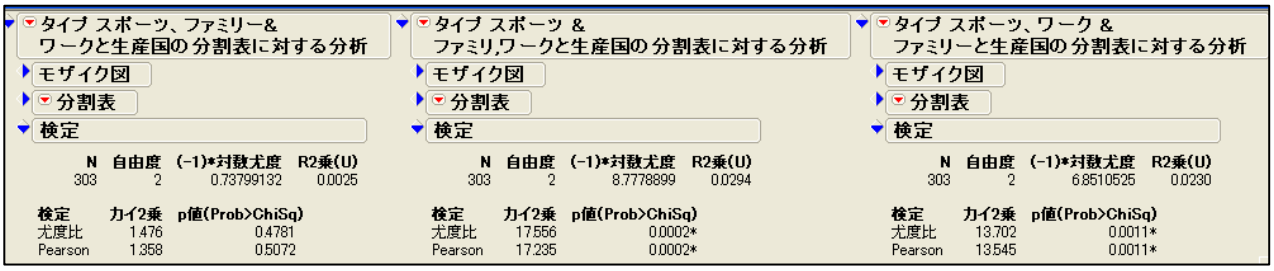


図 4

★「年齢」に対して

「年齢」は連続尺度で 18 から 60 までの値をとります。この範囲をある値を境にして 2 つにグループ分けし、同じ要領で尤度比カイ 2 乗を参照します。境界値を次々と変えていき、尤度比カイ 2 乗が最大になる境界値を見つけます。この例では、36 歳以上/未満

が境界値になり、このときの尤度比カイ 2 乗を求めると、図 5 のようになります。この値が、**「年齢」**の候補 G^2(=13.537)になります。

図 5

図 2 を参照しますと、候補 G^2 の値が最も大きいのは、**「タイプ」**です。そのため、**[分岐]** ボタンを押すと、「タイプ(スポーツ)」と「タイプ(ワーク、ファミリー)」で分岐します。(図 6)

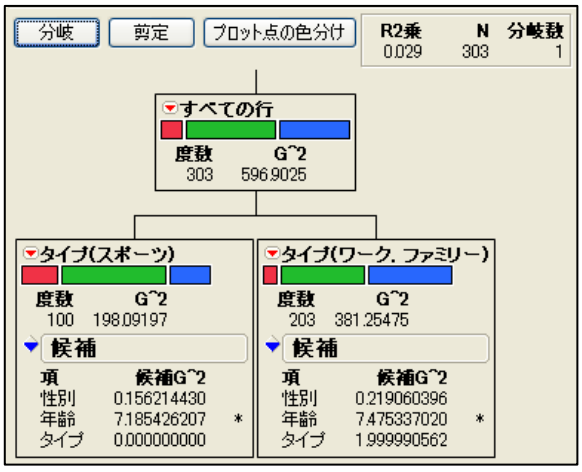


図 6

同じ要領で、「タイプ(スポーツ)」に属するデータについての候補 G^2、「タイプ(ワーク、ファミリー)」に属するデータについての候補 G^2 が表示されます。「タイプ(スポーツ)」は年齢の候補 G^2 (=7.1854)、「タイプ(ワーク、ファミリー)」は年齢の候補の G^2(=7.4533)です。候補 G^2 の値を比較すると、「タイプ(ワーク、ファミリー)」の値の方が大きいので、次は「タイプ(ワーク、ファミリー)」の**「年齢」**で分岐します。(図 7)

