IBM SPSS Decision Trees 21

注 ： この情報とサポートされている製品をご使用になる前に、「注意事項」(p.120) の一般情報をお読みください。

本版は IBM® SPSS® Statistics 21 ,および新版で指示されるまで後続するすべてのリリースおよび変更に対して適用されます。

Adobe 製品のスクリーンショットは Adobe Systems Incorporated の許可を得て転載しています。

Microsoft 製品のスクリーンショットは Microsoft 社の許可を得て転載しています。

Licensed Materials - Property of IBM

© Copyright IBM Corporation 1989, 2012.

U.S. Government Users Restricted Rights - Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

はじめに

IBM® SPSS® Statistics は、データ分析の包括的システムです。ディシジョン ツリー は、このマニュアルで説明されている追加の分析手法を提供するオプションのアドオン モジュールです。ディシジョン ツリー アドオン モジュールは SPSS Statistics Core システムと組み合わせて使用し、Core システムに 完全に統合されます。

IBM Business Analytics について

IBM Business Analytics ソフトウェアは、意思決定者がビジネス パフォーマンスを向上させるために信頼する完全で、一貫した正確な情報を提供します。ビジネス インテリジェンス、予測分析、財務実績および戦略管理、および 分析アプリケーションの包括的なポートフォリオを利用することによって、現在の実績を明確、迅速に理解し、将来の結果を予測することができます。豊富な業界のソリューション、実績ある実例、専門サービスと組み合わせ、さまざまな規模の組織が、高い生産性を実現、意思決定を自信を持って自動化し、より良い決定をもたらします。

このポートフォリオの一部として、IBM SPSS Predictive Analytics ソフトウェアを使用する組織は、将来のイベントを予測し、その洞察に基づいて積極的に行動し、より優れた業績を実現することができます。全世界の企業、政府、学術分野のお客様が IBM SPSS の技術を活用し、不正行為を減少させ、リスクを軽減させながら、顧客の獲得、保持、成長において、競争優位を高めることができます。IBM SPSS ソフトウェアを日々の業務に取り入れることによって、組織は業務目標を達成し、大きな競争的優位を獲得することができるよう、意思決定を方向付け、自動化することができるようになります。お問い合わせは、http://www.ibm.com/spss を参照してください。

テクニカル サポート

テクニカル サポートのサービスをご利用いただけます。IBM Corp.製品の使用方法や、対応しているハードウェア環境へのインストールに関して問い合わせることもできます。テクニカル サポートの詳細については、IBM Corp. Web サイト（http://www.ibm.com/support）を参照してください。連絡の際は、所属団体名、サポート契約などを確認できるよう、あらかじめ手元にご用意ください。

© Copyright IBM Corporation 1989, 2012.

学生向けテクニカル サポート

IBM SPSS ソフトウェア製品の Student 版、アカデミック版、Grad パック版を使用している学生の場合、学生用の特別オンライン ページ、Solutions for Education (http://www.ibm.com/spss/rd/students/) ページを参照してください。大学提供の IBM SPSS ソフトウェアのコピーを使用している場合、大学の IBM SPSS 製品コーディネータにお問い合わせください。

カスタマ サービス

配送やアカウントに関するご質問は、お近くの営業所にお問い合わせください。お問い合せの際には、シリアル番号をご用意ください。

トレーニング セミナー

IBM Corp. では一般公開およびオンサイトで トレーニング セミナーを実施しています。セミナーでは実践的な講習を行います。セミナーは主要都市で定期的に開催されます。セミナーに関する詳細については、http://www.ibm.com/software/analytics/spss/training を参照してください。

内容

パート I: ユーザー ガイド

パート II: 例

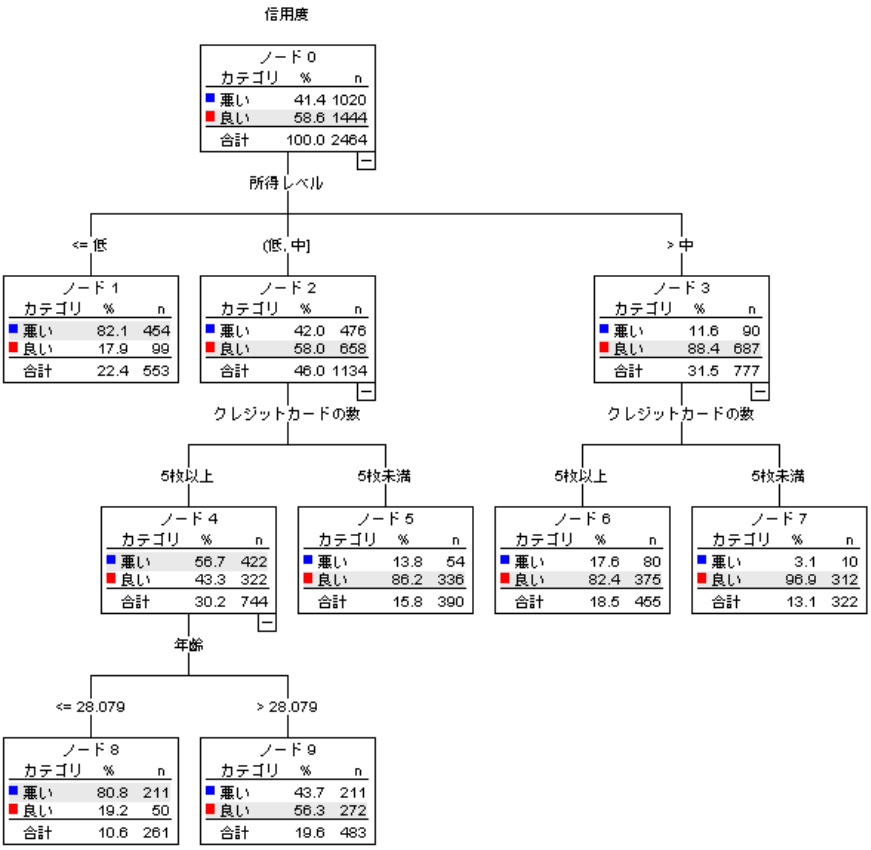
パート I:
ユーザー ガイド

章

1

ディシジョン ツリーの生成

図 1-1
ディシジョン ツリー



ディシジョン ツリー手続きで、ツリー ベースの分類モデルを作成します。ケースをグループに分類したり、独立（予測）変数の値を基に従属（目的）変数を予測するためのモデルです。この手続きには、分類を探索的、および確証的に分析するための検証ツールが用意されています。

この手続きの用途は以下のとおりです。

セグメンテーション。特定のグループに所属すると考えられる人物を特定します。

層化。リスクの高いグループ、中程度のグループ、低いグループなど、複数のカテゴリのうちの 1 つにケースを割り当てます。

© Copyright IBM Corporation 1989, 2012.

予測。ある人が債務不履行になる確率、車や家屋を転売する場合の潜在価値など、将来の出来事について、規則を作成して予測します。

データの分解と変数の選別。大量の変数から成るセットから、形式的パラメトリック モデルを構築する際に使用する有用な予測変数のサブセットを選択します。

交互作用の識別。特定のサブグループ間にのみ関連する関係を識別し、形式的パラメトリック モデルの中で指定します。

カテゴリの結合と連続変数の離散化。情報の損失を最小限にしつつ、グループの予測変数カテゴリと連続変数を再割り当てします。

例。ある銀行が、クレジットの申込者を信用リスクが適正かどうかによって分類するとします。過去の顧客の信用格付けの実績をはじめとするさまざまな要因を基に、新規顧客が債務不履行になるかどうかを予測するモデルを構築できます。

ツリー ベースの分析には、優れた機能が用意されています。

- 高リスク、または低リスクの等質なグループを識別できます。
- 個々のケースについて予測を立てるための規則を簡単に構築できます。

データの考慮事項

データ。 従属変数および独立変数は次のものを使用できます。

- **名義データ.** 値がランキングなどを持たないカテゴリを表しているとき､名義（変数）として取り扱うことができます。たとえば、従業員の会社の所属などです。名義変数の例としては、地域やジップ コードや所属宗教などがあります。
- **順序データ.** 値がランキングをもったカテゴリを表しているとき、変数を順序として取り扱うことができます。たとえば、「かなり不満」から「かなり満足」までのようなサービス満足度のレベルなどです。順序変数の例としては、満足度や信頼度を表す得点や嗜好得点などです。
- **スケール データ.** 値が有意な基準を持った順序カテゴリを表しているとき、変数をスケール（連続型）として扱うことができます。値間の距離の比較などに適切です。スケール変数の例としては、年齢や、千ドル単位で表した所得があります。

度数による重み付け。重み付けが有効な場合に、小数値の重みを最も近い整数に丸めます。これにより、重みの値が 0.5 未満のケースは、重み 0 として取り扱われることになり、分析から除外されます。

仮定。 この手続きでは、すべての分析変数に対して適切な尺度が割り当てられていることを前提にしています。一方、分析に含まれている従属変数のすべての値に値ラベルが定義されていることを前提にしている機能もあります。

- **測定レベル。** 尺度はツリーの計算に影響を及ぼすので、すべての変数を適切な尺度に割り当てる必要があります。デフォルトでは数値型変数がスケール変数、文字型変数が名義変数という前提なので、実際の尺度が正確に反映されない場合もあります。変数リストで各変数の隣にあるアイコンは、変数の型を表します。


スケール


名義


順序

ソース変数リストの変数を右クリックしてコンテキスト メニューから尺度を選択することで、その変数の尺度を一時的に変更できます。

- **値ラベル。** この手続きのダイアログ ボックス インターフェイスでは、カテゴリ（名義、順序）従属変数のすべての非欠損値で値ラベルが定義されているか、どの非欠損値でも一切定義されていないということを前提にしています。カテゴリ従属変数の非欠損値のうち、少なくとも 2 つで値ラベルを定義しないと、一部の機能を使用できません。少なくとも 2 つの非欠損値で値ラベルが定義されている場合、値ラベルのない値を含むケースはすべて分析から除外されます。

ディシジョン ツリーを行うには

▶ メニューから次の項目を選択します。
分析(A) > 分類 > ツリー...

図 1-2
[ディシジョン ツリー] ダイアログ ボックス

- ▶ [従属変数] ボックスに従属変数を選択します。
- ▶ 1 つ以上の独立変数を選択します。
- ▶ 成長手法を選択します。

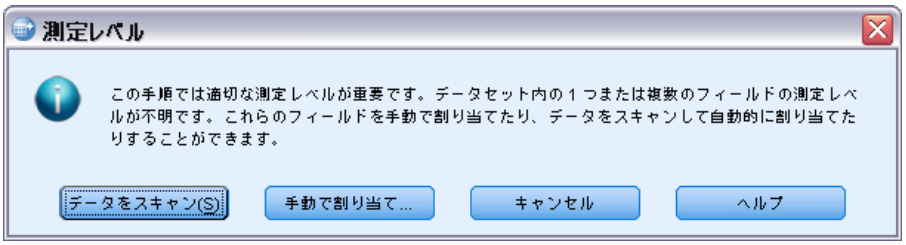
  オプションとして、次の選択が可能です。

  - ■ ソース リストの変数の尺度を変更できます。
  - ■ 独立変数リストの最初の変数を最初の分割変数として強制的にモデルに格納できます。
  - ■ あるケースがツリーの成長過程に及ぼす影響を示す影響度変数を選択できます。影響度の低いケースは影響が少なく、影響度が高くなるほど影響も強まります。影響度変数の値は、正の整数である必要があります。
  - ■ ツリーを検証できます。
  - ■ ツリーが成長する基準をカスタマイズできます。
  - ■ ターミナル ノード番号、予測値、および予測確率を変数として保存できます。
  - ■ モデルを XML (PMML) 形式で保存できます。

測定レベルが不明なフィールドです。

データセットの 1 つまたは複数の変数（フィールド）の尺度が不明な場合、尺度の警告が表示されます。尺度はこの手順の結果の計算に影響を与えるため、すべての変数に尺度を定義する必要があります。

図 1-3
尺度の警告



- **データをスキャン。** アクティブ データセットのデータを読み込み、デフォルトの尺度を尺度が現在不明なフィールドに割り当てます。データセットが大きい場合は時間がかかります。
- **手動で割り当てる。** 不明な尺度のフィールドをすべて表示するダイアログが開きます。このダイアログを使用して、尺度をこれらのフィールドに割り当てることができます。データ エディタの［変数ビュー］でも、尺度を割り当てることができます。

尺度がこの手順で重要であるため、すべてのフィールドに尺度が定義されるまで、ダイアログにアクセスしてこの手順を実行することはできません。

尺度の変更

▶ ソース リスト内の変数を右クリックします。

▶ ポップアップ コンテキスト メニューから尺度を選択します。

この操作で、ディシジョン ツリー手続きで使用される尺度が一時的に変更されます。

成長手法

利用できる成長手法は次のとおりです。

CHAID. カイ 2 乗自動反復検出。各ステップにおいて、CHAID は、従属変数と最も強い交互作用を持つ独立（予測）変数を選択します。従属変数の値に関する各予測変数のカテゴリーが著しく異ならない場合はそのカテゴリーは統合されます。

Exhaustive CHAID. 各予測変数に対して可能な限りの分割を調べる CHAID の改良版です。

CRT. 分類ツリーと回帰ツリー。CRT は、従属変数に関するデータはできるだけ等質のセグメントへ分割します。すべてのケースの従属変数が同じ値であるターミナル ノードは、等質であり、「純粋」ノードと言えます。

QUEST. 簡単、不偏、効率的な統計ツリーです。速く、多くのカテゴリーを持つ予測変数を肯定して他の方式の偏りを回避する方式。従属変数が名義の場合のみ、QUEST を指定することができます。

それぞれの成長手法には、次のような利点と制限があります。

	CHAID*	CRT	QUEST
カイ 2 乗を基にする**	X		
独立（予測）変数の代理変数		X	X
ツリーの剪定		X	X
ノードの多重分岐	X		
ノードの 2 分岐		X	X
影響度変数	X	X	
事前確率		X	X
誤分類コスト	X	X	X
高速計算	X		X

*Exhaustive CHAID を含みます。

**QUEST でも名義独立変数に対してカイ 2 乗の測度が使用されます。

カテゴリの選択

図 1-4
[カテゴリ] ダイアログ ボックス

カテゴリ（名義、順序）従属変数には、以下の操作を実行できます。

- 分析に含めるカテゴリの指定。
- 関心のある目標カテゴリの識別。

カテゴリの包含または除外

分析を従属変数の特定のカテゴリに制限できます。

- ［除外］リストの従属変数を値に持つケースは、分析に含まれません。
- 名義従属変数については、ユーザー欠損カテゴリを分析に含めることもできます。(デフォルトでは、ユーザー欠損カテゴリが［除外］リストに表示されます)。

目標カテゴリ

選択されている（オンになっている）カテゴリは、分析で最も注目しているカテゴリとして取り扱われます。たとえば、債務不履行になる可能性がある個人の識別に最も注目している場合、信用格付けが「悪い」というカテゴリを目標カテゴリに選択できます。

- デフォルトの目標カテゴリはありません。カテゴリが選択されていない場合、分類規則のオプションの一部とゲインに関連する出力を利用できません。
- 複数のカテゴリが選択されている場合、それぞれの目標カテゴリについて個別のゲイン テーブル、およびグラフが作成されます。
- 目標カテゴリに 1 つ以上のカテゴリを指定しても、ツリー モデル、リスク推定、誤分類の結果には影響しません。

カテゴリと値ラベル

このダイアログ ボックスには、従属変数に対する定義済みの値ラベルが必要です。カテゴリ従属変数の値のうち少なくとも 2 個の値が定義済みの値ラベルを持っていないと利用できません。

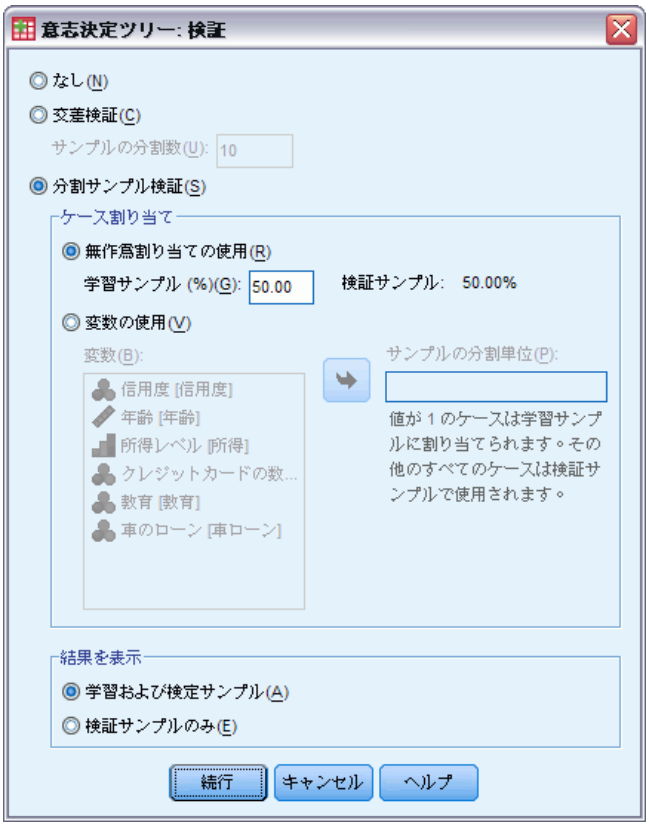
カテゴリを包含または除外し、目標カテゴリを選択するには

▶ ［ディシジョン ツリー］メイン ダイアログ ボックスで、2 つ以上の値ラベルが定義されているカテゴリ（名義、順序）従属変数を選択します。

▶ ［カテゴリ］をクリックします。

検証(V)

図 1-5
[検証] ダイアログ ボックス



ツリー構造が大きな母集団に対しても利用できる一般性をどの程度持っているか評価するために、検証を使用できます。検証は、交差検証と分割サンプル検証の 2 種類を使用できます。交差検証と分割サンプル検証の 2 つの検証方法を使用できます。

交差検証

交差検証では、サンプルを**群**と呼ばれる複数のサブサンプルに分割します。分割の後、ツリー モデルが生成されますが、各サブサンプルのデータは除外されます。つまり、最初のツリーは最初のサブサンプル以外のすべてのケースを基に生成され、2 番目のツリーは 2 番目のサブサンプル以外のすべてのケースを基に生成されます。それぞれのツリーを、そのツリーの生成時に除外したサブサンプルに適用し、誤分類のリスクを推定します。

- サブサンプル数は最高 25 回まで指定できます。サブサンプル数の値を大きくするほど、それぞれのツリー モデルから除外されるケースの数は減少します。
- 交差検証では、最終的なツリー モデルが 1 つ作成されます。交差検証の最終的なツリーによるリスク推定は、すべてのツリーのリスクの平均値を算出します。

分割サンプル検証

分割サンプル検証を使用する場合、学習サンプルでモデルを生成し、提供用サンプルで検証します。

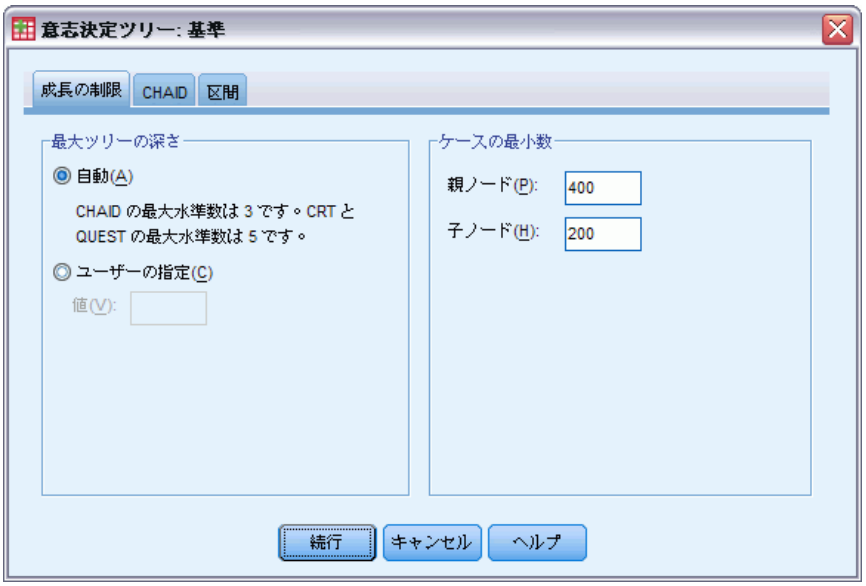
- 学習サンプルのサイズは、合計サンプル サイズに対する割合、またはサンプルを学習用と検証用に分割する変数で指定できます。
- 変数を使用して学習用と検証用のサンプルを定義する場合、変数の値が 1 のケースは学習サンプルになり、それ以外のケースは検証サンプルになります。その際、従属変数、重み付け変数、影響度変数、強制独立変数は使用できません。
- 学習用と検証用のサンプルの結果をともに表示すること、または検証サンプルのみの結果を表示することができます。
- 分割サンプル検証を小さいデータ ファイル（ケース数が少ないデータ ファイル）に対して使用する場合は注意が必要です。学習サンプルのサイズが小さいと、カテゴリによってはツリーを十分に成長させるだけのケースが不足する場合もあり、作成されるモデルが十分でない可能性があります。

ツリーの成長基準

成長手法、従属変数の尺度、またはその 2 つの組み合せによって、使用できる成長基準は異なります。

成長の制限

図 1-6
[基準] ダイアログ ボックスの [成長の制限] タブ



[成長の制限] タブでは、ツリーのレベル数を制限し、親子ノードのケースの最小値を制限できます。

ツリーの最大の深さ。 ルート ノードの下に最大で何水準まで成長するかを制御します。[自動] 設定により、レベルは CHAID 手法と Exhaustive CHAID 手法の場合ルート ノードの下 3 つ、CRT 手法と QUEST 手法の場合は 5 つまでに制限されます。

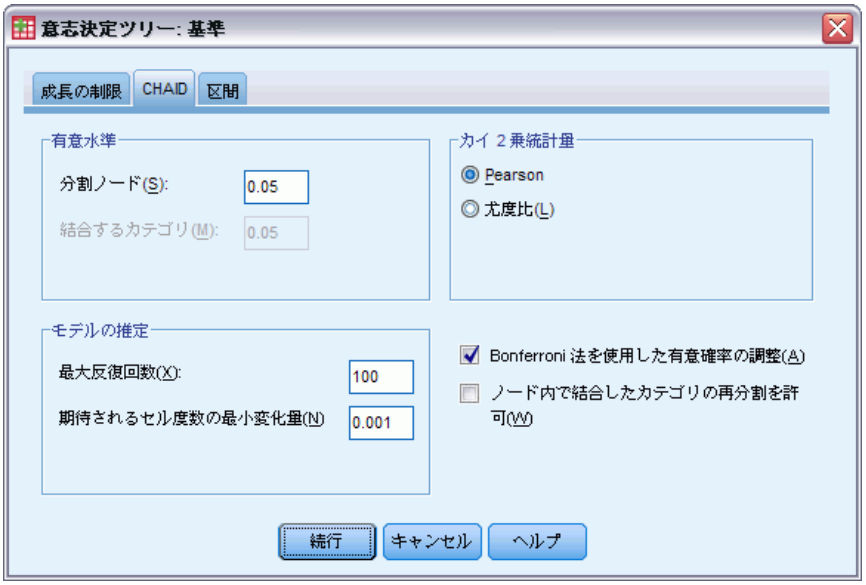
ケースの最小数。 ノードのケースの最小数を制御します。この基準を満たさないノードは分割されません。

- 最小値を高くすると、ノードの少ないツリーができます。
- 最小値を低くすると、ノードの多いツリーができます。

ケース数が少ないデータ ファイルの場合、親ノードでケース 100 個、子ノードで 50 個というデフォルト値の作用で、ルート ノードの下にノードを持たないツリーができあがることがあります。このような場合は、最小値を低くすると、よりよい結果が得られます。

CHAID 基準

図 1-7
[基準] ダイアログ ボックス -> [CHAID] タブ



CHAID 手法および Exhaustive CHAID 手法の場合、以下の項目を制御できます。

有意水準。 ノードを分割したりカテゴリを結合したりするための有意確率を制御できます。いずれの基準でも、デフォルトの有意水準は 0.05 です。

- ノードの分割の場合、0 より大きく 1 未満の範囲で値を指定する必要があります。値を下げるとノードの少ないツリーができます。
- カテゴリの結合の場合、0 より大きく 1 以下の範囲で値を指定する必要があります。カテゴリを結合しないようにするには、値を 1 に指定します。スケール独立変数の場合、最終的なツリーの変数のカテゴリは、指定した区間数（デフォルトでは 10 区間）で繰り返されます。詳細は、 p. 12 CHAID 分析のスケールの区間 を参照してください。