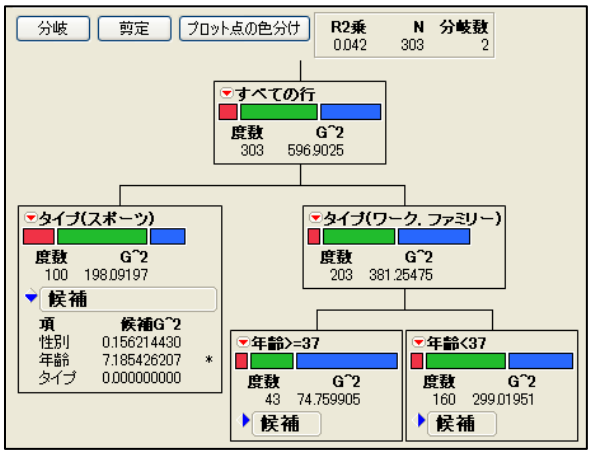


図 7

分岐基準として、**[有意度を最大化]**を選択した場合は、「候補」の欄に「対数価値」が表示されます。この値は、尤度比検定の p 値を調整した調整済みの p 値に対し、3 章で紹介した式(1)のように、負の対数をとったものになります。(図 8)

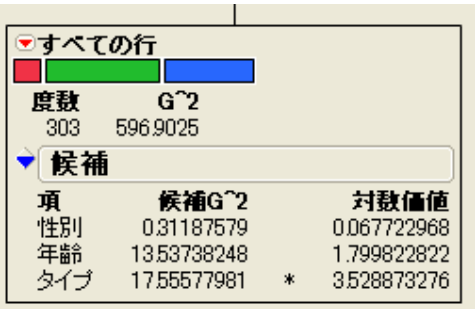


図 8

注意:最適な分岐点にはアスタリスク(*)がつきますが、候補 G^2(連続の場合は、候補 SS)と対数価値の最適な分岐点が異なる場合は、「<」(候補 G^2 または候補 SS)が最大の項、「>」(対数価値が最大の項)というように、別々に表示されます。アスタリスクは、候補 G^2(または候補 SS)が最大の項と対数価値が最大の項が一致するときに表示されます。

■応答が連続変数のとき

●使用する JMP のサンプルデータ :**「ボストンの住宅.jmp」**

●パーティションでの列の指定:

[Y, 目的変数]:**「持ち家の価格」**

[X, 説明変数]:**「犯罪率」、「区画」、・・・、「低所得者」**

基準は、**「分岐統計量を最大化」**を選択します。

「すべての行」に対する「候補」を表示させた最初のパーティションのレポートは、図 9 のようになります。

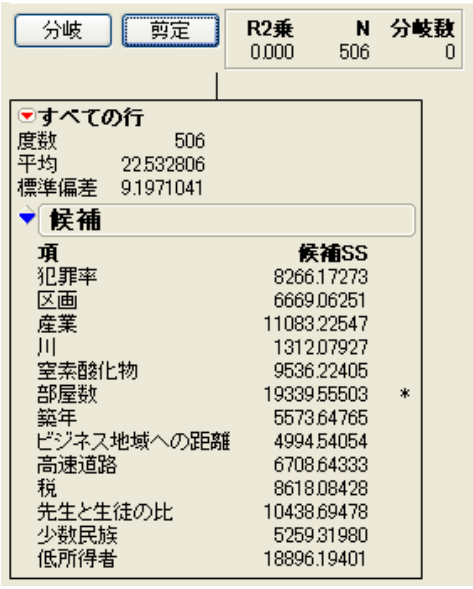
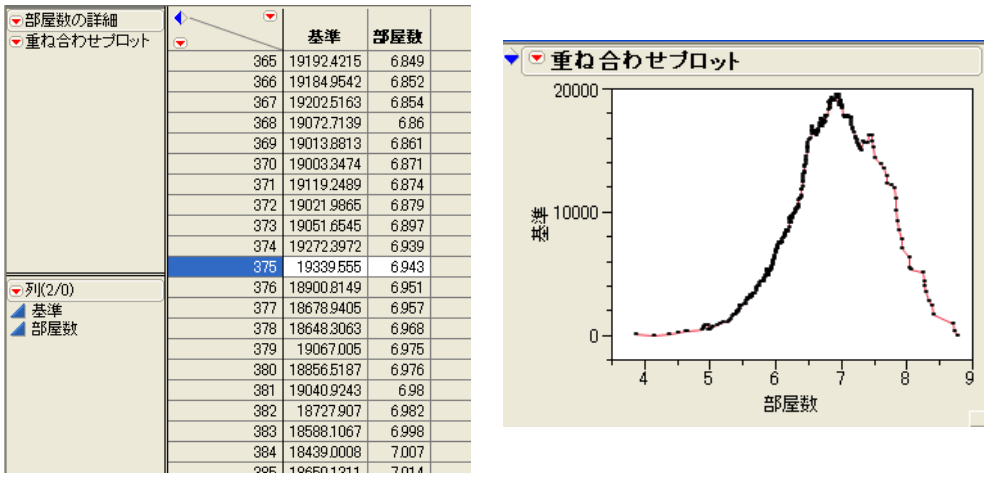


図 9

「候補」の欄を参照しますと、「部屋数」にアスタリスク(*)がついていることがわかります。そのため、ここでの最適な分岐は、「部屋数」になり、候補 SS は、19339.55 です。候補 SS は、説明変数のとりうる範囲を、その中でデータがとりうる値を境にして 2 つにグループ分けしたとき、各グループの平均をあてはめたときの平方和のうち最大のものを示します。

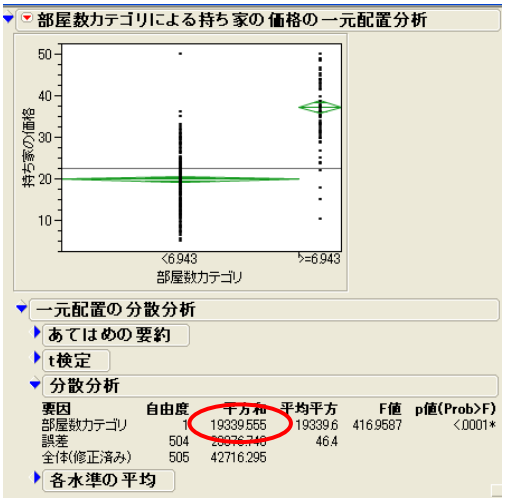
この例で**「部屋数」**は、連続尺度で 3.561 から 8.78 までの値をとります。この範囲(3.561,8.78)をデータがとりうる値を境に 2 つにグループ分けします。仮に 2 つのグループ分けを識別する新しい列(名義尺度)をつくったとします。このとき**[二変量の関係]**で、**「持ち家の価格」**を[Y,目的変数]、2 つにグループ分けした列を[X,説明変数]にして一元配置分散分析を行い、要因 X の平方和を参照します。それぞれの境界値に対して、この平方和が計算できますが、その中で一番大きい平方和が候補 SS(=19399.55)になります。この内容を確かめるには、「すべての行」の赤い三角ボタンをクリックし、**[詳細の表示]**を選択します。このとき、「どの項の詳細?」というタイトルのウィンドウが表示されますので、列**「部屋数」**を選択して**[OK]**ボタンをクリックすると、新しいデータテーブルが出力されます。このテーブルは、列**「部屋数」**が境界値を示し、この境界を基準にして 2 つのグループに分けたときの平方和が「基準」の列に表示されます。ここで、基準の値が最も大きい行は、375 行目の基準=19339.55 のときで、このときの部屋数は、6.943 となります。また、データテーブル左上のスクリプト「重ね合わせプロット」を実行しますと、データをグラフ表示することができ、部屋数の値に対する基準の値を視覚的に確認することができます。(図 10)



	基準	部屋数
365	19192.4215	6.849
366	19184.9542	6.852
367	19202.5163	6.854
368	19072.7139	6.86
369	19013.8813	6.861
370	19003.3474	6.871
371	19119.2489	6.874
372	19021.9865	6.879
373	19051.6545	6.897
374	19272.3972	6.939
375	19339.555	6.943
376	18900.8149	6.951
377	18678.9405	6.957
378	18648.3063	6.968
379	19067.005	6.975
380	18856.5187	6.976
381	19040.9243	6.98
382	18727.907	6.982
383	18588.1067	6.998
384	18439.0008	7.007