カイ 2 乗統計量。 順序従属変数では、ノード分割およびカテゴリ結合を判断するカイ 2 乗を尤度比法で計算します。名義従属変数では、次の方法を選択できます。

- **Pearsonの相関係数。** 計算は速くなりますが、小さいサンプルの場合には注意して使用する必要があります。これがデフォルトの方法となります。
- **尤度比。** Pearson の相関係数ほど注意して使用する必要がない一方、計算には時間がかかります。小さいサンプルに適した方法です。

モデルの推定。 名義従属変数または順序従属変数の場合に次のオプションを指定できます。

- **最大反復回数。** デフォルトは 100 回です。ツリーが最大反復回数に達して成長が止まる場合、この回数を増やすか、またはツリーの成長を制御する他の基準を変更できます。
- **期待されるセル度数の最小変化量。** 0 より大きく 1 未満の範囲で値を指定する必要があります。デフォルトは 0.05 です。値を下げるとノードの少ないツリーができます。

Bonferroni 法を使用した有意確率の調整。 多重比較のときに、結合基準、および分割基準の有意確率を Bonferroni 法を使用して調整します。これはデフォルトです。

ノード内で結合したカテゴリの再分割を許可。 カテゴリの結合を明示的に回避しない限り、モデルを記述するツリーが最も単純化されるように独立（予測）変数のカテゴリが結合されます。このオプションを選択すると、結合されたカテゴリを再分割することでより適切な解を得られる場合は、再分割が実行されます。

CHAID 分析のスケールの区間

図 1-8
[基準] ダイアログ ボックス -> [区間] タブ

CHAID 分析では、分析の前に常にスケール独立（予測）変数が個別のグループにまとめられます（たとえば、0 ～ 10、11 ～ 20、21 ～ 30 など）。最初の分割の後に連続するグループが結合されることがありますが、グループの初期値と最大値は次のオプションで制御できます。

- **固定数。** すべてのスケール独立変数は、最初に同数ずつのグループにまとめられます。デフォルトは 10 です。
- **ユーザー指定。** 各スケール独立変数は、最初にその変数に対して指定した数のグループにまとめられます。

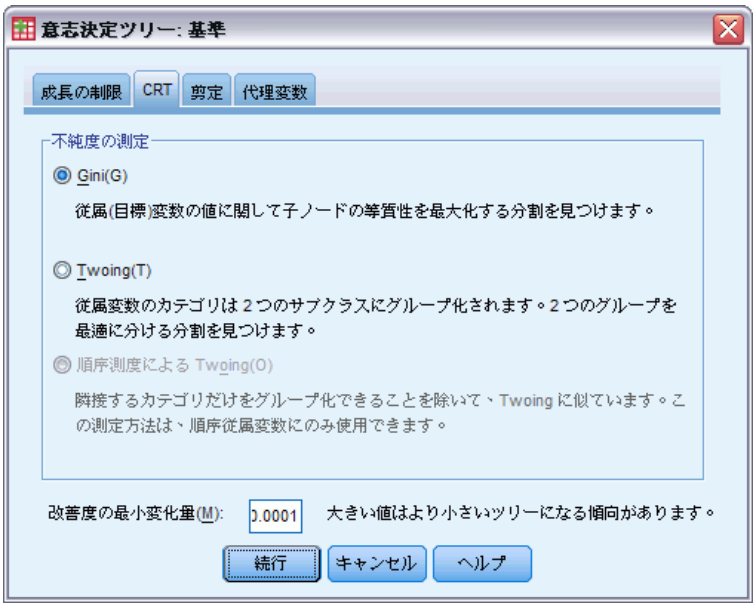
スケール独立変数の区間を指定するには

1. [ディシジョン ツリー] メイン ダイアログ ボックスで、スケール独立変数を 1 つ以上選択します。
2. 成長手法に [CHAID] または [Exhaustive CHAID] を選択します。
3. [基準] をクリックします。
4. [区間] タブをクリックします。

CRT 分析と QUEST 分析では、分割は必ず 2 分割で、スケール独立変数と順序独立変数が同様に扱われます。したがって、スケール独立変数に多数の区間を指定することはできません。

CRT の基準

図 1-9
[基準] ダイアログ ボックス -> [CRT] タブ



CRT 成長手法は、ノード内の等質性を最大限にします。あるノードでケースの等質なサブグループが現れない度合いを**不純度**で表します。たとえば、すべてのケースの従属変数が同じ値であるターミナル ノードは「純粋」なので、これ以上分割の必要がない等質なノードといえます。

不純度の測定に使用する方法、およびノードの分割に必要な不純度の最小減少量を選択できます。

不純度の測定。 スケール従属変数には、最小 2 乗偏差（LSD）という不純度の測定法が使用されます。この値は、度数による重み付けまたは影響度に合せて調整された、ノード内分散として計算されます。

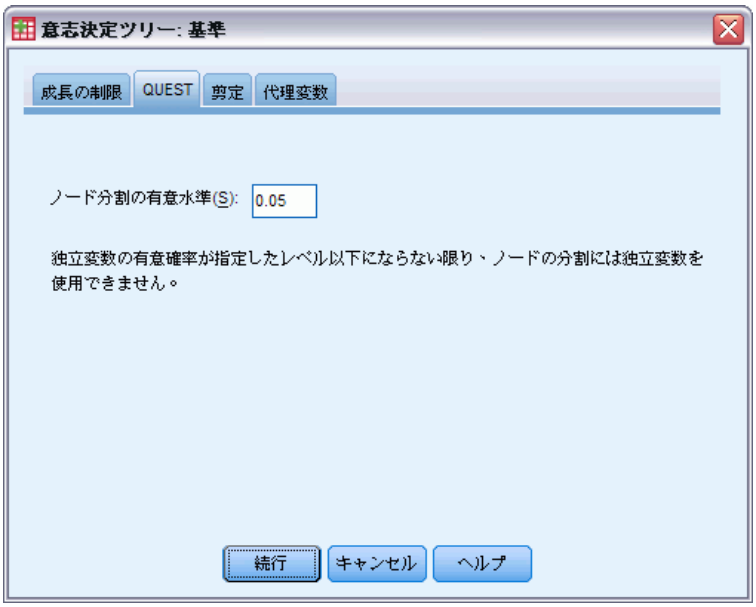
カテゴリ（名義、順序）従属変数には、次の不純度の測定法を選択できます。

- **Gini。** 子ノードの、従属変数の値に関する等質性が最大になる分割が見つかります。Gini は従属変数の各カテゴリに属するメンバーの確率の 2 乗を基に算出されます。ノード内のすべてのケースが 1 つのカテゴリにまとまったときが最小値（0）です。これがデフォルトの測定法となります。
- **Twoing。** 従属変数のカテゴリは 2 つのサブクラスにグループ化されます。2 つのグループを最適に分ける分割が見つかります。
- **順序測度による Twoing。** Twoing と同様ですが、隣接するカテゴリしかグループ化されない点が異なります。この測定法は、順序従属変数にだけ使用できます。

改善度の最小変化量。 ノードの分割に必要な不純度の最小の減少量です。デフォルトは 0.0001 です。値を上げると、ノードの少ないツリーができます。

QUEST 基準

図 1-10
[基準] ダイアログ ボックス -> [QUEST] タブ



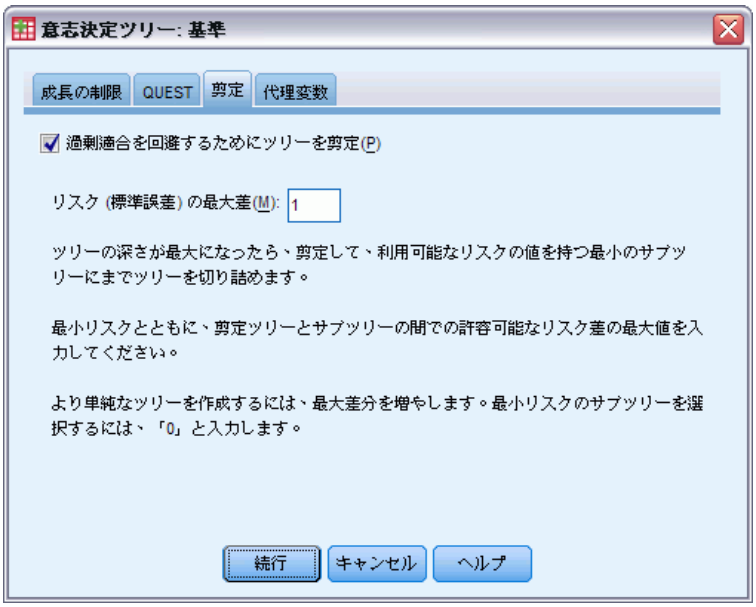
QUEST 手法では、ノードを分割するための有意水準を指定できます。有意水準が指定した値を上回る場合、独立変数でノードを分割することはできません。値は 0 より大きく 1 未満にする必要があります。デフォルトは 0.05 です。値が小さいと、最終的なモデルから除外される独立変数が多くなります。

QUEST 基準を指定するには

- ▶ [ディシジョン ツリー] メイン ダイアログ ボックスで、名義独立変数を選択します。
- ▶ 成長手法の場合は、[QUEST] を選択します。
- ▶ [基準] をクリックします。
- ▶ [QUEST] タブをクリックします。

ツリーの剪定

図 1-11
[基準] ダイアログ ボックス -> [剪定] タブ



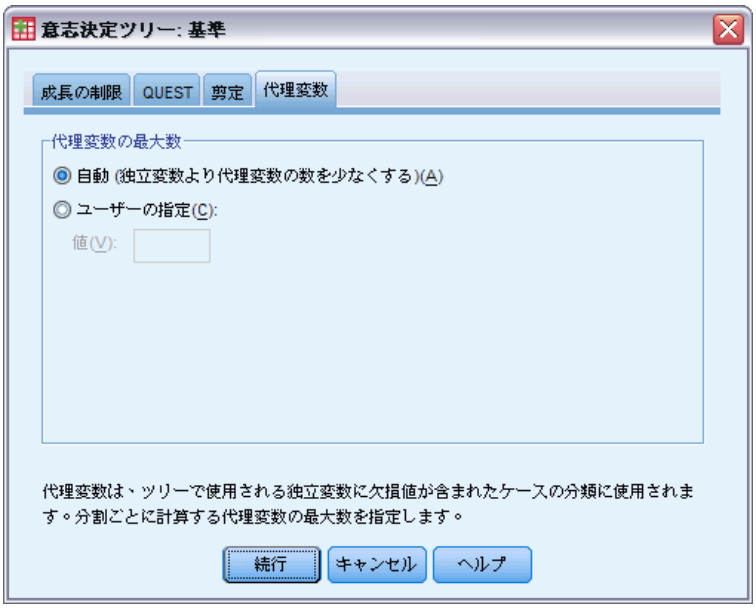
CRT 手法と QUEST 手法を使用し、ツリーを**剪定**してモデルの過剰適合を回避できます。ツリーは停止基準を満たすまで拡大し、リスクの指定した最大差分に基づいて、最も小さいサブツリーに自動的に剪定されます。リスクの値は標準誤差で表現されます。デフォルトは 1 です。負でない値を指定する必要があります。サブツリーのリスクを最小にするには、0 を指定します。

剪定とノードの非表示

剪定したツリーを作成すると、最終的なツリーではツリーから剪定されたノードを使用できません。最終的なツリーで子ノードを選択して、ユーザーの操作で非表示にしたり表示したりできますが、ツリーの作成過程で剪定したノードは表示できません。 詳細は、 2 章 p. 42 ツリーエディタ を参照してください。

代理変数

図 1-12
[基準] ダイアログ ボックスの [代理変数] タブ



CRT と QUEST では、独立（予測）変数に**代理変数**を使用できます。独立変数の値が欠損している場合、その変数と強く関連する別の独立変数が分類に使用されます。このような代替予測変数を代理変数と呼びます。モデルで使用する代理変数の個数の上限を指定できます。

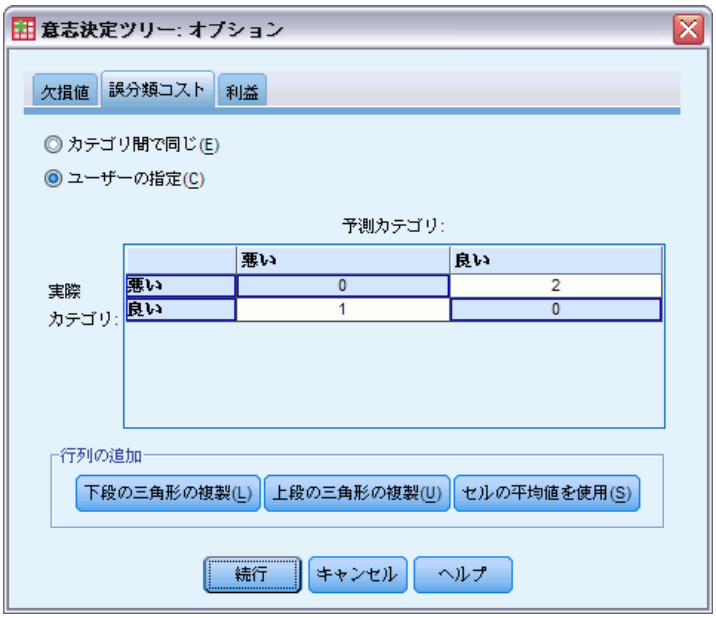
- デフォルトでは、独立変数の個数から 1 を引いた個数が代理変数の上限です。つまり、それぞれの独立変数は、他のすべての独立変数を代理変数として使用できます。
- モデルで代理変数を使用しない場合、代理変数の個数に 0 を指定します。

オプション

成長手法、従属変数の尺度、従属変数の値に対して定義された値ラベルの有無、この 3 つのいずれか、またはその組み合せによって、使用できるオプションは異なります。

誤分類コスト

図 1-13
[オプション] ダイアログ ボックスの [誤分類コスト] タブ



カテゴリ（名義、順序）従属変数では、誤分類コストを使用することで、不適切な分類に関する相対ペナルティについての情報を含めることができます。次に例を示します。

- 信用格付けの高い顧客を信用しないコストは、債務不履行になる顧客を過剰に信用するコストとは通常、異なります。
- 心臓病になるリスクが高い人物をリスクが低いと誤分類するコストは、低リスクの人物を高リスクと誤分類するコストよりもおそらく高いでしょう。
- 返信率が低い相手にダイレクト メールを送信するコストはきわめて低い一方、返信率が高い相手にメールを送信しないコストは（利益損失という点で）高くなります。

誤分類コストと値ラベル

カテゴリ従属変数の値のうち少なくとも 2 個の値が定義済みの値ラベルを持っていないと、このダイアログ ボックスは利用できません。

誤分類コストを指定するには

▶ [ディシジョン ツリー] メイン ダイアログ ボックスで、2 つ以上の値ラベルが定義されているカテゴリ（名義、順序）従属変数を選択します。

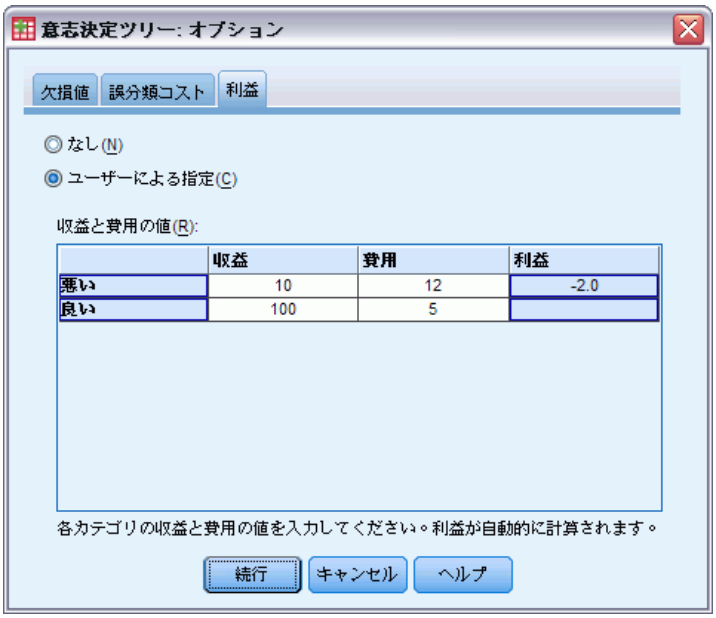
- ▶ [オプション] をクリックします。
- ▶ [誤分類コスト] タブをクリックします。
- ▶ [ユーザー指定] をクリックします。
- ▶ グリッドに誤分類コストを 1 つ以上入力します。負でない値を指定する必要があります。(正分類は対角線で示され、常に 0 です)。

行列の追加。 多くの場合、コストは対称にすることができます。対称とは、A を B と誤分類するコストと、B を A と誤分類するコストが同じである状態です。次のコントロールを使用すると、対称なコストの行列を容易に指定できます。

- **下段の三角形の複製。** 行列の下段の三角形（対角線の下）の値を、対応する上段の三角形のセルに複製します。
- **上段の三角形の複製。** 行列の上段の三角形（対角線の上）の値を、対応する下段の三角形のセルに複製します。
- **セルの平均値を使用。** 行列を半分にしたそれぞれの各セルに、2 つの値（上段の三角形と下段の三角形）の代わりに両方の値の平均値を格納します。たとえば、A を B と誤分類するコストが 1 で、B を A と誤分類するコストが 3 とすると、このコントロールを使用した場合は 2 つの値が平均値（1 + 3）/ 2 = 2 に置き換わります。

利益

図 1-14
[オプション] ダイアログ ボックス -> [利益] タブ



カテゴリ従属変数では、収益と費用の値を従属変数の水準に割り当てることができます。

- 利益は収益と費用の差として計算されます。
- 利益の値は、ゲイン テーブルの平均利益、および ROI（投資収益率）の値には影響しますが、。ツリー モデルの基本構造には影響しません。
- 収益と費用の値は数値にする必要があり、かつグリッドに表示される従属変数のすべてのカテゴリで指定する必要があります。

利益と値ラベル

このダイアログ ボックスには、従属変数に対する定義済みの値ラベルが必要です。カテゴリ従属変数の値のうち少なくとも 2 個の値が定義済みの値ラベルを持っていないと利用できません。

利益を指定するには

▶ [ディシジョン ツリー] メイン ダイアログ ボックスで、2 つ以上の値ラベルが定義されているカテゴリ（名義、順序）従属変数を選択します。

▶ [**オプション**] をクリックします。

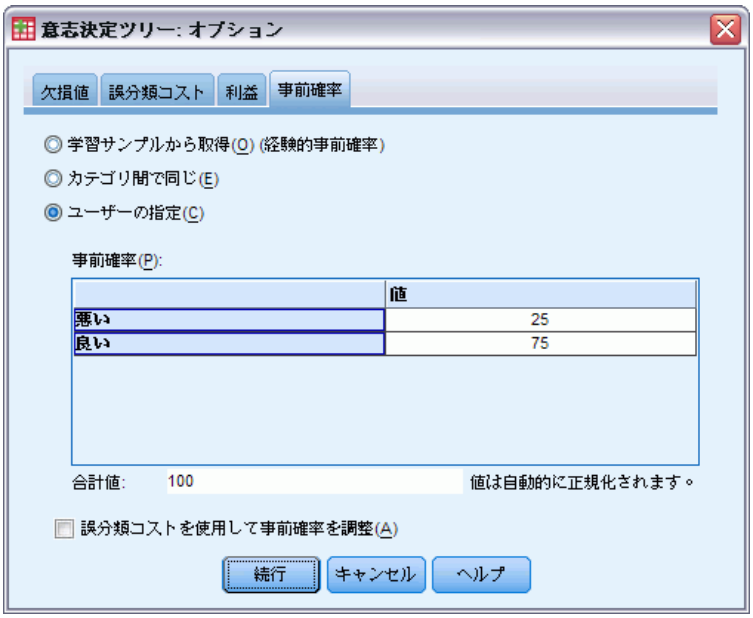
▶ [**利益**] タブをクリックします。

▶ [**ユーザー指定**] をクリックします。

▶ グリッドに表示されているすべての従属変数カテゴリに対し、収益と費用の値を入力します。

事前確率

図 1-15
[オプション] ダイアログ ボックス -> [事前確率] タブ



カテゴリ従属変数を含んだ CRT ツリーおよび QUEST ツリーに対して、所属グループの事前確率を指定できます。**事前確率**とは、独立（予測）変数の値について何かを知るより前に従属変数の各カテゴリの全体的な相対度数を推定したものです。事前確率を使用すると、母集団全体を代表しているわけではないサンプルのデータから起こったツリーの成長が補正されます。

学習サンプル（事前の経験）から取得。 データ ファイル内の従属変数の値の分布が母集団の分布を示している場合に使用します。分割サンプル検証を使用している場合は、学習サンプルのケースの分布が使用されます。

注 ：分割サンプル検証では、ケースがランダムに学習サンプルに割り当てられるので、サンプル内で実際にケースがどのように分布しているかを事前に知ることはできません。 詳細は、 p. 8 検証(V) を参照してください。

カテゴリ間で同じ。 従属変数のカテゴリが母集団の中で均等に表現されている場合に使用します。たとえば、4 つのカテゴリがある場合、それぞれのカテゴリにはケースのおよそ 25% が所属します。

ユーザー指定。 グリッドに表示されている従属変数のそれぞれのカテゴリに、負でない値を入力します。入力する値は、比率、パーセント、度数、またはそれ以外でカテゴリでの値の分布を表す値のいずれかです。

誤分類コストを使用して事前確率を調整。 誤分類コストをカスタムで定義した場合、コストを基に事前確率を調整できます。 詳細は、 p. 18 誤分類コスト を参照してください。

利益と値ラベル

このダイアログ ボックスには、従属変数に対する定義済みの値ラベルが必要です。カテゴリ従属変数の値のうち少なくとも 2 個の値が定義済みの値ラベルを持っていないと利用できません。

事前確率を指定するには

- ［ディシジョン ツリー］メイン ダイアログ ボックスで、2 つ以上の値ラベルが定義されているカテゴリ（名義、順序）従属変数を選択します。
- 成長手法に［CRT］または［QUEST］を選択します。
- ［オプション］をクリックします。
- ［事前確率］タブをクリックします。

得点

図 1-16
［オプション］ダイアログ ボックス -> ［得点］タブ

順序従属変数を含んだ CHAID および Exhaustive CHAID では、従属変数の各カテゴリにユーザー指定の得点を割り当てることができます。得点によって、従属変数のカテゴリの順序とカテゴリどうしの距離を定義します。得点を使用して、順序の値の相対距離を伸縮したり、順序を変更したりできます。

- **各カテゴリに一般的な順位を使用。** 従属変数の最も低いカテゴリに得点 1 を割り当て、次に低いカテゴリに得点 2 を割り当て、以下同様に続きます。これはデフォルトです。
- **ユーザー指定。** グリッドに表示されている従属変数のそれぞれのカテゴリに、得点を数値で入力します。

例

値ラベル	元の値	得点
非熟練者	1	1
熟練肉体労働者	2	4
事務員	3	4.5
Professional	4	7
管理職	5	6

- このように得点を定めると、非熟練者と 熟練肉体労働者の相対距離が伸び、熟練肉体労働者と 事務員の相対距離が縮みます。
- また、管理職と 専門職の順序が逆転します。

得点と値ラベル

このダイアログ ボックスには、従属変数に対する定義済みの値ラベルが必要です。カテゴリ従属変数の値のうち少なくとも 2 個の値が定義済みの値ラベルを持っていないと利用できません。

得点を指定するには

▶ [ディシジョン ツリー] メイン ダイアログ ボックスで、2 つ以上の値ラベルが定義されている順序従属変数を選択します。

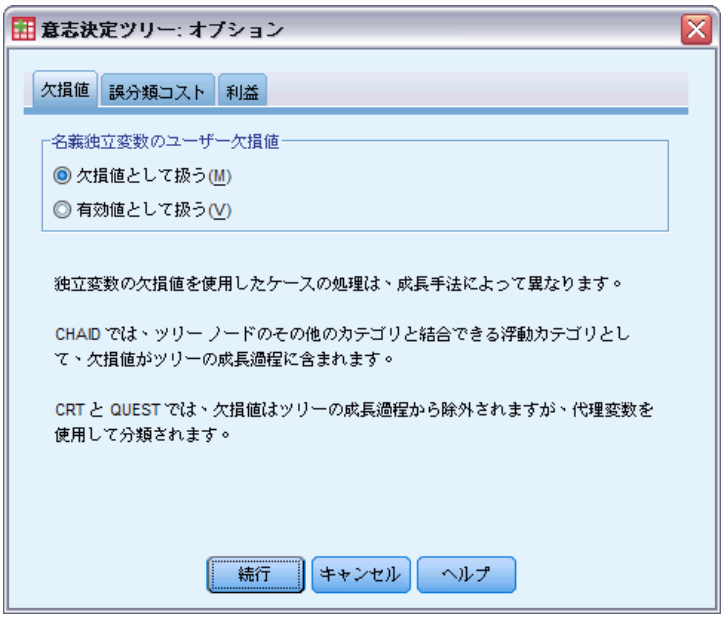
▶ 成長手法に [CHAID] または [Exhaustive CHAID] を選択します。

▶ [**オプション**] をクリックします。

▶ [**得点**] タブをクリックします。

欠損値

図 1-17
[オプション] ダイアログ ボックス -> [欠損値] タブ



[欠損値] タブでは、名義独立（予測）変数のユーザー欠損値の処理方法を制御します。

- 順序独立変数およびスケール独立変数のユーザー欠損値の処理方法は、成長手法によって異なります。
- 名義従属変数の処理方法は、[カテゴリ] ダイアログ ボックスで指定します。 詳細は、 p.6 カテゴリの選択 を参照してください。
- 順序従属変数およびスケール従属変数の場合、従属変数のシステム欠損値かユーザー欠損値があるケースは常に除外されます。

欠損値として扱う。 ユーザー欠損値とシステム欠損値を同様に扱います。システム欠損値の処理方法は成長手法によって異なります。

有効値として扱う。 ツリーの成長と分類で、名義独立変数のユーザー欠損値を通常の値として扱います。

方法依存規則

独立変数の値の一部（すべてではありません）がシステム欠損値、またはユーザー欠損値である場合、次のように処理されます。

- CHAID、および Exhaustive CHAID では、独立変数のシステム欠損値とユーザー欠損値が、1 つに結合したカテゴリとして分析に取り込まれます。スケール独立変数と順序独立変数では、まず有効な値でカテゴリが生成されてから、欠損するカテゴリを最も類似する（有効な）カテゴリと結合するのか、別のカテゴリとして保持するのかが決定されます。
- CRT、および QUEST では、独立変数の欠損値があるケースはツリーの成長過程から除外されます。ただし、代理変数が成長手法に含まれている場合は、代理変数で分類されます。名義ユーザー欠損値を欠損値として扱う場合も、同様に処理されます。 詳細は、 p. 17 代理変数 を参照してください。

名義独立変数のユーザー欠損値の処理方法を指定するには

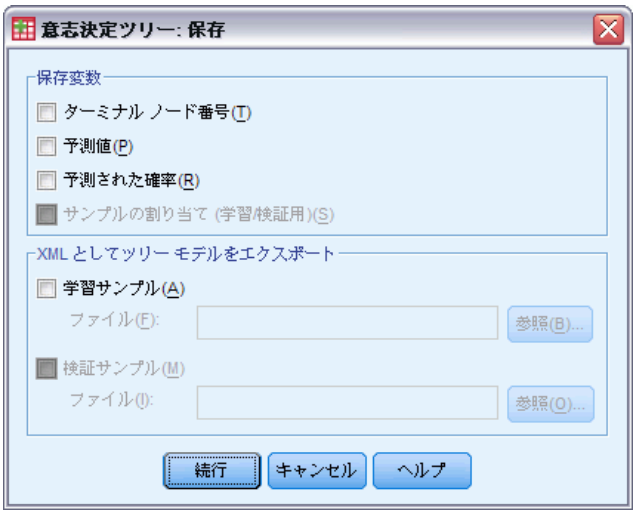
▶ [ディシジョン ツリー] メイン ダイアログ ボックスで、名義独立変数を 1 つ以上選択します。

▶ **[オプション]** をクリックします。

▶ **[欠損値]** タブをクリックします。

モデル情報の保存

図 1-18
[保存] ダイアログ ボックス



モデルの情報を変数として作業中のデータ ファイルに保存できます。また、モデル全体を XML（PMML）形式で外部ファイルに保存できます。

保存変数

ターミナル ノード番号。 各ケースが割り当てられているターミナル ノード。この値はツリーのノード番号です。

予測値。 モデルにより予測される従属変数のクラス（グループ）、または値。

予測確率。 モデルの予測に関する確率。従属変数のカテゴリごとに変数が 1 つ保存されます。スケール従属変数には使用できません。

サンプルの割り当て（学習/検証用）。 分割サンプル検証では、ケースが学習用サンプルで使用されたのか検証用サンプルで使用されたのかが、この変数からわかります。学習用サンプルは値が 1 で、検証用サンプルの場合は 0 です。分割サンプル検証を選択していない場合は使用できません。詳細は、 p.8 検証(V) を参照してください。

XML としてツリー モデルをエクスポート

ツリー モデル全体を XML（PMML）形式で保存できます。このモデル ファイルを使用して、得点付けのために他のデータ ファイルにモデル情報を適用できます。

学習サンプル。 指定されたファイルにモデルを書き込みます。分割サンプル検証を適用したツリーの場合、このモデルが学習用サンプルになります。

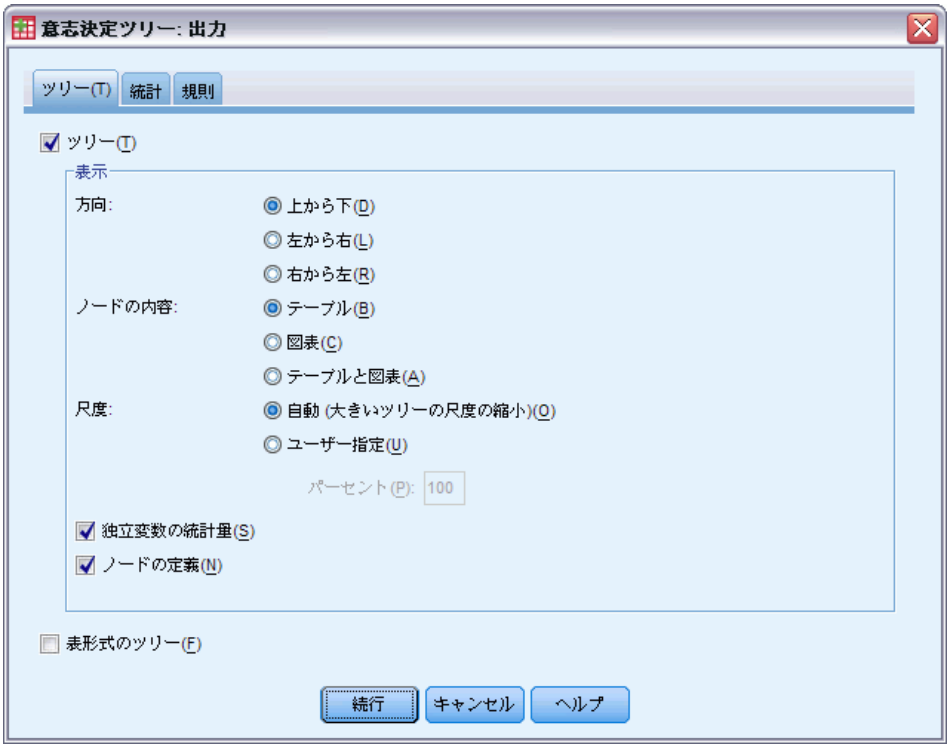
検証サンプル。 指定されたファイルに検証用サンプルのモデルを書き込みます。分割サンプル検証を選択していない場合は使用できません。

出力

成長手法、従属変数の尺度、およびその他の設定によって、使用できる出力オプションが異なります。

ツリー表示

図 1-19
[出力] ダイアログ ボックスの [ツリー] タブ



ツリーの外観の初期状態を制御したり、ツリーを完全に非表示にすることができます。

ツリー。 デフォルトでは、ビューアに表示される出力にツリー図が表示されます。出力からツリー図を除外するには、このオプションの選択を解除します（チェックを外します）。

表示。 ビューアに表示されるツリー図の外観の初期状態を制御するオプションです。生成されたツリーを編集することで、すべての属性を変更できます。

- **方向。** ツリーは、ルート ノードが最上部にきて下に伸びる形、左から右、右から左、のいずれかで表示できます。
- **ノードの内容。** テーブル、グラフ、またはその両方の形式でノードを表示できます。カテゴリ従属変数の場合は、テーブルに度数とパーセントが表示され、グラフは棒グラフが表示されます。スケール変数の場合は、テーブルに平均値、標準偏差、ケース数、予測値が表示され、グラフはヒストグラムが表示されます。

- **スケール。** デフォルトでは、ツリーが大きい場合、ツリーがページに合せて自動的に縮小表示されます。スケールの割合は、200% までの範囲で自由に指定できます。
- **独立変数の統計量。** CHAID および Exhaustive CHAID では、有意確率と自由度以外に、F 値（スケール従属変数の場合）、またはカイ 2 乗値（カテゴリ従属変数の場合）が統計に含まれます。CRT では、改善度が表示されます。QUEST では、スケール独立変数と順序独立変数の場合は F、有意確率、自由度が表示されます。名義独立変数の場合はカイ 2 乗、有意確率、自由度が表示されます。
- **ノードの定義。** ノードの定義には、各ノード分割で使用される独立変数の値が表示されます。

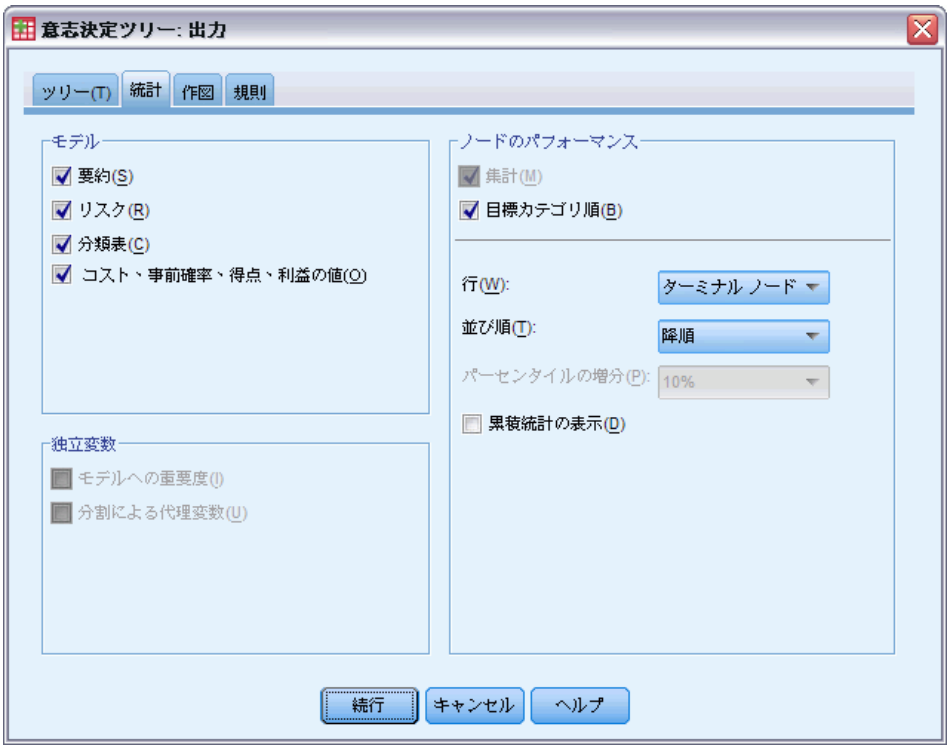
表形式のツリー。 親ノードの番号、独立変数の統計量、ノードの独立変数の値、スケール従属変数の平均値と標準偏差、カテゴリ従属変数の度数とパーセントなど、ツリーの各ノードの概要です。

図 1-20
表形式のツリー(F)

	悪い		良い		合計				1 次独立変数				
ノード	パーセント	度数	パーセント	度数	度数	パーセント	予測力カテゴリ	親ノード	変数	有意確率[a]	カイ 2 乗	自由度	分割値
0	41.4%	1020	58.6%	1444	2464	100.0%	良い						
1	82.1%	454	17.9%	99	553	22.4%	悪い	0	所得レベル	.000	662.457	2	<= 低
2	42.0%	476	58.0%	658	1134	46.0%	良い	0	所得レベル	.000	662.457	2	(低, 中]
3	11.6%	90	88.4%	687	777	31.5%	良い	0	所得レベル	.000	662.457	2	> 中
4	56.7%	422	43.3%	322	744	30.2%	悪い	2	クレジットカードの数	.000	193.113	1	5枚以上
5	13.8%	54	86.2%	336	390	15.8%	良い	2	クレジットカードの数	.000	193.113	1	5枚未満
6	17.6%	80	82.4%	375	455	18.5%	良い	3	クレジットカードの数	.000	38.587	1	5枚以上
7	3.1%	10	96.9%	312	322	13.1%	良い	3	クレジットカードの数	.000	38.587	1	5枚未満
8	80.8%	211	19.2%	50	261	10.6%	悪い	4	年齢	.000	95.299	1	<= 28.079205819899676
9	43.7%	211	56.3%	272	483	19.6%	良い	4	年齢	.000	95.299	1	...

統計

図 1-21
[分類ツリー: 出力] ダイアログ ボックスの [統計] タブ



従属変数の尺度、成長手法、およびその他の設定によって、使用できる統計テーブルが異なります。

モデル

要約。 集計には、使用している成長手法、モデルに含まれている変数、モデルに含まれていなくても指定されている変数が含まれます。

図 1-22
モデルの要約表

指定	成長方法	CHAID
	従属変数	信用度
	独立変数	年齢, 所得レベル, クレジットカードの数, 教育, 車のローン
	検証	なし
	ツリーの最大の深さ	3
	親ノードの最小ケース	400
	子ノードの最小ケース	200
結果	含まれている独立変数	所得レベル, クレジットカードの数, 年齢
	ノードの数	10
	ターミナルノードの数	6
	ツリーの深さ	3

リスク。 リスク推定値と標準誤差。ツリーの予測精度を測定する基準です。

- カテゴリ従属変数の場合、リスク推定値とは事前確率と誤分類コストを調整した後で誤って分類されたケースの割合です。
- スケール従属変数の場合、リスク推定値とはノード内の偏差です。

分類表。 カテゴリ（名義、順序）従属変数の場合、その従属変数の各カテゴリで正しく分類されたケースと誤って分類されたケースの数が表示されます。スケール従属変数には使用できません。

図 1-23
誤差テーブルと分類テーブル

相対リスク

推定値	標準誤差
.205	.008

成長手法: CHAID
従属変数: 信用度

分類

	予測値		
観測	悪い	良い	正解の割合
悪い	665	355	65.2%
良い	149	1295	89.7%
全体のパーセント	33.0%	67.0%	79.5%

成長手法: CHAID
従属変数: 信用度

コスト、事前確率、得点、利益の値。 カテゴリ従属変数の場合、分析に使用したコスト、事前確率、得点、利益の値が表示されます。スケール従属変数には使用できません。

独立変数(V)

モデルへの重要度。 CRT 成長手法の場合に、モデルへの重要度に応じて各独立（予測）変数の順位を付けます。QUEST 成長手法および CHAID 成長手法では使用できません。

分割による代理変数。 CRT 成長手法、および QUEST 成長手法で、モデルに代理変数が含まれている場合に、ツリーの分割ごとに代理変数を列挙します。CHAID 成長手法には利用できません。 詳細は、 p. 17 代理変数 を参照してください。

ノードのパフォーマンス

要約。 スケール従属変数の場合、ノード番号、ケース数、および従属変数の平均値が表示されます。利益が定義されているカテゴリ従属変数の場合、ノード番号、ケース数、平均利益、および ROI（投資収益率）の値が表示されます。利益が定義されていないカテゴリ従属変数には使用できません。 詳細は、 p. 19 利益 を参照してください。

図 1-24
ノードとパーセンタイルを示すゲインの要約テーブル

ノードのゲイン要約

ノード	N	パーセント	利益	ROI
7	322	13.1%	77.826	377.4%
5	390	15.8%	70.308	308.8%
6	455	18.5%	67.692	287.9%
9	483	19.6%	49.420	172.0%
8	261	10.6%	23.410	64.7%
1	553	22.4%	22.532	61.9%

パーセンタイルのゲイン要約

パーセンタイル	ノード	N	利益	ROI
10	7	246	77.826	377.4%
20	7 ; 5	493	75.218	352.0%
30	5 ; 6	739	73.488	336.2%
40	6	986	72.036	323.4%
50	6 ; 9	1232	70.205	307.9%
60	9	1478	66.745	280.6%
70	9 ; 8	1725	63.134	254.4%
80	8 ; 1	1971	58.149	221.6%
90	1	2218	54.183	197.9%
100	1	2464	51.023	180.4%

目標カテゴリ順。 目標カテゴリが定義されているカテゴリ従属変数の場合、ノードまたはパーセンタイル グループごとのパーセント ゲイン、応答割合、インデックス割合（リフト）が表示されます。目標カテゴリごとに独立したテーブルが作成されます。目標カテゴリが定義されていないスケール従属変数およびカテゴリ従属変数には使用できません。 詳細は、 p. 6 カテゴリの選択 を参照してください。

図 1-25
ノードとパーセンタイルを示す目標カテゴリ ゲイン

Target Category: 悪い

ノードのゲイン

	ノード		ゲイン			
ノード	N	パーセント	N	パーセント	回答	インデックス
1	553	22.4%	454	44.5%	82.1%	198.3%
8	261	10.6%	211	20.7%	80.8%	195.3%
9	483	19.6%	211	20.7%	43.7%	105.5%
6	455	18.5%	80	7.8%	17.6%	42.5%
5	390	15.8%	54	5.3%	13.8%	33.4%
7	322	13.1%	10	1.0%	3.1%	7.5%

パーセンタイルのゲイン

			ゲイン			
パーセンタイル	ノード	N	N	パーセント	回答	インデックス
10	1	246	202	19.8%	82.1%	198.3%
20	1	493	405	39.7%	82.1%	198.3%
30	1 ; 8	739	604	59.3%	81.8%	197.6%
40	8 ; 9	986	740	72.6%	75.1%	181.3%
50	9	1232	848	83.1%	68.8%	166.2%
60	9 ; 6	1478	908	89.0%	61.4%	148.4%
70	6	1725	951	93.3%	55.1%	133.2%
80	6 ; 5	1971	986	96.7%	50.0%	120.9%
90	5 ; 7	2218	1012	99.3%	45.6%	110.3%
100	7	2464	1020	100.0%	41.4%	100.0%

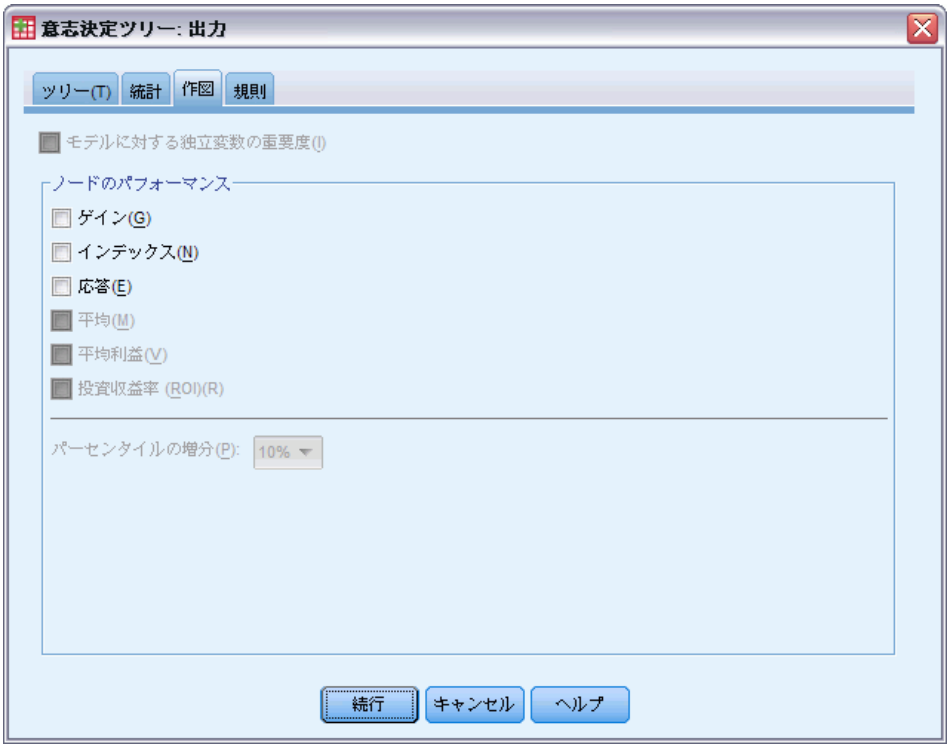
行。 ノードのパフォーマンス テーブルには、ターミナル ノード、パーセンタイル、またはその両方で結果を表示できます。両方を選択すると、目標カテゴリごとに 2 つのテーブルが作成されます。パーセンタイル テーブルには、各パーセンタイルの値が、並べ替え順に従って累積されて表示されます。

パーセンタイルの増分。 パーセンタイル テーブルについて、1、2、5、10、20、または 25 からパーセンタイルの増分を選択できます。

累積統計の表示。 ターミナル ノード テーブルの場合に、それぞれのテーブルに列を追加して、そこに累積した結果を表示します。

図表

図 1-26
[出力] ダイアログ ボックスの [作図] タブ



従属変数の尺度、成長手法、およびその他の設定によって、使用できるグラフが異なります。

モデルに対する独立変数の重要度。 独立変数（予測変数）のモデルの重要度を示す棒グラフ。これを利用できるのは、CRT 成長手法だけです。

ノードのパフォーマンス

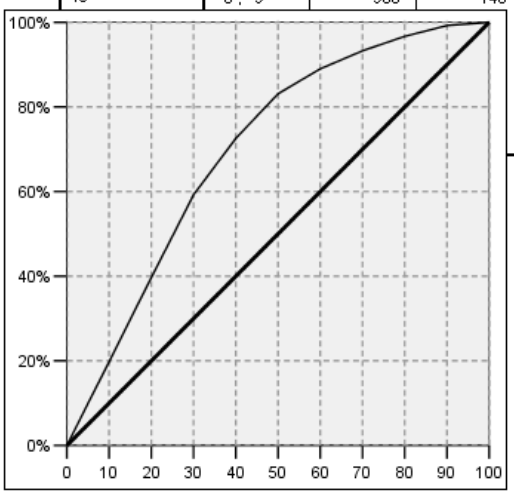
ゲイン。： ゲインは、各ノードの目標カテゴリの合計ケースのパーセントです。(node target n / total target n) x 100 の数式で算出します。ゲイン グラフはパーセンタイル ゲインの累計を表した折れ線グラフのことで、(目標の累積パーセンテージ n/目標の合計 n) x 100 で計算します。目標カテゴリごとに独立した折れ線グラフが作成されます。目標カテゴリが定義されたカテゴリ従属変数にのみ使用できます。 詳細は、 p.6 カテゴリの選択 を参照してください。

ゲイン グラフには、パーセンタイル テーブルのゲインの [ゲインのパーセント] 列と同じ値がプロットされます。この列も累積値を表しています。

図 1-27
パーセンタイル テーブルのゲインとゲイン グラフ

パーセンタイルのゲイン

パーセンタイル	ノード	N	ゲイン N	ゲイン パーセント	回答	インデックス
10	1	246	202	19.8%	82.1%	198.3%
20	1	493	405	39.7%	82.1%	198.3%
30	1 ; 8	739	604	59.3%	81.8%	197.6%
40	8 ; 9	986	740	72.6%	75.1%	181.3%
				83.1%	68.8%	166.2%
				89.0%	61.4%	148.4%
				93.3%	55.1%	133.2%
				96.7%	50.0%	120.9%
				99.3%	45.6%	110.3%
				100.0%	41.4%	100.0%



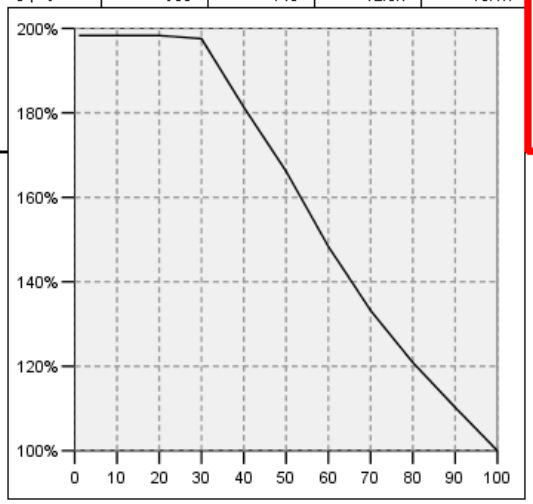
インデックス. インデックスは比較される目標カテゴリのノード応答パーセントをすべてのサンプルのすべての目標カテゴリ応答パーセントで割った比です。 インデックス グラフは累積パーセンタイルのインデックス値を表す折れ線グラフです。カテゴリ従属変数にのみ使用できます。積パーセンタイルのインデックスは、（cumulative percentile response percent / total response percent）x 100 で計算します。目標カテゴリごとに独立した折れ線グラフが作成され、目標カテゴリを定義する必要があります。