図 10

図 11 は、「部屋数」を 6.943 未満/6.943 以上の 2 つのカテゴリに分け(列名:「部屋数カテゴリ」)、上記のとおり、一元配置分散分析を行った結果になります。



要因	自由度	平方和	平均平方	F値	p値(Prob>F)
部屋数カテゴリ	1	19339.555	19339.6	416.9587	<.0001*
誤差	504	23376.740	46.4		
全体(修正済み)	505	42716.295			

図 11

「部屋数カテゴリ」の平方和は、19339.55 と表示され、これは、候補 SS に表示される値と一致します。図 12 は、1 回分岐したときの図で、確かに、”部屋数<6.943”と”部屋数>=6.943” で分岐しています。

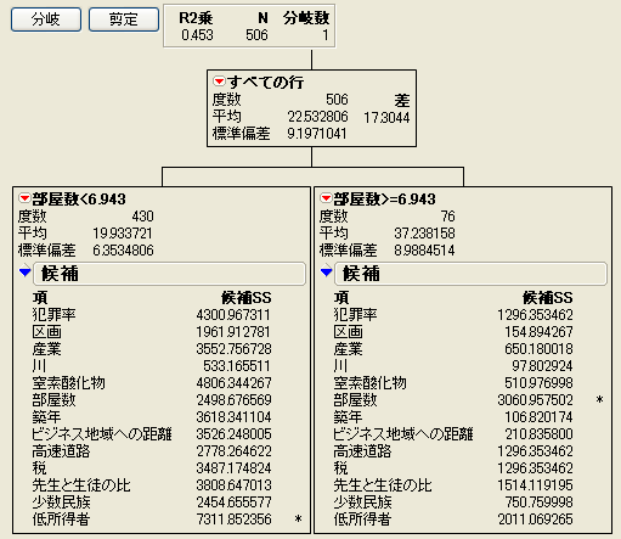


図 12

次は、応答がカテゴリのときと同様に、葉の中で候補 SS が最大になる値がアスタリスク表示されますので、すべての葉の中で、アスタリスクの行に表示される候補 SS が最大になる箇所で分岐します。図 11 より、「部屋数<6.943」の葉での候補 SS の最大値は、「低所得者」の 7311.85 で、一方、「部屋数>=6.943」の葉での候補 SS の最大値は、「部屋数」の 3060.95 です。これより、次は「部屋数<6.943」の変数「低所得者」で分岐します。(図 13)

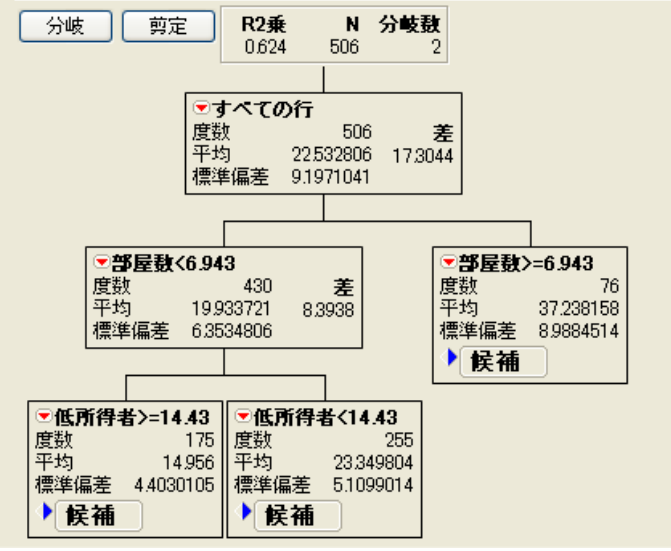


図 13

分岐基準として、**[有意度を最大化]**を選択した場合は、「候補」の欄に「対数価値」が表示されます。この値は、分散分析における F 検定の p 値を調整した調整済みの p 値に対し、3 章で紹介した式(1)のように、負の対数をとったものになります。