インデックス グラフには、パーセンタイル テーブルのゲインの［インデックス］列と同じ値がプロットされます。

図 1-28
パーセンタイル テーブルのゲインとインデックス グラフ

パーセンタイルのゲイン

パーセンタイル	ノード	N	ゲイン N	ゲイン パーセント	回答	インデックス
10	1	246	202	19.8%	82.1%	198.3%
20	1	493	405	39.7%	82.1%	198.3%
30	1 ; 8	739	604	59.3%	81.8%	197.6%
40	8 ; 9	986	740	72.6%	75.1%	181.3%
50						166.2%
60						148.4%
70						133.2%
80						120.9%
90						110.3%
100						100.0%



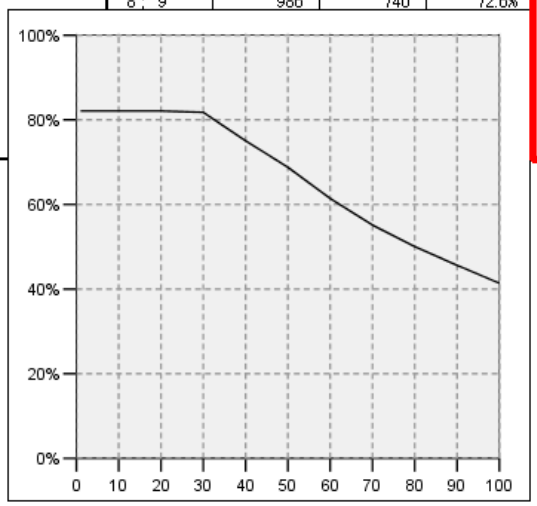
応答. 指定された目標カテゴリのノードにおけるケースのパーセントです。レスポンス グラフはパーセンタイル レスポンスの累計を表した折れ線グラフのことで、(目標の累積パーセンタイル n / 累積合計パーセンタイル n) x 100 で計算します。目標カテゴリが定義されたカテゴリ従属変数にのみ使用できます。

応答グラフには、パーセンタイル テーブルのゲインの［応答］列と同じ値がプロットされます。

図 1-29
パーセンタイル テーブルのゲインと応答グラフ

パーセンタイルのゲイン

パーセンタイル	ノード	N	ゲイン N	ゲイン パーセント	回答	インデックス
10	1	246	202	19.8%	82.1%	198.3%
20	1	493	405	39.7%	82.1%	198.3%
30	1 ; 8	739	604	59.3%	81.8%	197.6%
40	8 ; 9	986	740	72.6%	75.1%	181.3%
50					68.8%	166.2%
60					61.4%	148.4%
70					55.1%	133.2%
80					50.0%	120.9%
90					45.6%	110.3%
100					41.4%	100.0%



平均値。 従属変数の累積パーセンタイルの平均値を表す折れ線グラフ。スケール従属変数にのみ使用できます。

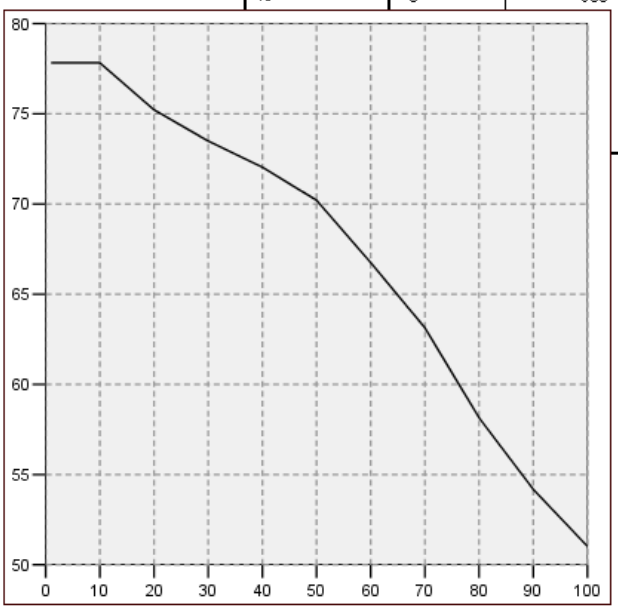
平均利益。 平均利益の累積を表す折れ線グラフ。利益が定義されているカテゴリ従属変数にのみ使用できます。 詳細は、 p. 19 利益 を参照してください。

平均利益グラフには、パーセンタイル テーブルのゲインの要約の［利益］列と同じ値がプロットされます。

図 1-30
パーセンタイル テーブルのゲインの要約と平均利益グラフ

パーセンタイルのゲイン要約

パーセンタイル	ノード	N	利益	ROI
10	7	246	77.826	377.4%
20	7 ; 5	493	75.218	352.0%
30	5 ; 6	739	73.488	336.2%
40	6	986	72.036	323.4%
			70.205	307.9%
			66.745	280.6%
			63.134	254.4%
			58.149	221.6%
			54.183	197.9%
			51.023	180.4%



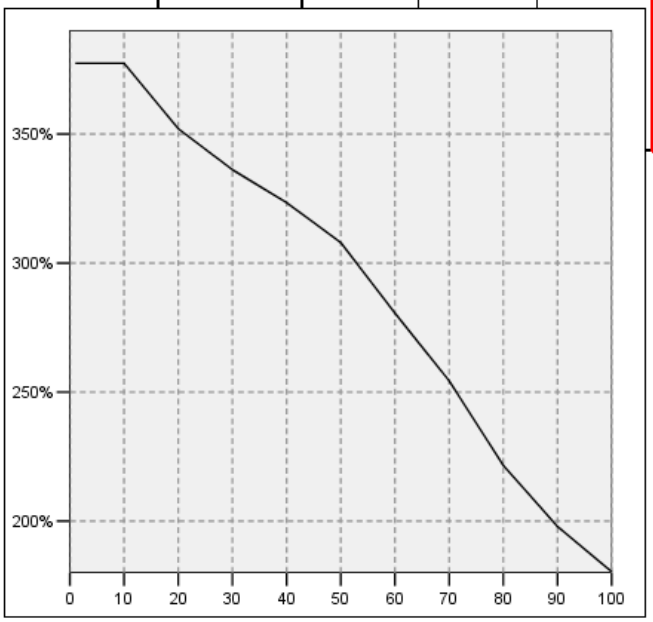
投資収益率 (ROI)。 ROI（投資におけるリターン）の累積を表す折れ線グラフ。ROI は利益対出費の比率で計算します。利益が定義されているカテゴリ従属変数にのみ使用できます。

ROI グラフには、パーセンタイル テーブルのゲインの要約の［ROI］列と同じ値がプロットされます。

図 1-31
パーセンタイル テーブルのゲインの要約と ROI グラフ

パーセンタイルのゲイン要約

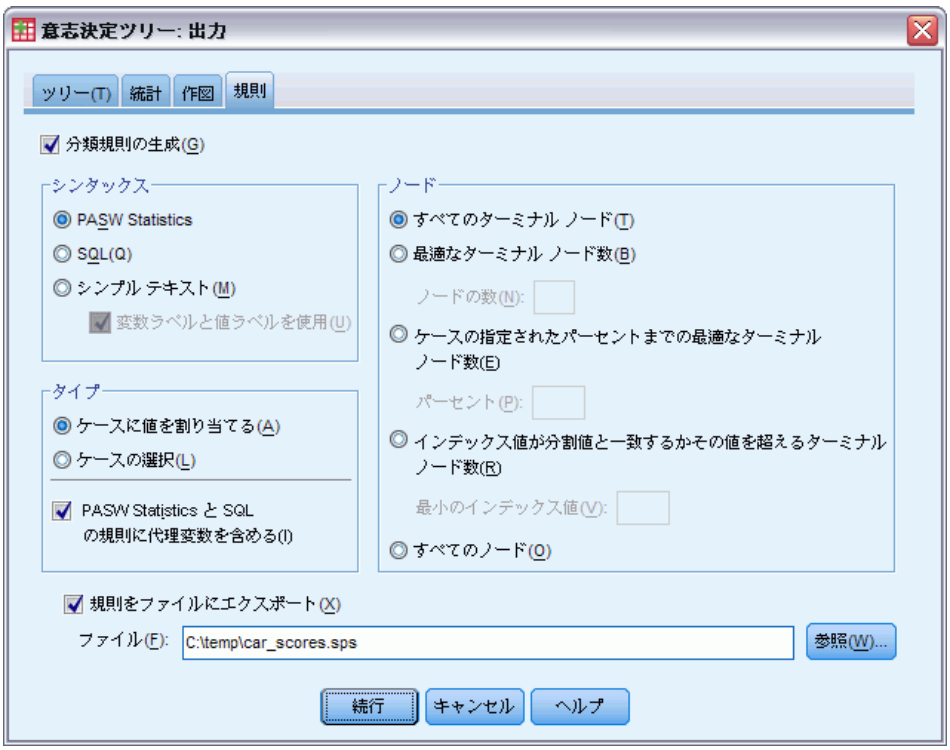
パーセンタイル	ノード	N	利益	ROI
10	7	246	77.826	377.4%
20	7 ; 5	493	75.218	352.0%
30	5 ; 6	739	73.488	336.2%
40	6	986	72.036	323.4%
				307.9%
				280.6%
				254.4%
				221.6%
				197.9%
				180.4%



パーセンタイルの増分。すべてのパーセンタイル グラフについて、この設定は、グラフに表示されるパーセンタイルの増分（1、2、5、10、20、または 25）を制御できます。

選択規則と得点規則

図 1-32
[分類ツリー: 出力] ダイアログ ボックスの [規則] タブ



[規則] タブでは、選択、分類、予測の規則をコマンド シンタックス、SQL、および通常の（平易な英語の）テキストで生成できます。生成した規則は、ビューアに表示したり、外部ファイルに保存したりできます。

シンタックス。 ビューアに表示される出力と外部ファイルに保存された選択規則の両方で選択規則の形式を指定します。

- **IBM® SPSS® Statistics。**コマンド シンタックス言語規則は、ケースのサブグループ選択に使用できるフィルタ条件を定義するための一連のコマンドとして表されるか、またはケースの得点付けに使用できる COMPUTE ステートメントとして表されます。
- **SQL。** 標準の SQL 規則を生成して、データベースからレコードの選択や抽出を行ったり、そのようなレコードに値を割り当てたりします。生成された SQL の規則は、テーブル名やその他のデータ ソース情報を一切含みません。
- **シンプル テキスト。** 平易な英語の擬似コード。規則は、ノードごとのモデルの分類や予測を記述する「if...then」という論理ステートメントのセットとして表されます。この形式の規則では、定義済みの変数ラベルと値ラベル、または変数名とデータ値を使用できます。

型。SPSS Statistics と SQL の規則では、生成される規則の種類、つまり選択規則または得点規則を指定します。

- **ケースに値を割り当てる。** 規則を使用して、ノードの所属条件を満たすケースにモデルの予測を割り当てることができます。ノードの所属条件を満たすノードごとに個別の規則が生成されます。
- **ケースの選択。** 規則を使用して、ノードの所属条件を満たすケースを選択できます。SPSS Statistics と SQL の規則では、1 つの規則を生成して、選択条件を満たすすべてのケースを選択します。

[SPSS Statistics と SQL の規則に代理変数を含める]CRT と QUEST の場合、規則にはモデルから代理予測変数を含めることができます。代理変数を含む規則は、非常に複雑になる場合があります。一般に、ツリーに関する概念情報を単に得る場合は、代理変数を除外します。一部のケースに不完全な独立変数（予測変数）があり、ツリーを模倣する規則が必要な場合は、代理変数を含めます。 詳細は、 p. 17 代理変数 を参照してください。

ノード。 生成する規則の範囲を制御します。範囲内のそれぞれのノードごとに、個別の規則が生成されます。

- **すべてのターミナル ノード。** ターミナル ノードごとに規則を生成します。
- **最適なターミナル ノード数。** インデックス値の上位 n 個のターミナル ノードに対して規則を生成します。この数値がツリー内のターミナル ノードの数より大きい場合、すべてのターミナル ノードに対して規則が生成されます。(下記の注を参照してください)。
- **ケースの指定されたパーセントまでの最適なターミナル ノード数。** インデックス値の上位 n 個のターミナル ノードに対して規則を生成します (下記の注を参照してください)。
- **インデックス値が分割値と一致するかその値を超えるターミナル ノード数。** インデックス値が指定した値以上のすべてのターミナル ノードに対して規則を生成します。インデックス値が 100 を超えると、そのノードの目標カテゴリのケースのパーセントが、ルート ノードのパーセントを超えていると見なされます。(下記の注を参照してください)。
- **すべてのノード。** すべてのノードに規則を生成します。

注 1 :インデックス値に基づくノードの選択は、目標カテゴリが定義されているカテゴリ従属変数にのみ使用できます。目標カテゴリを複数指定している場合、目標カテゴリごとに個別の規則のセットが生成されます。

注 2 :SPSS Statistics および SQL で (値を割り当てる規則ではなく) ケースを選択するための規則に、**[すべてのノード]** と **[すべてのターミナル ノード]** を選択すると、分析に使用するすべてのケースが選択される規則を効果的に生成できます。

規則をファイルにエクスポート。 規則を外部テキスト ファイルに保存します。

最終的なツリー モデルで選択されているノードを基に、選択規則や得点規則をダイアログから生成および保存することもできます。 詳細は、 2 章 p. 49 ケースの選択規則と得点規則 を参照してください。

注 :コマンド シンタックス形式の規則を別のデータ ファイルに適用する場合、そのデータ ファイルには、最終モデルに含まれる独立変数と同じ名前の変数が含まれている必要があります。これらの変数は、同じ測定基準で測定され、ユーザー定義の欠損値がある場合は同じ欠損値を持っている必要があります。

章

2

ツリー エディタ

ツリー エディタでは、次の操作ができます。

- 選択したツリーの枝の表示と非表示を切り替える。
- ノードの内容、ノードの分割で表示される統計量、および他の情報の表示を制御する。
- ノード、背景、罫線、図表、フォントの色を変更する。
- フォントのスタイルやサイズを変更する。
- ツリーの位置合わせを変更する。
- 選択したノードに基づいてケースのサブグループを選択して、さらに分析を進める。
- 選択したノードに基づいてケースの選択や得点を行う規則を作成して保存する。

ツリー モデルを編集するには、次の手順を実行します。

▶ [ビューア] ウィンドウでツリー モデルをダブルクリックします。

または

▶ [編集] メニューまたはマウス右ボタン コンテキスト メニューから次の項目を選択します。
内容編集(O)... > 別ウィンドウ(W)

ノードの表示と非表示の切り替え

親ノードの下にある 1 つの枝のすべての子ノードを非表示にする（閉じる）には、次の手順を実行します。

▶ 親ノードの右下隅にある小さなボックスのマイナス符号 (-) をクリックします。

その枝の親ノードより下にあるすべてのノードが非表示になります。

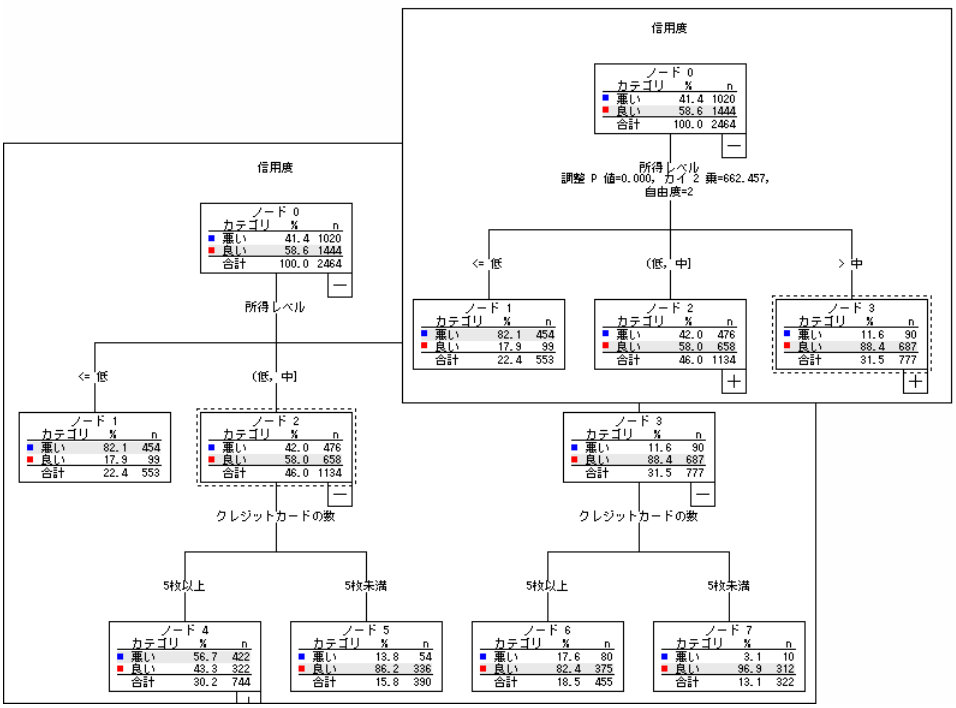
親ノードの下にある 1 つの枝の子ノードを表示する（開く）には、次の手順を実行します。

▶ 親ノードの右下隅にある小さなボックスのプラス符号 (+) をクリックします。

© Copyright IBM Corporation 1989, 2012.

注 :枝の子ノードを非表示にすることは、ツリーの剪定とは異なります。剪定されたツリーが必要な場合は、ツリーを作成する前に剪定を要求する必要があります。また、剪定された枝は最終的なツリーには含まれません。詳細は、 1 章 p. 16 ツリーの剪定 を参照してください。

図 2-1
展開ツリーと縮小ツリー



複数のノードの選択

ケースの選択、得点規則と選択規則の生成、および現在選択しているノードに基づいた他の操作を実行できます。複数のノードを選択するには、次の手順を実行します。

- ▶ 選択するノードをクリックします。
- ▶ Ctrl キーを押しながら、選択する他のノードをクリックします。

ある枝の兄弟ノードと親ノードの両方またはいずれか、および別の枝の子ノードを複数選択できます。ただし、同じノードの枝の親ノードと子または孫では、複数選択を使用できません。

大きなツリーを使用した作業

ツリー モデルには、完全なサイズでツリー全体を表示することが困難または不可能である非常に多くのノードや枝が含まれている場合があります。次に示すように、大きなツリーを使用して作業する際に役に立つ機能が多数あります。

- **ツリー マップ。** 通常のツリーより大幅に小さい、ツリーの簡略版であるツリー マップを使用して、ツリーを移動し、ノードを選択できます。詳細は、 p. 44 ツリー マップ を参照してください。
- **尺度変更。** ツリー表示の尺度パーセントを変更して、縮小および拡大できます。 詳細は、 p. 45 ツリー表示の尺度変更 を参照してください。
- **ノードと枝の表示。** ノードにテーブルまたは図表のみを表示したり、ノード ラベルまたは独立変数情報の表示を抑制したりすることで、ツリーをコンパクトにすることができます。 詳細は、 p. 47 ツリーに表示される情報の制御 を参照してください。

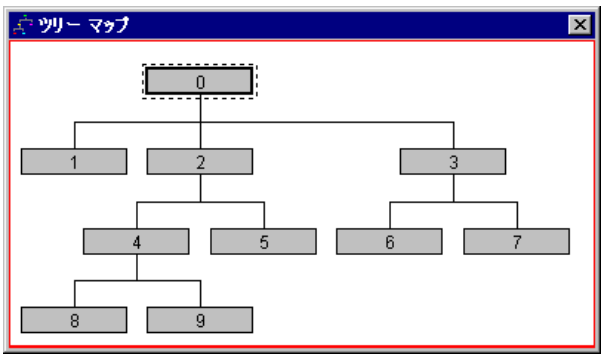
ツリー マップ

ツリー マップでは、ツリーの移動やノードの選択に使用できるツリーがコンパクトに簡素化されて表示されます。

ツリー マップのウィンドウを使用するには、次の手順を実行します。

▶ ［ツリー エディタ］メニューから次の項目を選択します。
表示 > ツリー マップ

図 2-2
ツリー マップのウィンドウ



- 現在選択されているノードは、ツリー モデル エディタとツリー マップのウィンドウの両方で強調表示されます。

- 現在ツリー モデル エディタの表示領域に表示されているツリーの部分は、ツリー マップでは赤色の四角形で示されます。表示領域に表示されているツリーのセクションを変更するには、四角形を右クリックしてドラッグします。
- 現在ツリー エディタの表示領域に表示されていないノードをツリー マップで選択すると、選択したノードもその表示領域に表示されます。
- 複数のノードを選択するには、ツリー エディタでツリー マップを選択する場合と同じように、Ctrl キーを押しながら複数のノードをクリックします。同じノードの枝の親ノードと子または孫では、複数選択を使用できません。

ツリー表示の尺度変更

デフォルトでは、ツリーは［ビューア］ウィンドウに適合するように自動的に尺度が変更されます。そのため、一部のツリーは最初は読み取りにくくなる可能性があります。定義済みの尺度設定を選択するか、5 ～ 200% のユーザー指定の尺度値を独自に入力できます。

ツリーの尺度を変更するには、次の手順を実行します。

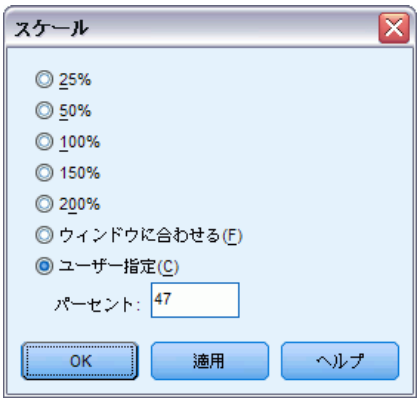
▶ ツールバーに表示されるドロップダウン リストから尺度パーセントを選択するか、ユーザー指定のパーセント値を入力します。

または

▶ ［ツリー エディタ］メニューから次の項目を選択します。

表示 > 尺度(C)...

図 2-3
[スケール] ダイアログ ボックス



ツリー モデルを作成する前に尺度値を指定することもできます。 詳細は、 1 章 p. 26 出力 を参照してください。

ノードの要約ウィンドウ

ノードの要約ウィンドウには、選択したノードが大きく表示されます。また、要約ウィンドウを使用すると、選択したノードに基づいて選択規則または得点規則を表示、適用、保存できます。

- 要約表、図表、または規則の表示を切り替えるには、ノードの要約ウィンドウの［表示］メニューを使用します。
- 表示する規則の種類を選択するには、ノードの要約ウィンドウの［規則］メニューを使用します。 詳細は、 p. 49 ケースの選択規則と得点規則 を参照してください。
- ノードの要約ウィンドウのすべての表示には、選択されたすべてのノードの要約が結合されて反映されます。

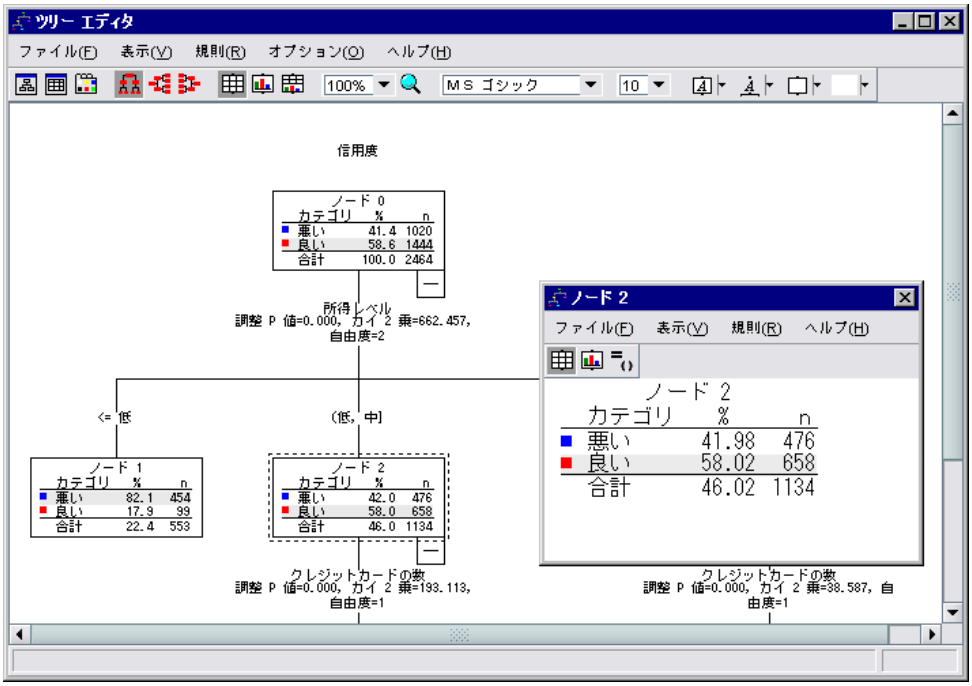
ノードの要約ウィンドウを使用するには、次の手順を実行します。

▶ ツリー エディタでノードを選択します。複数のノードを選択するには、Ctrl キーを押しながらクリックします。

▶ メニューから次の項目を選択します。

表示 > 要約表

図 2-4
［要約］ウィンドウ

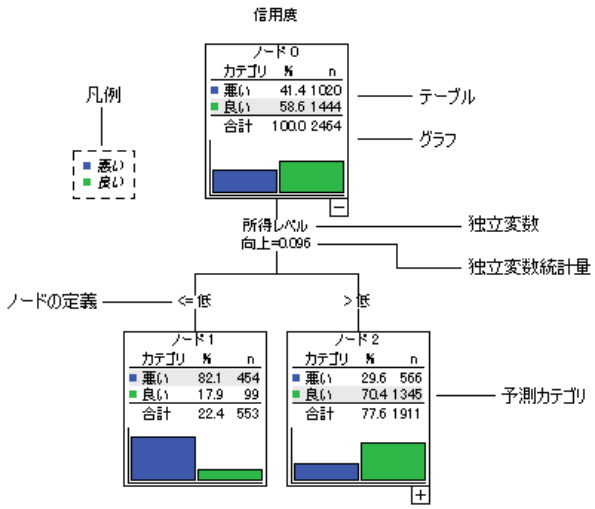


ツリーに表示される情報の制御

ツリー エディタの［オプション］メニューでは、ノードの内容の表示、独立変数（予測変数）の名前と統計量、ノードの定義、その他の設定を制御できます。これらの設定の多くは、ツールバーからも制御できます。

設定	［オプション］メニュー選択
予測カテゴリ（カテゴリ従属変数）の強調	予測値のハイライト表示(H)
ノードのテーブルまたは図表、あるいは両方	ノードの内容(N)
有意差検定値と p 値	独立変数の統計(I)
独立（予測）変数名	独立変数(V)
ノードの独立（予測）変数値	ノードの定義(D)
配置（上から下、左から右、右から左）	方向
図表の凡例	凡例

図 2-5
ツリー要素



ツリーの色とテキスト フォントの変更

ツリーで次の色を変更できます。

- ノードの罫線、背景、およびテキストの色
- 枝の色と枝のテキストの色
- ツリーの背景色
- 予測カテゴリの強調表示色（カテゴリ従属変数）
- ノード図表の色

また、ツリー内のすべてのテキストのフォントの種類、スタイル、およびサイズを変更できます。

注 ：ノードや枝の色またはフォント属性を個別に変更することはできません。色の変更は同じ種類のすべての要素に適用され、フォントの変更（色以外）はすべての図表要素に適用されます。

色およびテキストのフォント属性を変更するには、次の手順を実行します。

- ツールバーを使用して、ツリー全体のフォント属性、またはさまざまなツリー要素の色を変更します。ツールバーに表示される各コントロールの上にマウス カーソルを置くと、各コントロールの説明が表示されます。

  または

- ツリー エディタ内の任意の場所をダブルクリックして［プロパティ］ウィンドウを開くか、メニューから次の項目を選択します。
  表示 > プロパティ

- 罫線、枝、ノードの背景、予測カテゴリ、ツリーの背景の場合は、[色] タブをクリックします。

- フォントの色と属性の場合は、[テキスト] タブをクリックします。

- ノード図表の色の場合は、[ノード図表] タブをクリックします。

図 2-6
[プロパティ] ウィンドウ -> [色] タブ

図 2-7
[プロパティ] ウィンドウ -> [テキスト] タブ

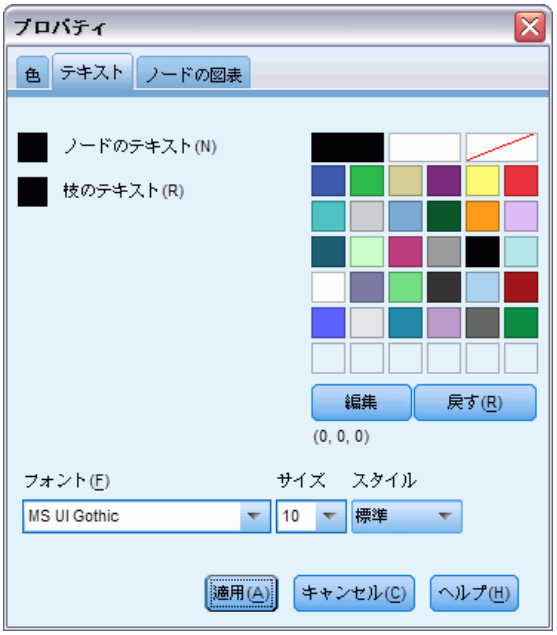


図 2-8
[プロパティ] ウィンドウ -> [ノード図表] タブ

ケースの選択規則と得点規則

ツリー エディタを使用すると、次のことができます。

- 選択したノードに基づいて、ケースのサブグループを選択します。 詳細は、 p. 50 分析からのケースの除外 を参照してください。
- IBM® SPSS® Statistics コマンド シンタックスまたは SQL 形式で、ケース選択規則または得点規則を生成します。 詳細は、 p. 50 選択規則と得点規則の保存 を参照してください。

また、ディシジョン ツリー手続きを実行してツリー モデルを作成するときに、さまざまな条件に基づいて規則を自動的に保存することもできます。 詳細は、 1 章 p. 39 選択規則と得点規則 を参照してください。

分析からのケースの除外

特定のノードまたはノードのグループのケースについてより詳しく理解するには、さらに分析を行うために、選択したノードに基づいてケースのサブグループを選択できます。

▶ ツリー エディタでノードを選択します。複数のノードを選択するには、Ctrl キーを押しながらクリックします。

▶ メニューから次の項目を選択します。
規則 > ケースのフィルタリング(F)...

▶ フィルタ変数の名前を入力します。選択したノードのケースには、この変数に対し値 1 が返されます。その他のすべてのケースには値 0 が返されます。これらのケースは、フィルタの状態（ケースの除外の状態）を変更するまで、それ以降の分析から除外されます。

▶ [OK] をクリックします。

図 2-9
[ケースのフィルタリング] ダイアログ ボックス

選択規則と得点規則の保存

ケースの選択規則または得点規則を外部ファイルに保存して、それらの規則を別のデータ ソースに適用できます。これらの規則は、ツリー エディタで選択したノードに基づきます。

シンタックス。 ビューアに表示される出力と外部ファイルに保存された選択規則の両方で選択規則の形式を指定します。

- **IBM® SPSS® Statistics。**コマンド シンタックス言語規則は、ケースのサブグループ選択に使用できるフィルタ条件を定義するための一連のコマンドとして表されるか、またはケースの得点付けに使用できる COMPUTE ステートメントとして表されます。
- **SQL。** 標準の SQL 規則を生成して、データベースからレコードの選択や抽出を行ったり、そのようなレコードに値を割り当てたりします。生成された SQL の規則は、テーブル名やその他のデータ ソース情報を一切含みません。

型。 選択規則または得点規則を作成できます。

- **ケースの選択。** 規則を使用して、ノードの所属条件を満たすケースを選択できます。SPSS Statistics と SQL の規則では、1 つの規則を生成して、選択条件を満たすすべてのケースを選択します。
- **ケースに値を割り当てる。** 規則を使用して、ノードの所属条件を満たすケースにモデルの予測を割り当てることができます。ノードの所属条件を満たすノードごとに個別の規則が生成されます。

代理変数を含める。 CRT と QUEST の場合、規則にはモデルから代理予測変数を含めることができます。代理変数を含む規則は、非常に複雑になる場合があります。一般に、ツリーに関する概念情報を単に得る場合は、代理変数を除外します。一部のケースに不完全な独立変数（予測変数）があり、ツリーを模倣する規則が必要な場合は、代理変数を含めます。 詳細は、 1 章 p.17 代理変数 を参照してください。

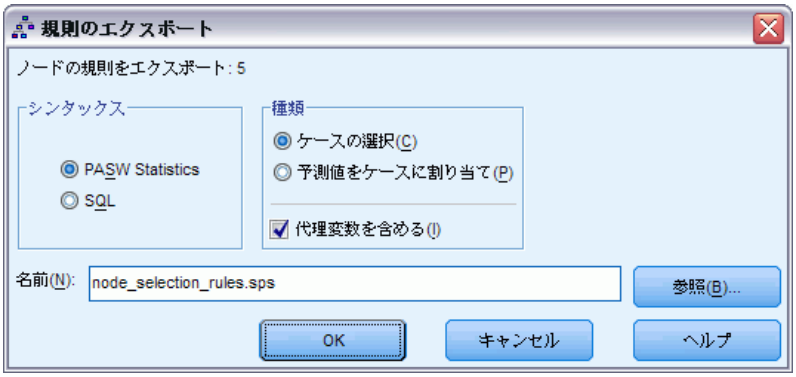
ケースの選択規則または得点規則を保存するには、次の手順を実行します。

▶ ツリー エディタでノードを選択します。複数のノードを選択するには、Ctrl キーを押しながらクリックします。

▶ メニューから次の項目を選択します。
規則 > エクスポート(E)...

▶ 必要な規則の種類を選択し、ファイル名を入力します。

図 2-10
[規則のエクスポート] ダイアログ ボックス



注 ：コマンド シンタックス形式の規則を別のデータ ファイルに適用する場合、そのデータ ファイルには、最終モデルに含まれる独立変数と同じ名前の変数が含まれている必要があります。これらの変数は、同じ測定基準で測定され、ユーザー定義の欠損値がある場合は同じ欠損値を持っている必要があります。

パート II:
例

章 3

データの仮定事項と必要条件

ディシジョン ツリー手続きでは、次の項目を仮定しています。

- 適切な尺度がすべての分析変数に割り当てられている。
- カテゴリ（**名義**、**順序**）従属変数の場合、分析の対象となるすべてのカテゴリに値ラベルが定義されている。

ファイル tree_textdata.sav を使用して、上記 2 つの必要条件の重要性を説明します。このデータ ファイルは、尺度や値ラベルなどの属性を定義する前の、読み込むか入力したデフォルト状態のデータを表しています。 詳細は、 A 付録 サンプル ファイル in IBM SPSS Decision Trees 21 を参照してください。

ツリー モデルでの尺度の効果

このデータ ファイルに含まれている変数は両方とも数値型で、どちらにも**スケール**尺度が割り当てられます。ただし、後述するように、両変数は実際には数値コードに依存してカテゴリ値を表すカテゴリ変数です。

▶ ディシジョン ツリー分析を実行するには、メニューから次の項目を選択します。

分析(A) > 分類 > ツリー...

© Copyright IBM Corporation 1989, 2012.

ソース変数リストに並ぶ 2 つの変数の左に付いているアイコンは、各変数がスケール変数として扱われることを示しています。

図 3-1
2 つのスケール変数が表示された [ディシジョン ツリー] メイン ダイアログ ボックス

- ［従属変数］として「従属」を選択します。
- ［独立変数］として「独立」を選択します。
- [OK] をクリックして手続きを実行します。
- ［ディシジョン ツリー］ダイアログ ボックスを再び開いて、[戻す] をクリックします。
- ソース リストの「従属」を右クリックし、コンテキスト メニューから [名義] を選択します。
- ソース リストの「独立」に対して同じ操作を行います。

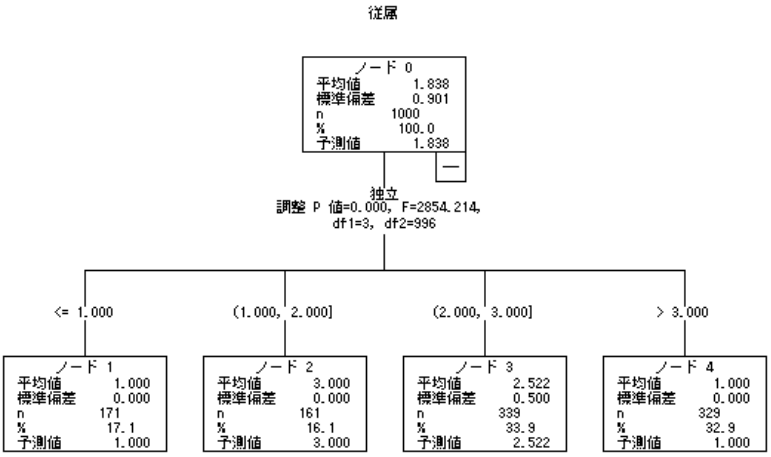
これで、2 つの変数の左にあるアイコンが変わり、各変数が名義変数として扱われることが示されます。

図 3-2
ソース リストの名義アイコン

▶ [従属変数] として「従属」を、[独立変数] として「独立」を選択し、[OK] をクリックして手続きを再実行します。

それでは、2 つのツリーを比較してみましょう。まず、2 つの数値型変数がスケール変数として扱われているツリーを見てみます。

図 3-3
両変数をスケール変数として扱うツリー

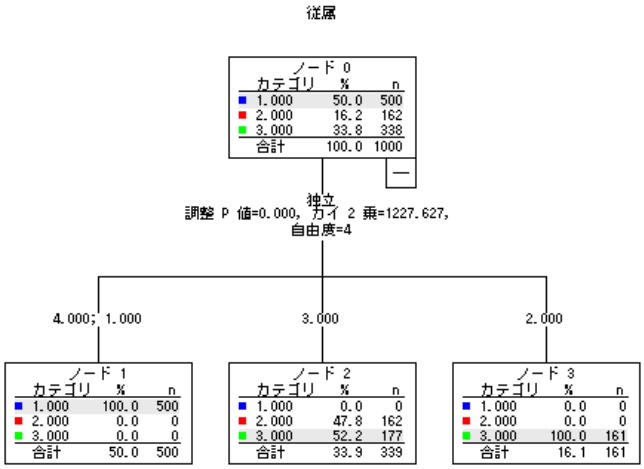


- ツリーの各ノードは、「予測」値を示しています。予測値は各ノードにおける従属変数の平均値を表します。変数が実際にはカテゴリ変数の場合、平均値は意味のある統計量にならない場合があります。
- ツリーには、独立変数の値ごとに 4 つの子ノードがあります。

ツリー モデルでは、類似したノードが結合されることがよくありますが、スケール変数の場合は連続値だけが結合されます。この例では、類似しているとみなされる連続値はなく、ノードは結合されていません。

両変数が名義変数として扱われるツリーでは、異なる点がいくつかあります。

図 3-4
両変数を名義変数として扱うツリー



- 各ノードにあるのは、予測値ではなく度数分布表です。この度数分布表では、従属変数のカテゴリごとのケース数（度数およびパーセント）が示されています。
- 「予測」カテゴリ、つまり各ノードで最高度数を持つカテゴリが強調表示されます。たとえば、ノード 2 の予測カテゴリはカテゴリ 3 です。
- 子ノードは 4 つではなく 3 つになっており、独立変数の 2 つの値が 1 つのノードに結合されています。

同一ノードに結合された独立変数の 2 つの値は、1 と 4 です。定義上、名義値に固有の順序はないので、連続していない値を結合できます。

[尺度] を永続的に割り当てる方法

[ディシジョン ツリー] ダイアログ ボックスで変数の尺度を変更する場合、その変更は一時的にしか適用されず、データ ファイルには保存されません。また、すべての変数に対して正しい尺度がわかっているとは限りません。

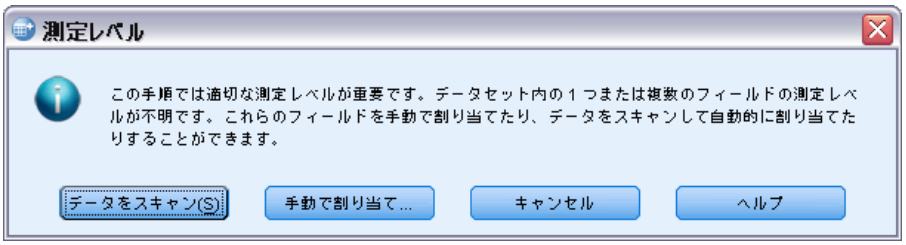
[変数プロパティの定義] を使えば、各変数の正しい測定レベルの特定と、割り当てた測定レベルの永続的な変更を行えます。[変数プロパティの定義] を使用するには、次の手順を実行します。

▶ メニューから次の項目を選択します。
データ > 変数プロパティの定義(V)...

不明な尺度の変数

データセットの 1 つまたは複数の変数（フィールド）の尺度が不明な場合、尺度の警告が表示されます。尺度はこの手順の結果の計算に影響を与えるため、すべての変数に尺度を定義する必要があります。

図 3-5
尺度の警告



- **データをスキャン。** アクティブ データセットのデータを読み込み、デフォルトの尺度を尺度が現在不明なフィールドに割り当てます。データセットが大きい場合は時間がかかります。
- **手動で割り当てる。** 不明な尺度のフィールドをすべて表示するダイアログが開きます。このダイアログを使用して、尺度をこれらのフィールドに割り当てることができます。データ エディタの［変数ビュー］でも、尺度を割り当てることができます。

尺度がこの手順で重要であるため、すべてのフィールドに尺度が定義されるまで、ダイアログにアクセスしてこの手順を実行することはできません。

ツリー モデルの値ラベルの効果

［ディシジョン ツリー］ダイアログ ボックスのインターフェイスでは、カテゴリ（名義、順序）従属変数のすべての非欠損値で値ラベルが定義されているか、あるいは値ラベルがいっさい定義されていないと仮定しています。 カテゴリ従属変数の少なくとも 2 つの非欠損値が値ラベルを持っていないと、一部の機能が利用できなくなります。少なくとも 2 つの非欠損値で値ラベルが定義されている場合、値ラベルのない値を含むケースはすべて分析から除外されます。

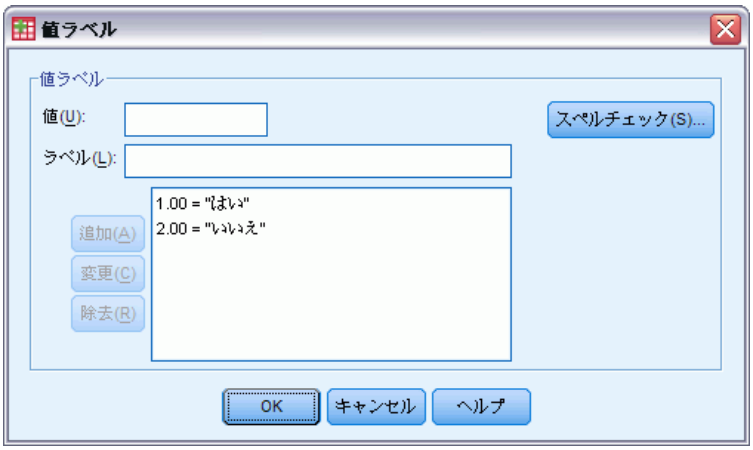
この例で使用するデータ ファイルは、初期状態では値ラベルが定義されていません。従属変数を名義変数として扱う場合、ツリー モデルではすべての非欠損値が分析に使用されます。この例で該当する値は 1、2、3 です。

では、従属変数の一部の値だけに値ラベルを定義した場合（すべての値には定義しない）、どのような現象が生じるでしょうか。

▶ ［データ エディタ］ウィンドウで、**［変数ビュー］**タブをクリックします。

▶ 変数「従属」の［値］セルをクリックします。

図 3-6
従属変数の値ラベルの定義



▶ まず、［値］に**「1」**、［値ラベル］に**「はい」**と入力し、［**追加**］をクリックします。

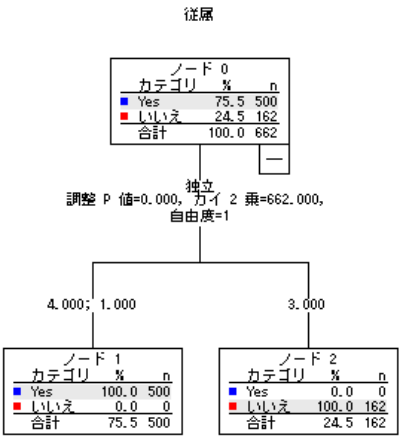
▶ 次に、［値］に**「2」**、［値ラベル］に**「いいえ」**と入力し、［**追加**］をクリックします。

▶ ［OK］をクリックします。

▶ ［ディシジョン ツリー］ダイアログ ボックスを再び開きます。ダイアログ ボックスでは、［従属変数］として、名義尺度で「従属」が選択されているはずです。

▶ ［OK］をクリックして手続きを再実行します。

図 3-7
一部が値ラベルを持つ名義従属変数のツリー



今度は、値ラベルが定義された 2 つの従属変数だけがツリー モデルに含まれています。データを詳細に理解していないとわかりにくいかもしれませんが、従属変数に値 3 を持つケースはすべて除外されています。

[値ラベル] をすべての値に割り当てる方法

有効なカテゴリ値が誤って分析から除外されないようにするには、[変数プロパティの定義] を使用して、データ内に存在するすべての従属変数値に値ラベルを割り当てます。

［変数プロパティの定義］ダイアログ ボックスに変数「従属」のデータ辞書情報を表示させると、値 3 を持つケースが 300 以上あるのに、その値に対して値ラベルが定義されていないことがわかります。

図 3-8
［変数プロパティの定義］ダイアログ ボックス（部分的に値ラベルを持つ変数）

章
4

ディシジョン ツリーを使用した信用リスクの評価

銀行では、銀行からローンを借り入れた顧客がそのローンを返済したのか、あるいは返済の履行を怠ったのかなど、顧客に関する履歴情報のデータベースを管理しています。ツリー モデルを使用すると、2 つのグループの顧客特性を分析し、ローン希望者が債務不履行を起こす可能性を予測するモデルを構築できます。

信用データは、tree_credit.sav に保存されています。 詳細は、 A 付録 サンプル ファイル in IBM SPSS Decision Trees 21 を参照してください。

モデルの作成

ディシジョン ツリー手続きには、ツリー モデルを作成する方法がいくつかあります。今回の例では、次のデフォルトの方法を使用します。

CHAID. カイ 2 乗自動反復検出。各ステップにおいて、CHAID は、従属変数と最も強い交互作用を持つ独立（予測）変数を選択します。従属変数の値に関する各予測変数のカテゴリーが著しく異ならない場合はそのカテゴリーは統合されます。

CHAID ツリー モデルの構築

▶ ディシジョン ツリー分析を実行するには、メニューから次の項目を選択します。

分析(A) > 分類 > ツリー...

© Copyright IBM Corporation 1989, 2012.

図 4-1
[ディシジョン ツリー] ダイアログ ボックス

- 従属変数として［信用度］を選択します。
- 残りの変数をすべて独立変数として選択します。（この手続きでは、最終的なモデルに対して有意な寄与率を持たない変数は、すべて自動的に除外されます。）

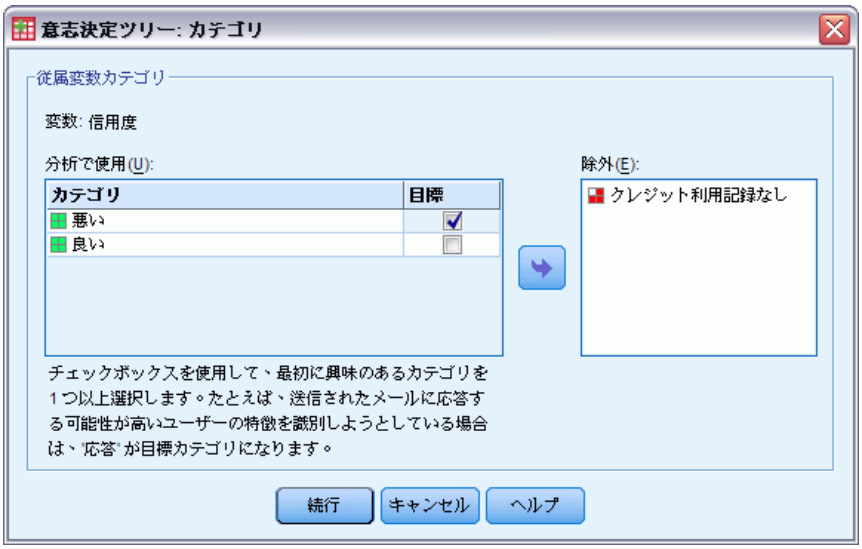
  この段階でも、手続きを実行して基本的なツリー モデルを作成することはできますが、今回は、追加出力を一部選択し、モデルの作成で使用する条件を微調整します。

目標カテゴリの選択

- 選択した従属変数のすぐ下にある［**カテゴリ**］ボタンをクリックします。

［カテゴリ］ダイアログ ボックスが開きます。このダイアログ ボックスでは、対象となる従属変数の目標カテゴリを指定できます。目標カテゴリはツリー モデルには影響しませんが、出力やオプションの中には、目標カテゴリを選択した場合にしか使用できないものもあります。

図 4-2
[カテゴリ] ダイアログ ボックス



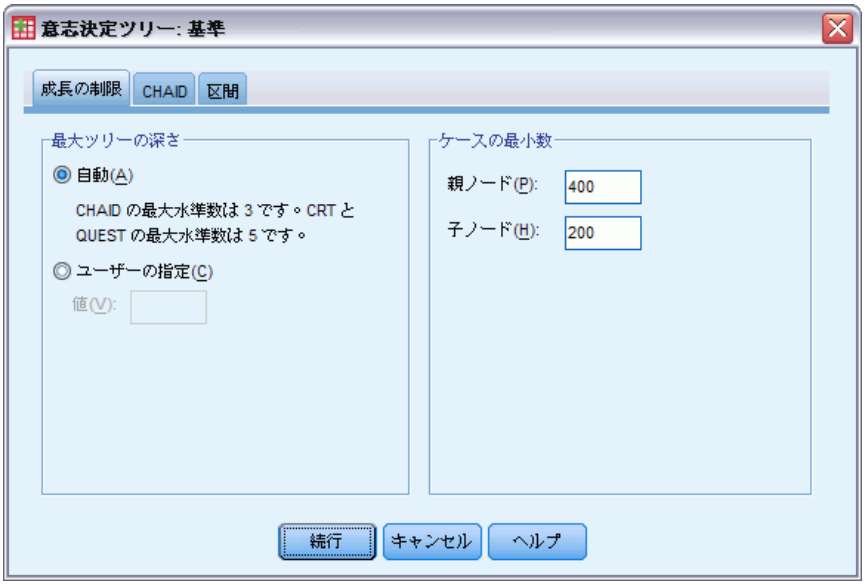
- カテゴリ「悪い」の［目標］チェック ボックスを選択（チェック）します。信用度の値が「悪い」の顧客（ローンの債務不履行を起こした顧客）は、対象目標カテゴリとして扱われます。
- ［続行］をクリックします。

ツリー成長基準の指定

今回の例では、ツリーをできる限り単純なものにしておきたいので、親ノードと子ノードの最小ケース数を増やしてツリーの成長を制限します。

- ［ディシジョン ツリー］メイン ダイアログ ボックスの［基準］をクリックします。

図 4-3
[基準] ダイアログ ボックスの [成長の制限] タブ



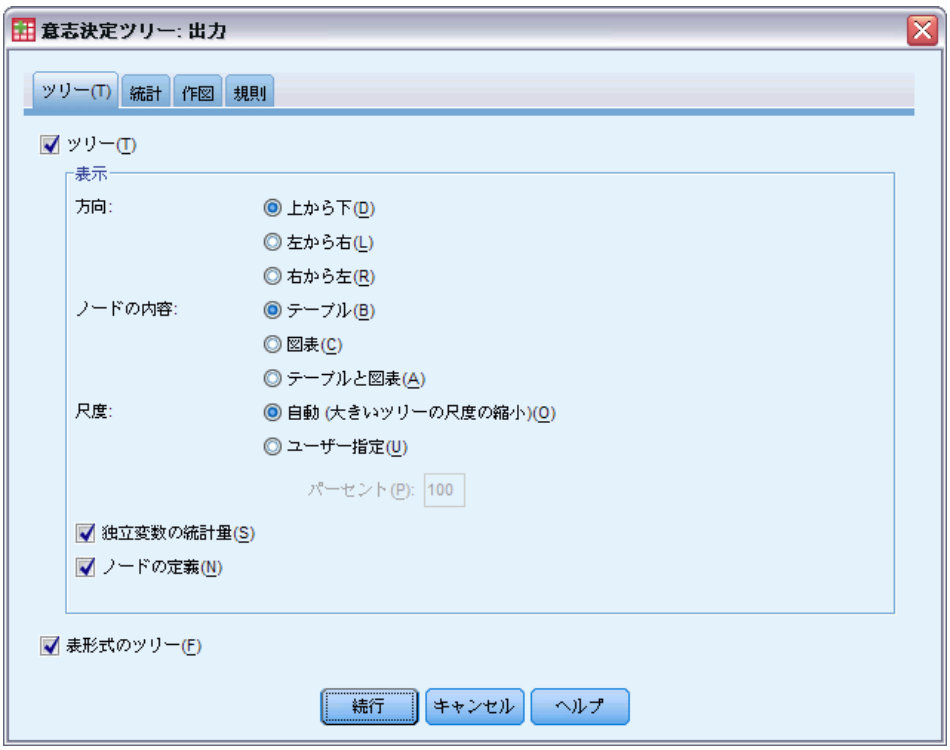
- ［ケースの最小数］グループで、［親ノード］に「400」と入力し、［子ノード］に「200」と入力します。
- ［続行］をクリックします。

追加出力の選択

- ［ディシジョン ツリー］メイン ダイアログ ボックスの［出力］をクリックします。

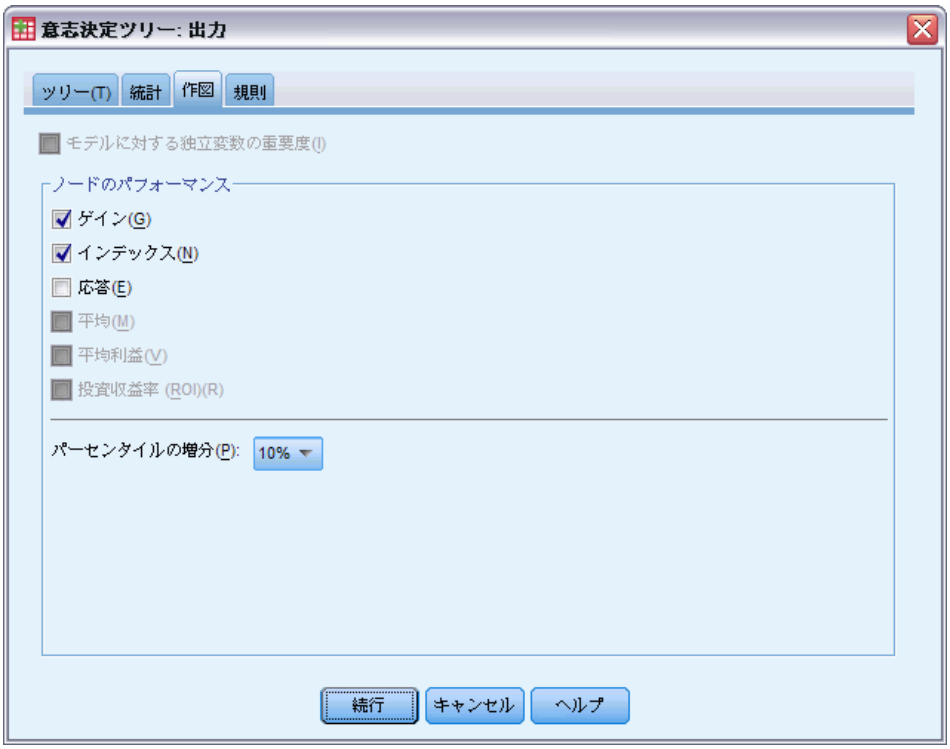
タブ付きダイアログ ボックスが開きます。このダイアログ ボックスでは、さまざまな種類の追加出力を選択できます。

図 4-4
[出力] ダイアログ ボックスの [ツリー] タブ



- ［ツリー］タブで、[表形式のツリー] を選択（チェック）します。
- 次に、[作図] タブをクリックします。

図 4-5
[出力] ダイアログ ボックスの [作図] タブ



▶ [**ゲイン**] と [**インデックス**] を選択（チェック）します。

注: これらの図表には、従属変数に目標カテゴリが必要となります。今回の例では、目標カテゴリを 1 つ以上指定するまで［作図］タブは使用できません。

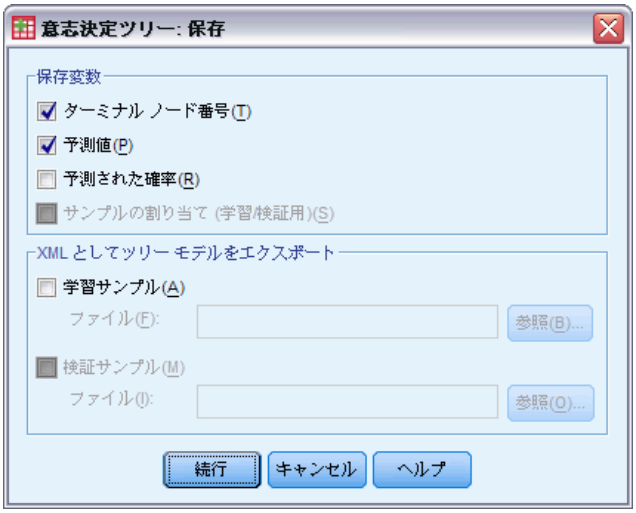
▶ [**続行**] をクリックします。

予測値の保存

モデルの予測に関する情報を持つ変数を保存できます。たとえば、ケースごとに信用度の予測値を保存して、その予測値を実際の信用度と比較できます。

▶ ［ディシジョン ツリー］メイン ダイアログ ボックスの [**保存**] をクリックします。

図 4-6
[保存] ダイアログ ボックス



- [ターミナル ノード番号]、[予測値]、および [予測された確率] を選択（チェック）します。
- [続行] をクリックします。
- [ディシジョン ツリー] メイン ダイアログ ボックスの [OK] をクリックして手続きを実行します。

モデルの評価

今回の例では、モデルの結果に次の項目が含まれます。

- モデルに関する情報を示すテーブル。
- ツリー図。
- モデルのパフォーマンスを示す図表。
- アクティブなデータ セットに追加されたモデルの予測変数。

モデルの要約表

図 4-7
モデルの要約(M)

指定	成長方法	CHAID
	従属変数	信用度
	独立変数	年齢, 所得レベル, クレジットカードの数, 教育, 車のローン
	検証	なし
	ツリーの最大の深さ	3
	親ノードの最小ケース	400
	子ノードの最小ケース	200
結果	含まれている独立変数	所得レベル, クレジットカードの数, 年齢
	ノードの数	10
	ターミナル ノードの数	6
	ツリーの深さ	3

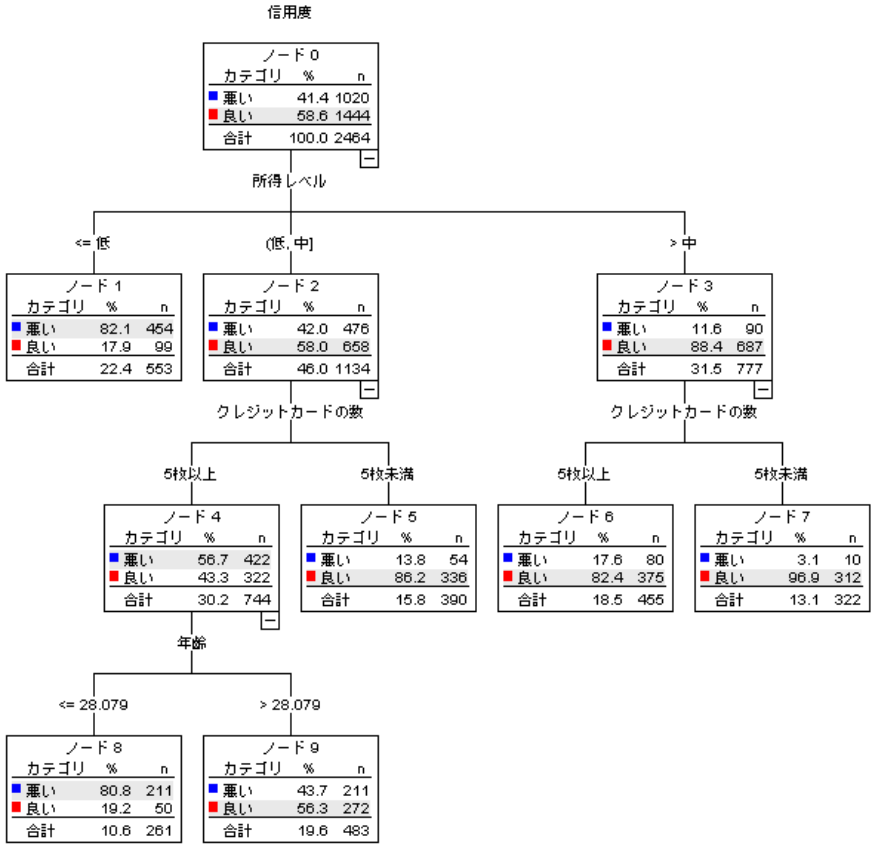
モデルの要約表には、モデルの構築で使用した指定や構築されたモデルに関する概略的な情報が表示されます。

- ［指定］には、分析で使用される変数など、ツリー モデルの作成で使用した設定に関する情報が表示されます。
- ［結果］には、ノードの合計数とターミナル ノードの数、ツリーの深度(ルート ノードより下のレベル数)、および最終モデルに含まれる独立変数に関する情報が表示されます。

5 つの独立変数を指定しましたが、最終モデルに含まれたのは 3 つだけです。変数「教育」、および変数「車のローン」の現在数は、モデルに対して有意な寄与率を持たず、最終モデルから自動的に除外されています。

ツリー図

図 4-8
信用度モデルのツリー図



ツリー図はツリー モデルを図示したものです。このツリー図は、次の内容を示します。

- CHAID 手法を使用すると、「信用度」の最適予測変数は「所得レベル」になります。
- 低所得カテゴリでは、「信用度」の有意予測変数は「所得レベル」だけです。このカテゴリに属する銀行顧客は、82% が債務不履行を起こしています。子ノードが下にないので、このカテゴリは **最終**ノードと見なされます。
- 中高所得カテゴリでは、「クレジットカードの数」が 2 番目の最適予測変数になります。
- 中所得の顧客の場合、クレジット カードを 5 枚以上所有している顧客に対しては、「年齢」が予測変数として追加されます。中所得でクレジット カードを 5 枚以上所有している顧客は、28 歳以下の場合、80%

超が「悪い」信用度になりますが、28 歳より年長の場合、「悪い」信用度になる人は半数に達しません。

ツリー エディタを使用すると、選択した枝の表示と非表示、色やフォントの変更、および選択したノードに基づくケースのサブグループの選択を行うことができます。 詳細は、 p. 78 ノード内のケースの選択 を参照してください。

ツリー表

図 4-9
信用度のツリー表

	悪い		良い		合計			
ノード	パーセント	度数	パーセント	度数	度数	パーセント	予測力テゴリ	親ノード
0	41.4%	1020	58.6%	1444	2464	100.0%	良い	
1	82.1%	454	17.9%	99	553	22.4%	悪い	0
2	42.0%	476	58.0%	658	1134	46.0%	良い	0
3	11.6%	90	88.4%	687	777	31.5%	良い	0
4	56.7%	422	43.3%	322	744	30.2%	悪い	2
5	13.8%	54	86.2%	336	390	15.8%	良い	2
6	17.6%	80	82.4%	375	455	18.5%	良い	3
7	3.1%	10	96.9%	312	322	13.1%	良い	3
8	80.8%	211	19.2%	50	261	10.6%	悪い	4
9	43.7%	211	56.3%	272	483	19.6%	良い	4

名前のとおり、ツリー表には、ツリー図に関する重要な情報の大部分が表形式で表示されます。この表には、ノードごとに次の内容が表示されます。

- 従属変数の各カテゴリに属するケースの数とパーセント。
- 従属変数の予測カテゴリ。今回の例の予測カテゴリは、対象ノードのケースの 50% 以上を含む「信用度」カテゴリです（可能な信用度が 2 つしかないので、どちらかが過半数になります)。
- ツリー内の各ノードの親ノード。低所得レベルのノードであるノード 1 が、どのノードの親ノードでもないことに注目してください。ノード 1 はターミナル ノードなので、子ノードはありません。

図 4-10
信用度のツリー表 (続き)

1 次独立変数				
変数	有意確率[a]	カイ 2 乗	自由度	分割値
所得レベル	.000	662.457	2	<= 低
所得レベル	.000	662.457	2	(低, 中]
所得レベル	.000	662.457	2	> 中
クレジットカードの数	.000	193.113	1	5枚以上
クレジットカードの数	.000	193.113	1	5枚未満
クレジットカードの数	.000	38.587	1	5枚以上
クレジットカードの数	.000	38.587	1	5枚未満
年齢	.000	95.299	1	<= 28.079205818990676
年齢	.000	95.299	1	> 28.079205818990676

- ノードの分割に使用される独立変数。
- カイ 2 乗値 (CHAID 手法を使用してツリーを作成したため)、自由度 (df)、および分割の有意水準 (Sig)。実際には、有意確率だけが関心の対象になる場合がほとんどですが、このモデルでは、すべての分割に対して有意確率は 0.0001 未満です。
- 対象ノードの独立変数の値。

注:順序独立変数とスケール独立変数の場合、ツリーおよびツリー表で範囲が表示されることがあります。範囲は、一般的な (値 1, 値 2] 形式で表され、基本的な意味は「値 1 より大きく、値 2 以下」です。 今回の例では、所得レベルとして可能な値は 3 つしかなく ([低]、[中]、および [高])、[(低, 中]] は単に [中] を意味します。同様に、>[> 中] は >[高] を意味します。

ノードのゲイン

図 4-11
ノードのゲイン

ノード	ノード		ゲイン			
	度数	パーセント	度数	パーセント	応答	インデックス
1	553	22.4%	454	44.5%	82.1%	198.3%
8	261	10.6%	211	20.7%	80.8%	195.3%
9	483	19.6%	211	20.7%	43.7%	105.5%
6	455	18.5%	80	7.8%	17.6%	42.5%
5	390	15.8%	54	5.3%	13.8%	33.4%
7	322	13.1%	10	1.0%	3.1%	7.5%

成長手法: CHAID
従属変数: 信用度

ノードのゲイン テーブルには、モデル内のターミナル ノードに関する概要が表示されます。

- このテーブルに表示されるのは、ツリーの成長が止まるノードであるターミナル ノードだけです。ターミナル ノードは対象モデルの最適分類予測を示すので、多くの場合、ターミナル ノードだけが関心の対象となります。
- ゲイン値には目標カテゴリの情報が含まれているので、このテーブルは 1 つ以上の目標カテゴリを指定した場合にだけ使用できます。今回の例では、目標カテゴリが 1 つしかないので、ノードのゲイン テーブルは 1 つしかありません。
- [ノード N] は各ターミナル ノード内のケースの数を示し、ノード パーセントは各ノード内のケースの合計数のパーセントを示します。
- [ゲイン N] は目標カテゴリに属する各ターミナル ノード内のケースの数を示し、[ゲイン パーセント] は目標カテゴリに属するケースの全体数に対する、目標カテゴリに属するケースのパーセントを示します。今回の例では、信用度が「悪い」ケースの数とパーセントになります。
- カテゴリ従属変数の場合、[応答] は、指定された目標カテゴリに属するノード内のケースのパーセントを示します。今回の例では、ツリー図でカテゴリ [悪い] に対して表示されるパーセントと同じになります。
- カテゴリ従属変数の場合、[インデックス] は、サンプル全体の応答のパーセントに対する、目標カテゴリの応答のパーセントの比率を示します。

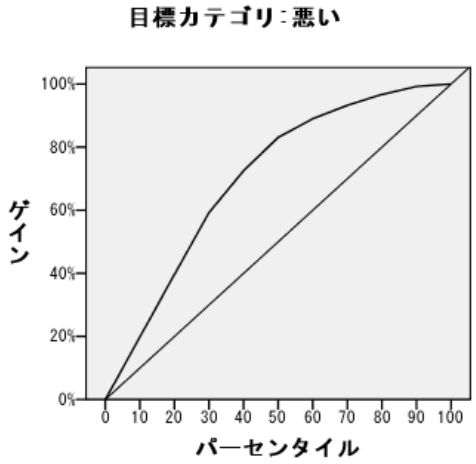
インデックス値

基本的に、インデックス値は、対象ノードの「観測された」目標カテゴリのパーセントが、「予測される」目標カテゴリのパーセントとどの程度異なるかを示す指標となります。 ルート ノード内の目標カテゴリのパーセントは、独立変数の効果をすべて考慮しない場合に予測されるパーセントを表します。

100% より大きなインデックス値は、目標カテゴリに属しているケースが、目標カテゴリ全体のパーセントよりも多いことを意味します。 逆に、100% 未満のインデックス値は、目標カテゴリに属しているケースが全体のパーセントよりも少ないことを意味します。

ゲイン グラフ

図 4-12
「悪い」信用度を目標カテゴリとしたゲイン グラフ

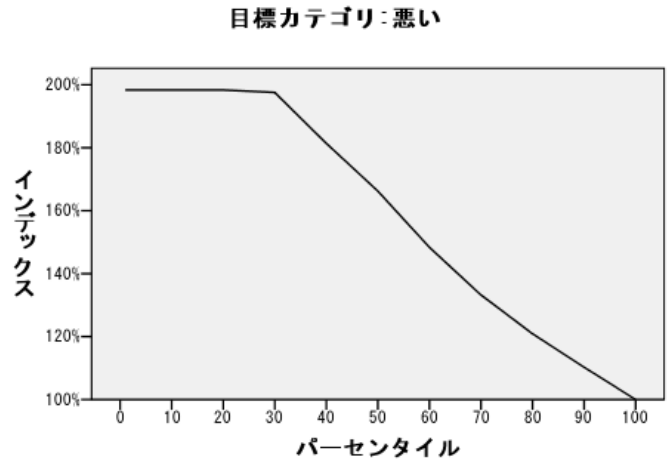


このゲイン グラフは、モデルが非常に適切であることを示します。

累積ゲイン グラフは、端から端まで見ていくと、必ず 0% から始まり 100% で終わります。適切なモデルの場合、ゲイン グラフは 100% に向けて急勾配で上昇し、その後、水平状態になります。情報を提供しないモデルの場合、対角線の参照線をたどるようになります。

インデックス グラフ

図 4-13
「悪い」信用度を目標カテゴリとしたインデックス グラフ



インデックス グラフも、今回のモデルが適切であることを示します。累積インデックス グラフは、100% 超から始まり、100% に達するまで徐々に低下します。

適切なモデルの場合、インデックス値は 100% をはるかに超える位置から始まり、高い位置で水平状態を保った後、100% に向けて急勾配で低下します。情報を提供しないモデルの場合、グラフの線全体が 100% 付近にとどまります。

リスク推定値と分類

図 4-14
誤差テーブルと分類テーブル

相対リスク

推定値	標準誤差
.205	.008

成長手法: CHAID
従属変数: 信用度

分類

	予測値		
観測	悪い	良い	正解の割合
悪い	665	355	65.2%
良い	149	1295	89.7%
全体のパーセント	33.0%	67.0%	79.5%

成長手法: CHAID
従属変数: 信用度

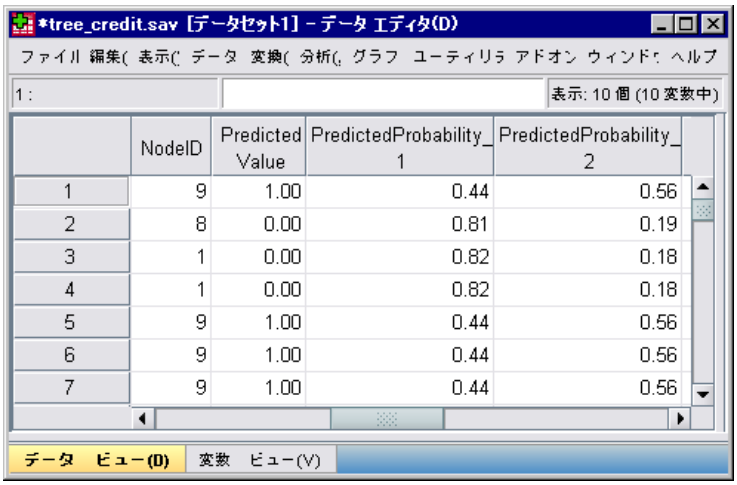
誤差テーブルと分類テーブルには、モデルの機能がどの程度優れているかについて簡単な評価が表示されます。

- 0.205 というリスク推定値は、モデルによって予測されたカテゴリ（信用度が「良い」のカテゴリまたは「悪い」のカテゴリ）が、ケースの 20.5% に対して間違っているということを示します。そのため、顧客を誤分類する「リスク」は約 21% になります。
- 分類テーブルに表示される結果は、リスク推定値と一致します。分類テーブルは、このモデルが顧客の約 79.5% を正しく分類することを示しています。

ただし、分類テーブルは、このモデルに潜在的な問題があることを明らかえにしています。信用格付けの悪い顧客については、65% だけの客に悪い信用格付けを予測しており、信用格付けの悪い顧客のうち 35% は、「良い」顧客として誤って分類されているということになります。

予測値

図 4-15
予測値と確率のための新変数



アクティブなデータ セットに、次の 4 つの変数が新しく作成されます。

NodeID。 各ケースのターミナル ノード番号。

PredictedValue。 各ケースの従属変数の予測値。従属変数は 0 = 「悪い」と 1 = 「良い」にコード化されているので、予測値が 0 であれば、そのケースの信用度が「悪い」と予測されることを示します。

PredictedProbability。 ケースが従属変数の各カテゴリに属する確率。従属変数には可能な値が 2 つしかないので、次の 2 つの変数が作成されます。

- **PredictedProbability_1。** ケースが信用格付けの悪いカテゴリに属する確率。
- **PredictedProbability_2。** ケースが信用格付けの良いカテゴリに属する確率。

予測確率とは、要するに、対象ケースが所属するターミナル ノードにおいて、ケースが従属変数のどのカテゴリに属しているかを示す比率です。たとえば、ノード 1 の場合、82% のケースがカテゴリ「悪い」に属し、18% がカテゴリ「良い」に属していて、予測確率はそれぞれ 0.82 と 0.18 になります。

カテゴリ従属変数の場合、予測値は、ターミナル ノード内のケースの属する比率が最も高いカテゴリになります。たとえば、ケース 1 の場合、ターミナル ノード内のケースの約 56% が「良い」信用度を持っているので、予測値は 1（信用度が「良い」）になります。 逆に、ケース 2 では、ターミナル ノード内のケースの約 81% が「悪い」信用度を持っているので、予測値は 0（信用度が「悪い」）になります。

ただし、コストを定義している場合は、予測カテゴリと予測確率の関係が複雑になることがあります。 詳細は、 p. 82 結果に対するコストの割り当て を参照してください。

モデルの改善

モデルの正分類率は、全体で約 80% です。この数値は、ターミナル ノードにも現れており、大部分のターミナル ノードにおいて、ノード内で強調表示される予測カテゴリは、80% 以上のケースで、実際に属するカテゴリと一致しています。

ただし、「悪い」信用度と「良い」信用度の間で、ケースがほとんど均等に分かれているターミナル ノードが 1 つあります。ノード 9 の予測信用度は「良い」ですが、実際に「良い」信用度を持つケースは、ノード内のケースの 56% しかありません。 つまり、ノード 9 内のケースの約半数 (44%) は、間違った予測カテゴリに分類されます。また、「悪い」信用リスクの識別に主な関心がある場合、このノードは適切に機能しません。

ノード内のケースの選択

ノード 9 内のケースを確認して、データから新たな役立つ情報が得られないか調べます。

- ビューア内のツリーをダブルクリックして、ツリー エディタを開きます。
- ノード 9 をクリックして選択します。(複数のノードを選択する場合は、Ctrl キーを押しながらクリックします。)
- [ツリー エディタ] メニューから次の項目を選択します。
  規則 > ケースのフィルタリング(F)...

図 4-16
[ケースのフィルタリング] ダイアログ ボックス

[ケースのフィルタリング] ダイアログ ボックスでは、フィルタ変数を作成し、その変数の値に基づいてケースをフィルタリングする設定を適用します。デフォルトのフィルタ変数名は、filter_$ です。

- 選択したノードのケースには、フィルタ変数の値 1 が返されます。
- その他のケースにはすべて値 0 が返され、フィルタの状態を変更するまで、以降の分析から除外されます。

今回の例では、ノード 9 に属さないケースはいったん分析から除外されます（ただし、削除はされません）。

▶ [OK] をクリックしてフィルタ変数を作成し、フィルタ条件を適用します。

図 4-17
データ エディタの分析から除外されたケース

データ エディタでは、分析から除外されたケースの行番号が斜線付きで示されます。ノード 9 に属さないケースが、分析から除外されています。ノード 9 内のケースは分析から除外されないので、以降の分析ではノード 9 内のケースだけが対象となります。

選択したケースの調査

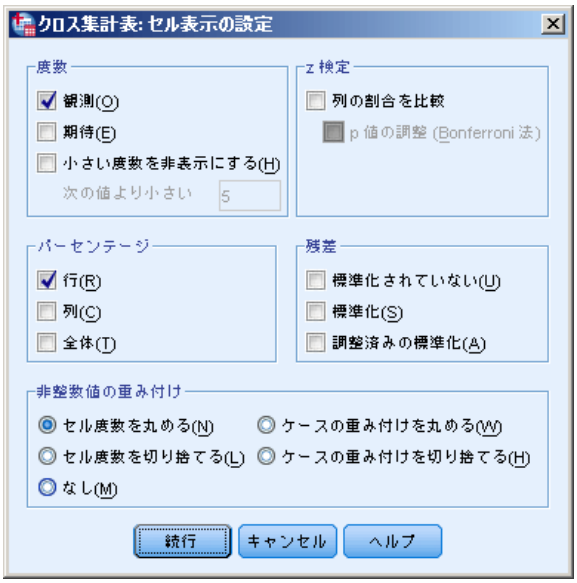
ノード 9 内のケースを調査するとき、まず、モデルで使用されなかった変数を試してみるのも 1 つの手段です。今回の例では、データ ファイル内のすべての変数を分析の対象としましたが、その中の「教育」と「車のローン」の 2 つの変数は最終モデルに含まれませんでした。 分類ツリー手続きが 2 つの変数を最終モデルから除外したことには正当な理由があるはずです。十分な情報は得られないかも知れませんが、調べてみましょう。

▶ メニューから次の項目を選択します。
分析(A) > 記述統計 > クロス集計表...

図 4-18
[クロス集計表] ダイアログ ボックス

- 「信用度」を［行変数］として選択します。
- 「教育」と「車のローン」を［列変数］として選択します。
- **［セル］**をクリックします。

図 4-19
[クロス集計表: セル表示の設定] ダイアログ ボックス



- ［パーセンテージ］グループで、[行] を選択（チェック）します。
- [続行] をクリックし、［クロス集計表］メイン ダイアログ ボックスの [OK] をクリックして手続きを実行します。

クロス集計表を調べると、モデルに含まれなかった 2 つの変数の場合、信用度カテゴリ「良い」に属するケースとカテゴリ「悪い」に属するケースの間でそれほど大きな差がないことがわかります。

図 4-20
選択したノード内のケースのクロス集計表

信用度 と 教育 のクロス表

			教育		
			高校	大学	合計
信用度	悪い	度数	110	101	211
		信用度 の %	52.1%	47.9%	100.0%
	良い	度数	128	144	272
		信用度 の %	47.1%	52.9%	100.0%
合計		度数	238	245	483
		信用度 の %	49.3%	50.7%	100.0%

信用度 と 車のローン のクロス表

			車のローン		
			1以下	2以上	合計
信用度	悪い	度数	18	193	211
		信用度 の %	8.5%	91.5%	100.0%
	良い	度数	39	233	272
		信用度 の %	14.3%	85.7%	100.0%
合計		度数	57	426	483
		信用度 の %	11.8%	88.2%	100.0%

- 「教育」では、「悪い」信用度を持つケースで、高卒が半数を少し上回り、「良い」信用度を持つケースで、大卒が半数を少し上回っています。しかし、この差異は統計的に有意ではありません。
- 「車のローン」では、車のローンが「1 以下」のとき、「良い」信用度を持つケースのパーセントは、「悪い」信用度を持つケースのパーセントより高くなっていますが、どちらのグループでも、ほとんどのケースが「2 以上」の自動車ローンを組んでいます。

これで、2 つの変数が最終モデルに含まれなかった理由がだいたいわかりました。しかし、ノード 9 の予測を改善する方法は残念ながらまったくわかっていません。分析用に指定されなかった変数が他にある場合は、先に進む前にその変数をいくつか調べてみるのもよいでしょう。

結果に対するコストの割り当て

すでに述べたように、ノード 9 内でケースのほぼ半数ずつが、2 つの信用度カテゴリに分類されるということも問題ですが、「悪い」信用リスクを正しく識別するモデルを構築することが主な目的の場合は、ノード 9 の予測カテゴリが「良い」になっていることが問題になります。 ノード 9 のパフォーマンスを向上できない可能性はありますが、モデルを改善して「悪い」信用度を持つケースの正分類率を上げる余地はまだあり

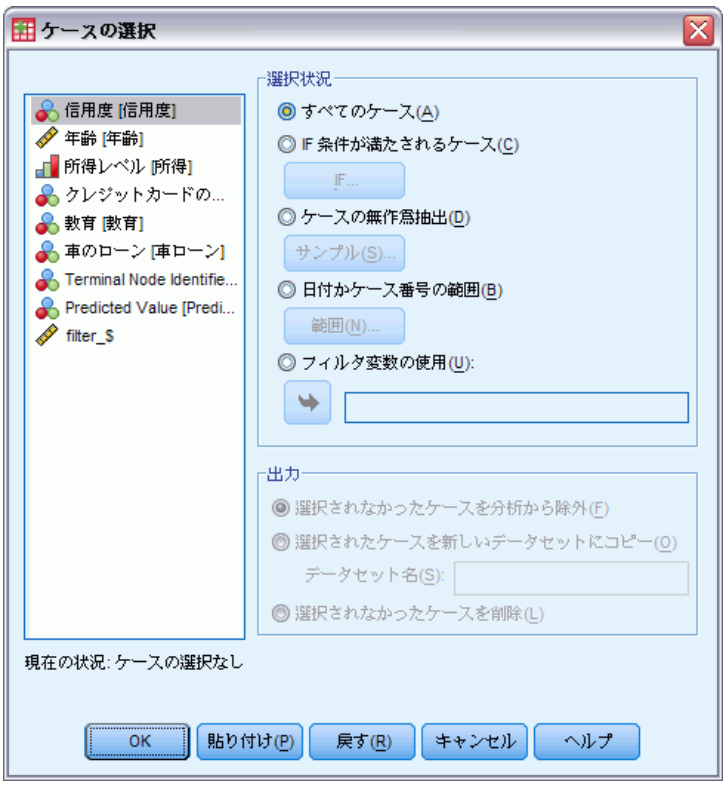
ます。ただし、この方法では、「良い」信用度を持つケースを誤分類する比率が高くなります。

まず、ケースのフィルタリングを無効にして、再びすべてのケースを分析の対象に戻す必要があります。

▶ メニューから次の項目を選択します。
データ > ケースの選択(S)...

▶ [ケースの選択] ダイアログ ボックスで、[**全てのケース**] を選択し、[OK] をクリックします。

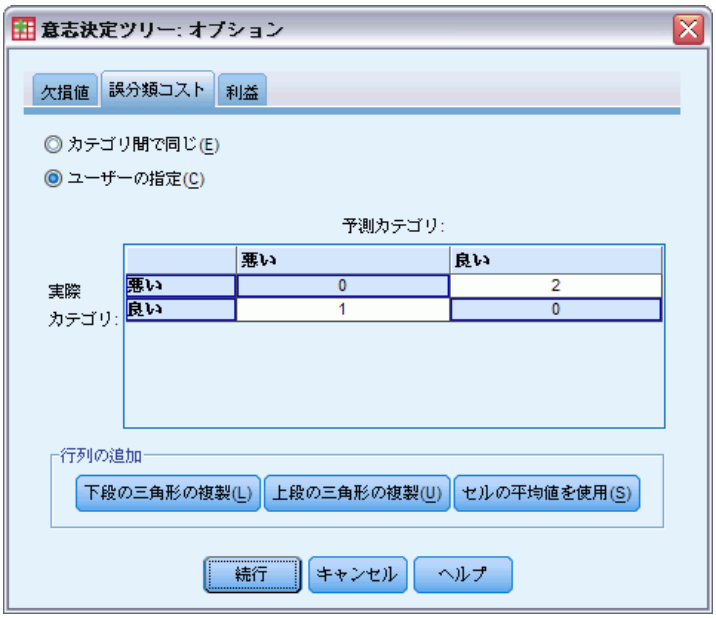
図 4-21
[ケースの選択] ダイアログ ボックス



▶ [ディシジョン ツリー] ダイアログ ボックスを再び開いて、[**オプション**] をクリックします。

▶ [誤分類コスト] タブをクリックします。

図 4-22
[オプション] ダイアログ ボックスの [誤分類コスト] タブ

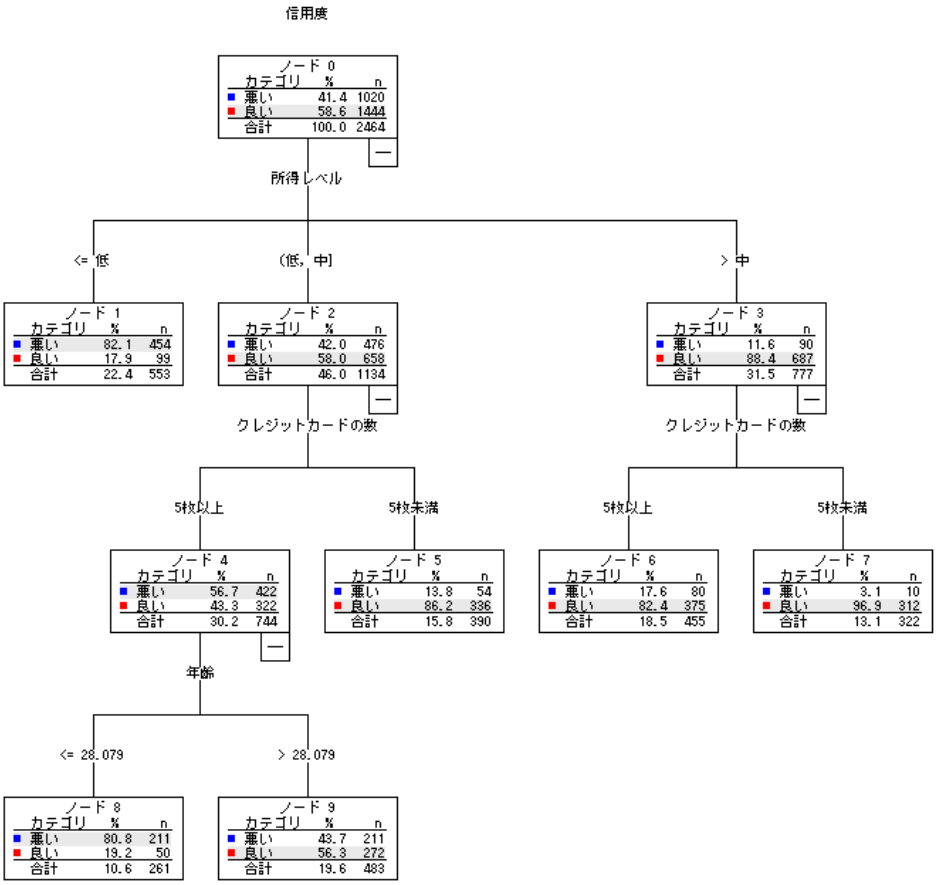


▶ [ユーザー指定] を選択し、[実際のカテゴリ] が [悪い] で [予測カテゴリ] が [良い] の箇所に、値 2 を入力します。

上記の操作は、「悪い」信用リスクを誤って「良い」信用リスクに分類した場合の「コスト」が、「良い」信用リスクを誤って「悪い」信用リスクに分類した場合の「コスト」の 2 倍になると指定したことになります。

▶ [続行] をクリックし、メイン ダイアログ ボックスの [OK] をクリックして、手続きを実行します。

図 4-23
調整済みのコスト値を含むツリー モデル



手続きによって生成されたツリーは、一見しただけでは、元のツリーとほとんど同じに見えます。しかし、詳しく調べると、各ノードでのケースの分布は変化していませんが、一部の予測カテゴリが変化したことがわかります。

ターミナル ノードについて言えば、予測カテゴリは、ノード 9 以外のすべてのノードで同じ状態です。ノード 9 では、半数を若干上回る数のケースがカテゴリ 「良い」に属しているのに、予測カテゴリは 「悪い」になりました。

「悪い」信用度を「良い」信用度に誤分類するとコストが高くなると指定したので、ケースが 2 つのカテゴリにほぼ均等に分布するノードでは、ケースの過半数がカテゴリ 「良い」に属する場合でも、予測カテゴリは 「悪い」になっています。

予測カテゴリでの変化の効果は、分類テーブルに現われています。

図 4-24
調整済みのコストに基づく誤差テーブルと分類テーブル

相対リスク

推定値	標準誤差
.288	.011

成長方法 CHAID
従属変数 信用度

分類

観測	予測値		
	悪い	良い	正確なパーセント
悪い	876	144	85.9%
良い	421	1023	70.8%
全体のパーセント	52.6%	47.4%	77.1%

成長方法 CHAID
従属変数 信用度

- 「悪い」信用リスクでは、正分類率が調整前のわずか 65% という状態から約 86% へと改善しています。
- 一方、「良い」信用リスクでは、正分類率は 90% から 71% まで低下し、全体の正分類率も 79.5% から 77.1% まで低下しています。

また、リスク推定値と全体の正分類率が一致しなくなったことに注意してください。全体の正分類率が 77.1% であれば、リスク推定値は 0.229 になると予想できます。 今回の例では、「悪い」信用度を持つケースの誤分類コストが増加したことで、リスクの値が膨張し、解釈が難しくなっています。

要約表

ツリー モデルを使用すると、ケースを特性に応じたグループに分類できます。今回の例では、銀行の信用記録が良い顧客と悪い顧客に関連する特性を利用しました。特定の予測結果が他の起こりうる結果よりも重要である場合、モデルを改善して、目的とする結果の誤分類コストを高めることができます。ただし、ある結果の誤分類率を下げると、別の結果の誤分類率が上がります。

章 5

得点モデルの作成

ディシジョン ツリー手続きの最も強力で役立つ機能の 1 つに、他のデータ ファイルに適用できるモデルを作成し、結果を予測する機能があります。たとえば、人口統計情報および車の購入価格に関する情報を持つデータ ファイルに基づいて、人口統計上の特性が似ている人が新車購入にいくら支払う可能性が高いかを予測する際に使用できるモデルを作成できます。モデルを作成すれば、人口統計情報はあるがこれまでの車の購入に関する情報のない他のデータ ファイルに、そのモデルを適用できます。

この例では、tree_car.sav を使用します。 詳細は、 A 付録 サンプル ファイル in IBM SPSS Decision Trees 21 を参照してください。

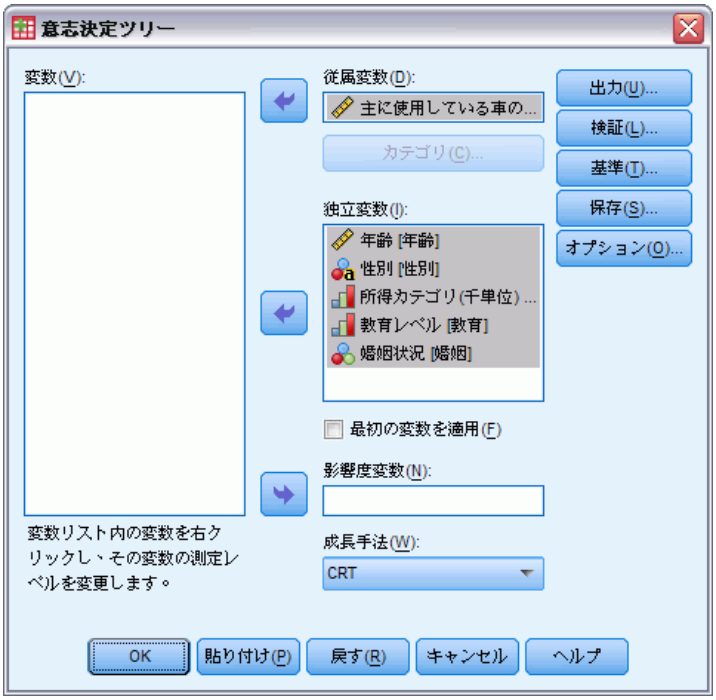
モデルの作成

- ディシジョン ツリー分析を実行するには、メニューから次の項目を選択します。

  分析(A) > 分類 > ツリー...

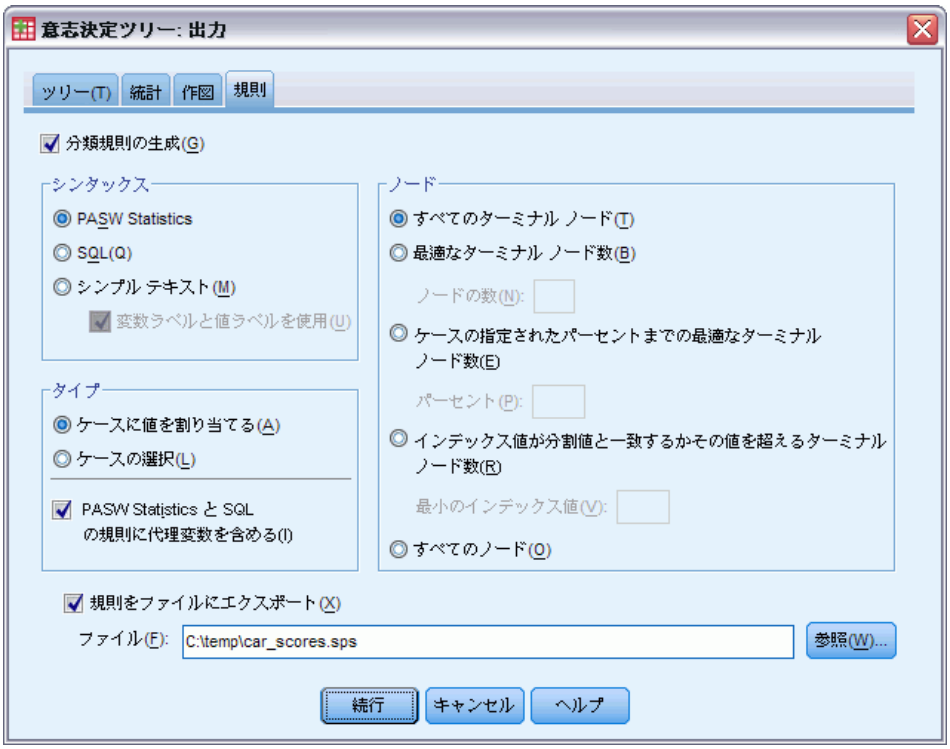
© Copyright IBM Corporation 1989, 2012.

図 5-1
[ディシジョン ツリー] ダイアログ ボックス



- 従属変数として「主に使用している車の価格」を選択します。
- 残りの変数をすべて独立変数として選択します。(この手続きでは、最終的なモデルに対して有意な寄与率を持たない変数は、すべて自動的に除外されます。)
- [成長手法] として [CRT] を選択します。
- [出力] をクリックします。

図 5-2
[分類ツリー: 出力] ダイアログ ボックスの [規則] タブ



- ［規則］タブをクリックします。
- ［分類規則の生成］を選択（チェック）します。
- ［シンタックス］として［IBM® SPSS® Statistics］を選択します。
- ［型］として［ケースに値を割り当てる］を選択します。
- ［規則をファイルにエクスポート］を選択（チェック）して、ファイル名およびディレクトリの場所を入力します。

  ファイル名とディレクトリの場所は後で必要になるため、覚えておくか、書き留めておいてください。ディレクトリ パスを指定しないと、ファイルの保存場所がわからなくなる可能性があります。［参照］ボタンを使用すれば、特定の（有効な）ディレクトリの場所を指定できます。
- ［続行］をクリックして、［OK］をクリックし、手続きを実行してツリー モデルを作成します。

モデルの評価

他のデータ ファイルにモデルを適用する前に、モデルの作成に使用した元のデータでモデルが適切なものか確認します。

モデルの要約

図 5-3
モデルの要約表

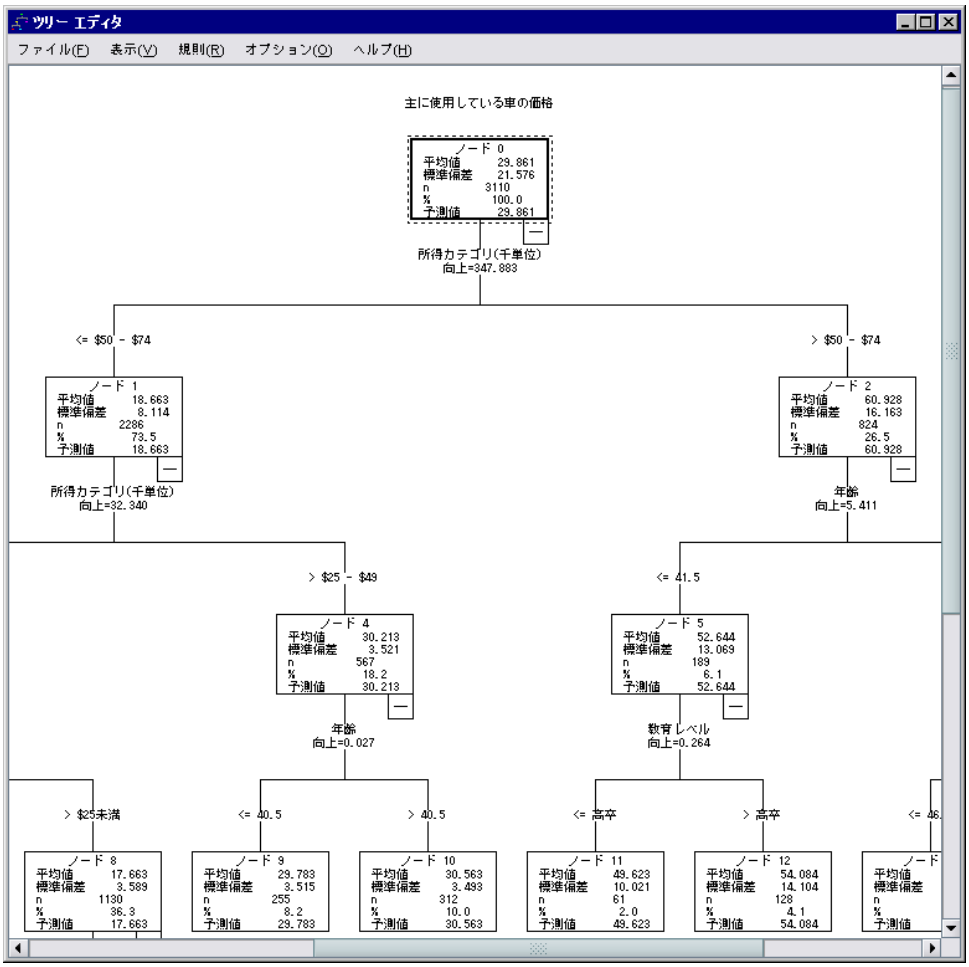
指定	成長方法	CRT
	従属変数	主に使用している車の価格
	独立変数	年齢, 性別, 所得カテゴリ(千単位), 教育レベル, 婚姻状況
	検証	NONE
	ツリーの最大の深さ	5
	親ノードの最小ケース	100
	子ノードの最小ケース	50
結果	含まれている独立変数	所得カテゴリ(千単位), 年齢, 教育レベル
	ノードの数	29
	ターミナルノードの数	15
	奥行き	5

モデルの要約表は、選択した独立変数のうち 3 つの変数（「所得」、「年齢」、および「教育」）だけが有意な寄与率を持ち、最終モデルに含まれていることを示しています。これは、このモデルを他のデータ ファイルに適用する上で、確認すべき重要な情報です。このモデルで使用された独立変数が、モデルを適用するデータ ファイルに存在する必要があるためです。

また、要約表は、ツリー モデルに 29 個のノードと 15 個のターミナル ノードがあり、このモデルがそれほど単純なモデルではないことを示しています。記述や説明のしやすい単純なモデルではなく、実用的に適用できる信頼性のあるモデルが必要な場合、モデルが複雑であっても問題ではないことがあります。もちろん、実用的な面から考えて、あまり多くの独立（予測）変数に依存しないモデルが必要になることもあります。今回の例では、最終モデルに含まれているのは 3 つの独立変数だけなので、問題ではありません。

ツリー モデル図

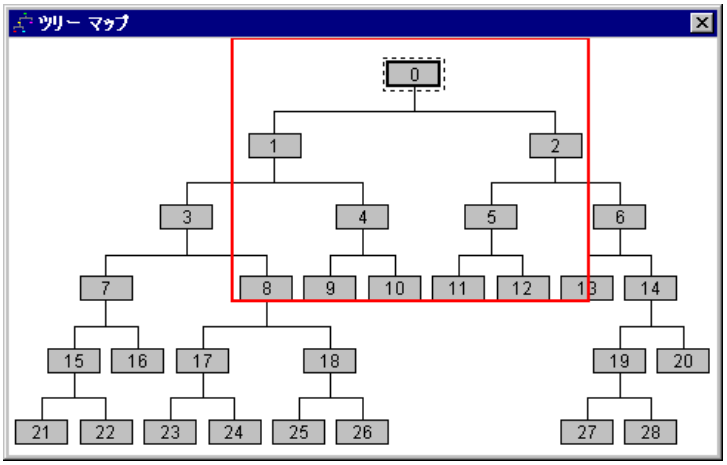
図 5-4
ツリー エディタのツリー モデル図



ツリー モデル図には、非常に多くのノードがあるので、ノードの内容情報を読み取れるサイズで、一度にモデル全体を表示することは困難な場合があります。ツリー マップを使用して、ツリー全体を表示できます。

- ビューア内のツリーをダブルクリックして、ツリー エディタを開きます。
- ［ツリー エディタ］メニューから次の項目を選択します。
  表示 > ツリー マップ

図 5-5
ツリー マップ



- ツリー マップでは、ツリー全体が表示されます。ツリー マップのウィンドウのサイズは変更できます。ツリーを表示する図は、ウィンドウのサイズに合わせて拡大または縮小されます。
- ツリー マップ内で強調表示されている領域は、ツリー エディタで現在表示されているツリーの領域です。
- ツリー マップを使用して、ツリーを移動し、ノードを選択できます。

詳細は、 2 章 p. 44 ツリー マップ を参照してください。

スケール従属変数の場合、各ノードには、従属変数の平均値と標準偏差が表示されます。ノード 0 には、車の購入価格の総平均値が約 29.9（1,000 ドル単位）で、標準偏差が約 21.6 と表示されます。

- ノード 1 は、収入が 75（1,000 ドル単位）未満のケースを示しており、このノードでは、車の平均価格は 18.7 ドルにしかなりません。
- 対照的に、ノード 2 は、収入が 75 以上のケースを表示しており、このノードでは、車の平均価格は 60.9 です。

ツリーをさらに詳しく調べると、「年齢」および「教育」も車の購入価格との関係を示していることがわかります。ただし、今回は、モデルを実用的に適用することが主な目的なので、モデルの個々の成分を詳細に調査することはしません。

リスク推定値

図 5-6
誤差テーブル

相対リスク

推定値	標準誤差
68.485	2.985

成長手法: CRT
従属変数: 主に使用している車の価格

これまでに調査した結果では、今回のモデルが特に適切なモデルかどうかわかりません。モデルのパフォーマンスを測る 1 つの指標として、リスク推定値があります。スケール従属変数の場合、リスク推定値はノード内分散の測定値です。この値自体ではそれほど多くのことがわからないこともあります。分散が低いことは、モデルがより適切であることを示しますが、分散は測定の単位と関係があります。たとえば、価格を千単位ではなく一単位で記録した場合、リスク推定値は千倍の大きさになります。

スケール従属変数に関するリスク推定値の解釈を意味のあるものにするには、いくつかの作業が必要になります。

- 全分散は、ノード内（エラー）分散とノード間（説明された）分散の和に等しくなります。
- ノード内分散は、リスク推定値の 68.485 です。
- 全分散は、独立変数を考慮せずに測った従属変数の分散で、ルート ノードでの分散になります。
- ルート ノードでの標準偏差は 21.576 と示されており、全分散はその値を 2 乗した 465.524 になります。
- エラーによる分散（説明されない分散）の比率は 68.485 / 465.524 = 0.147 です。
- モデルによって説明される分散の比率は 1 ～ 0.147 = 0.853 (85.3%) です。この比率は、モデルが非常に優れたモデルだということを示しています。（これは、カテゴリ従属変数の正分類率全体に対する解釈と同様の解釈になります。）

別のデータ ファイルへのモデルの適用

モデルが十分適していると判断した場合、同じ変数「年齢」、「所得」、および「教育」を含む他のデータ ファイルにモデルを適用し、データ ファイル内のケースごとに車の予測購入価格を示す新しい変数を作成できます。 このプロセスは、一般的に、**得点（スコアリング）**と呼ばれます。

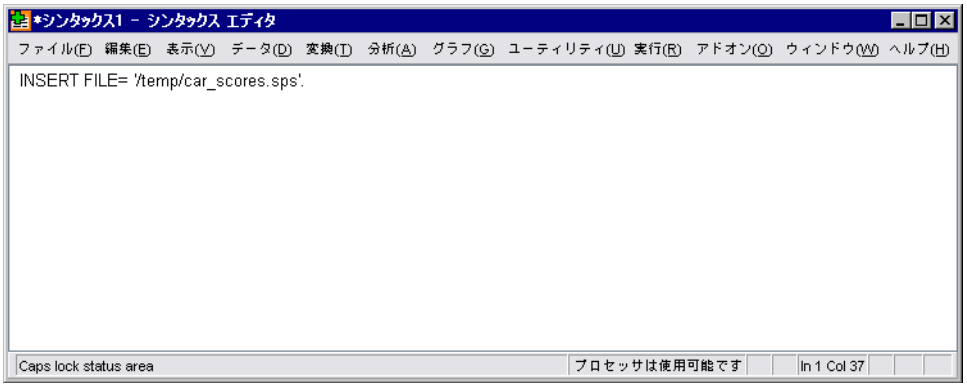
モデルを作成したとき、ケースに値を割り当てるための「規則」をコマンド シンタックスの形式でテキスト ファイルに保存するように指定しました。ここでは、保存したファイルのコマンドを使用して、別のデータ ファイルで得点を生成します。

- データ ファイル tree_score_car.sav を開きます。詳細は、 A 付録 サンプル ファイル in IBM SPSS Decision Trees 21 を参照してください。
- 次に、メニューから次の項目を選択します。
  ファイル(F) > 新規作成(N) > シンタックス
- コマンド シンタックス ウィンドウで、次のコマンドを入力します。

```
INSERT FILE=
 '/temp/car_scores.sps'.
```

別のファイル名や場所を使用した場合は、それにあわせて変更してください。

図 5-7
コマンド ファイルを実行する INSERT コマンドが入力されたシンタックス ウィンドウ



INSERT コマンドは、指定したファイル内でコマンドを実行します。ここでは、モデル作成時に生成した「規則」ファイルです。

- コマンド シンタックス ウィンドウのメニューから、次の項目を選択します。
  実行(R) > すべて

図 5-8
データ ファイルに追加される予測値

	所得	教育	婚姻	nod_001	pre_001	var	var
1	3.00	1	1	10.0000	30.4924		
2	4.00	1	0	14.0000	67.5392		
3	2.00	3	1	23.0000	15.6662		
4	2.00	4	1	24.0000	17.0631		
5	1.00	2	0	22.0000	10.1038		
6	3.00	2	0	9.0000	29.5262		
7	1.00	1	0	22.0000	10.1038		
8	4.00	3	1	11.0000	46.7857		
9	3.00	3	1	10.0000	30.4924		
10	4.00	4	1	20.0000	63.9424		

この操作により、データ ファイルに新しく 2 つの変数が追加されます。

- 「nod_001」には、ケースごとにモデルが予測したターミナル ノード番号が入っています。
- 「pre_001」には、ケースごとの車の購入価格の予測値が入っています。

ターミナル ノードに値を割り当てる規則を要求したので、取りうる予測値の数は、ターミナル ノードの数と同じになります。今回の場合は、15 個になります。たとえば、予測されたノード番号が 10 のケースでは、車の購入価格がすべて 30.56 という同じ予測価格になります。 これは、偶然ではなく、元のモデルのターミナル ノード 10 に報告された平均値です。

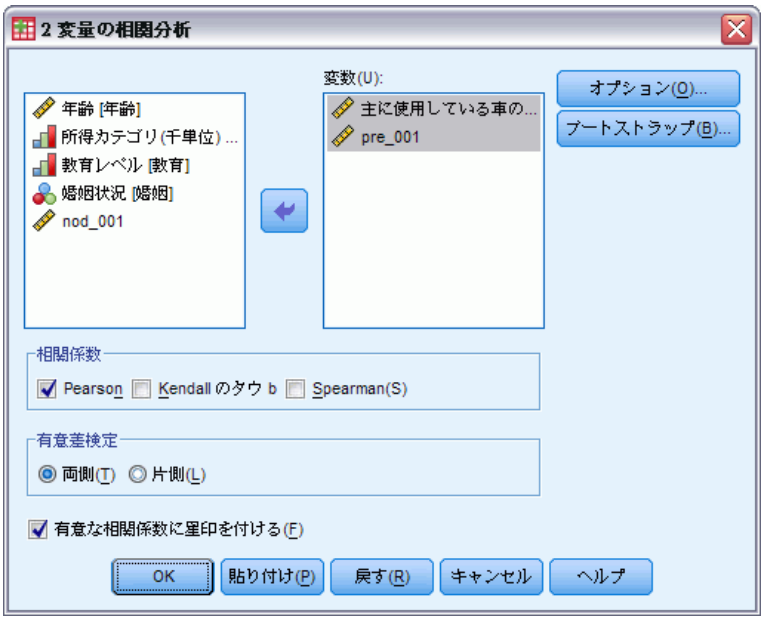
一般的に、従属変数の値が不明なデータにモデルを適用しますが、今回の例では、実はモデルを適用したデータ ファイルにその情報が含まれています。そのため、モデルの予測と実際の値を比較できます。

▶ メニューから次の項目を選択します。

分析(A) > 相関 > 2 変量...

▶ 「主に使用している車の価格」および「pre_001」を選択します。

図 5-9
[2 変量の相関分析] ダイアログ ボックス



▶ [OK] をクリックして手続きを実行します。

図 5-10
車の実際価格と予測価格の相関

		主に使用している車の価格	pre_001
主に使用している車の価格	Pearson の相関係数	1	.919**
	有意確率 (両側)		.000
	N	3290	3290
pre_001	Pearson の相関係数	.919**	1
	有意確率 (両側)	.000	
	N	3290	3290

**. 相関係数は 1% 水準で有意 (両側) です。

0.92 という相関は、車の実際価格と予測価格が非常に高い正の相関関係にあることを示します。この相関は、モデルが適切に機能していることを示します。

要約表

ディシジョン ツリー手続きを使用してモデルを作成すれば、それを他のデータ ファイルに適用して結果を予測できます。対象となるデータ ファイルには、最終モデルに含まれる独立変数と同じ名前の変数が必要です。こ

の変数は、同じ測定基準で測定され、ユーザー定義欠損値がある場合は同じ値でなければなりません。ただし、最終モデルから除外された従属変数と独立変数は、対象となるデータ ファイルに存在する必要はありません。

章 6

ツリー モデル内の欠損値

成長手法によって、独立（予測）変数の欠損値を処理する方法が異なります。

- CHAID と Exhaustive CHAID では、各独立変数のシステム欠損値とユーザー欠損値がすべて 1 つのカテゴリとして扱われます。スケール独立変数と順序独立変数の場合、そのカテゴリは、成長基準に応じて、対象となる独立変数の他のカテゴリに順次結合していくことがあります。
- CRT と QUEST では、独立（予測）変数に**代理変数**が使用されます。独立変数の値が欠損している場合、その変数と強く関連する別の独立変数が分類に使用されます。このような代替予測変数を代理変数と呼びます。

今回の例では、モデル内で使用する独立変数に欠損値がある場合の CHAID と CRT の違いを示します。

この例では、データ ファイル tree_missing_data.sav を使用します。詳細は、 A 付録 サンプル ファイル in IBM SPSS Decision Trees 21 を参照してください。

注:名義独立変数と名義従属変数の場合、**ユーザー欠損値**を有効値として扱うことも選択できます。その場合、ユーザー欠損値は他の非欠損値と同様に扱われます。 詳細は、 1 章 p.24 欠損値 を参照してください。

© Copyright IBM Corporation 1989, 2012.

CHAID を使用した場合の欠損値

図 6-1
欠損値のある信用データ

	信用度	年齢	所得	クレジットカード	教育	車ロー
1	0.00	36.22	2.00	.	2.00	
2	0.00	21.99	2.00	.	2.00	
3	0.00	29.17	.	2.00	1.00	
4	0.00	32.75	.	2.00	2.00	
5	0.00	36.77	2.00	.	2.00	
6	0.00	39.32	2.00	2.00	2.00	
7	0.00	31.70	2.00	2.00	2.00	
8	0.00	34.72	.	2.00	1.00	
9	0.00	31.53	1.00	2.00	1.00	
10	0.00	24.78	2.00	.	2.00	

信用リスクの例（詳細については4 章を参照してください）と同様に、この例では信用リスクを「良い」と「悪い」に分類するモデルを構築します。一番大きな違いは、今回のデータ ファイルでは、モデル内で使用する独立変数の一部に欠損値が含まれていることです。

- ディシジョン ツリー分析を実行するには、メニューから次の項目を選択します。

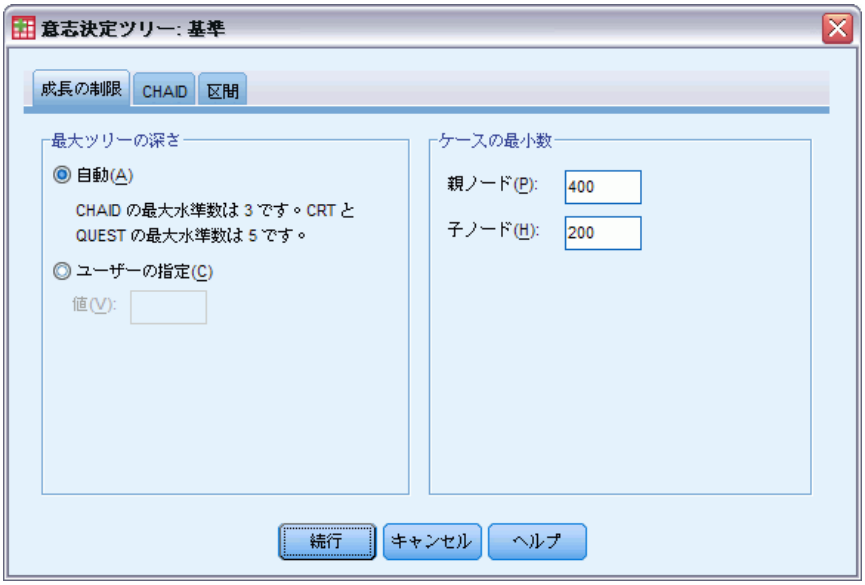
  分析(A) > 分類 > ツリー...

図 6-2
[ディシジョン ツリー] ダイアログ ボックス

- 従属変数として［信用度］を選択します。
- 残りの変数をすべて［独立変数］として選択します。（この手続きでは、最終的なモデルに対して有意な寄与率を持たない変数は、すべて自動的に除外されます。）
- ［成長手法］として［CHAID］を選択します。

  今回の例では、ツリーをできる限り単純なものにしておきたいので、親ノードと子ノードの最小ケース数を増やしてツリーの成長を制限します。
- ［ディシジョン ツリー］メイン ダイアログ ボックスの［基準］をクリックします。

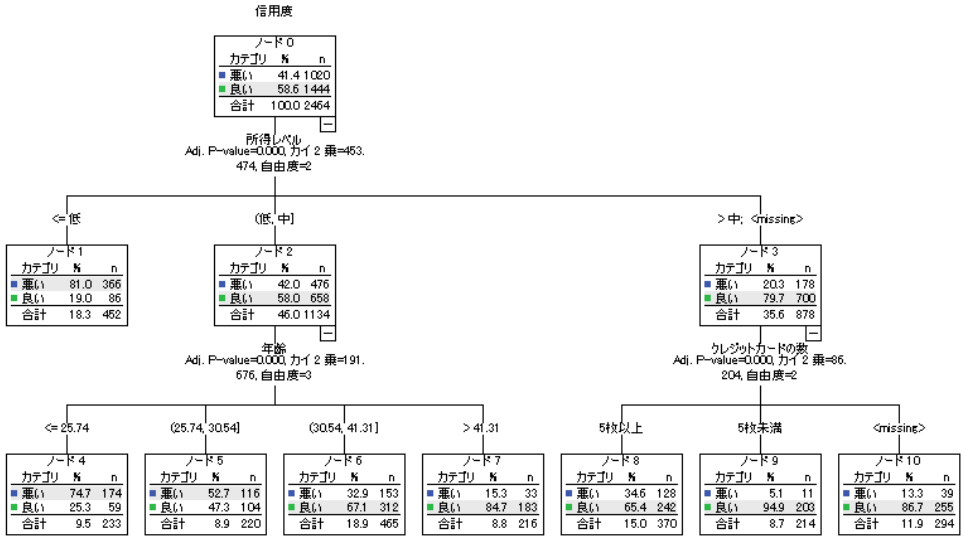
図 6-3
[基準] ダイアログ ボックスの [成長の制限] タブ



- ［ケースの最小数］の［親ノード］に「400」と入力し、［子ノード］に「200」と入力します。
- [続行] をクリックし、[OK] をクリックして手続きを実行します。

CHAID の結果

図 6-4
独立変数に欠損値のある CHAID ツリー



ノード 3 の「所得レベル」の値に、>[< 中; <欠損値>]> と表示されます。これは、ノード 3 に高収入カテゴリのケースと「所得レベル」に欠損値のあるすべてのケースが含まれていることを意味します。

ターミナル ノード 10 には、「クレジットカードの数」に欠損値のあるケースがあります。「良い」信用リスクの識別という点からすると、ノード 10 は確かに 2 番目に良いターミナル ノードとなっています。しかし、このモデルを「良い」信用リスクの予測に使用するときには問題となる可能性があります。対象ケースのクレジット カード数がわからないというだけの理由で、そのケースの信用度を「良い」と予測するモデルは役に立ちません。そのようなケースは年収レベルに関する情報が欠損していることもあります。

図 6-5
CHAID モデルの誤差テーブルと分類テーブル

相対リスク

推定値	標準誤差
.249	.009

成長方法: CHAID
従属変数: 信用度

分類

観測	予測値		
	悪い	良い	正確なパーセント
悪い	656	364	64.3%
良い	249	1195	82.8%
全体のパーセント	36.7%	63.3%	75.1%

成長方法: CHAID
従属変数: 信用度

誤差テーブルと分類テーブルは、CHAID モデルがケースの約 75% を正しく分類していることを示しています。これは悪い数値ではありませんが、良い数値でもありません。また、「良い」信用度を持つケースの正分類率が楽観的すぎることが疑われます。2 種類の独立変数（「所得レベル」と「クレジットカードの数」）に関する情報が欠損していると「良い」信用度が示されるという前提があるからです。

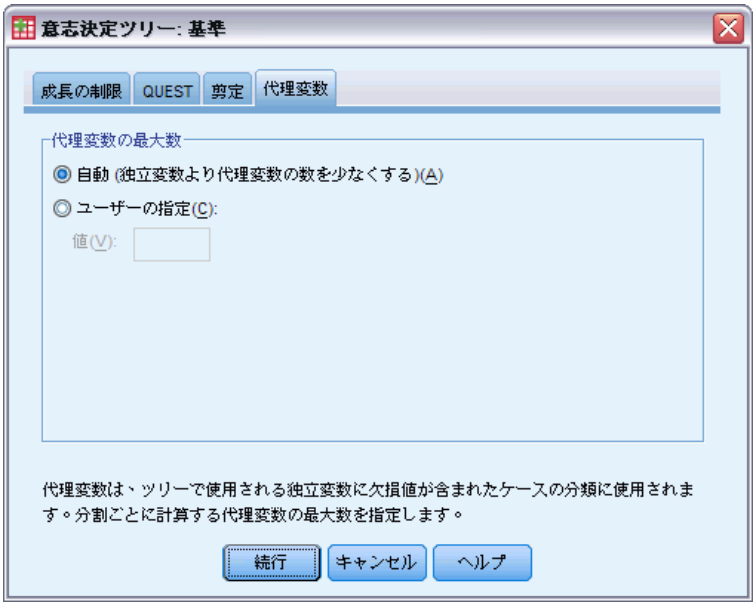
CRT を使用した場合の欠損値

成長手法として CRT を使用し、同じ基本分析を行います。

- ［ディシジョン ツリー］メイン ダイアログ ボックスで、［成長手法］として [CRT] を選択します。
- [基準] をクリックします。
- ケースの最小数がそのまま、親ノードは 400、子ノードは 200 に設定されていることを確認してください。
- [代理変数] タブをクリックします。

  注:成長手法として [CRT] または [QUEST] を選択していないと［代理変数］タブは表示されません。

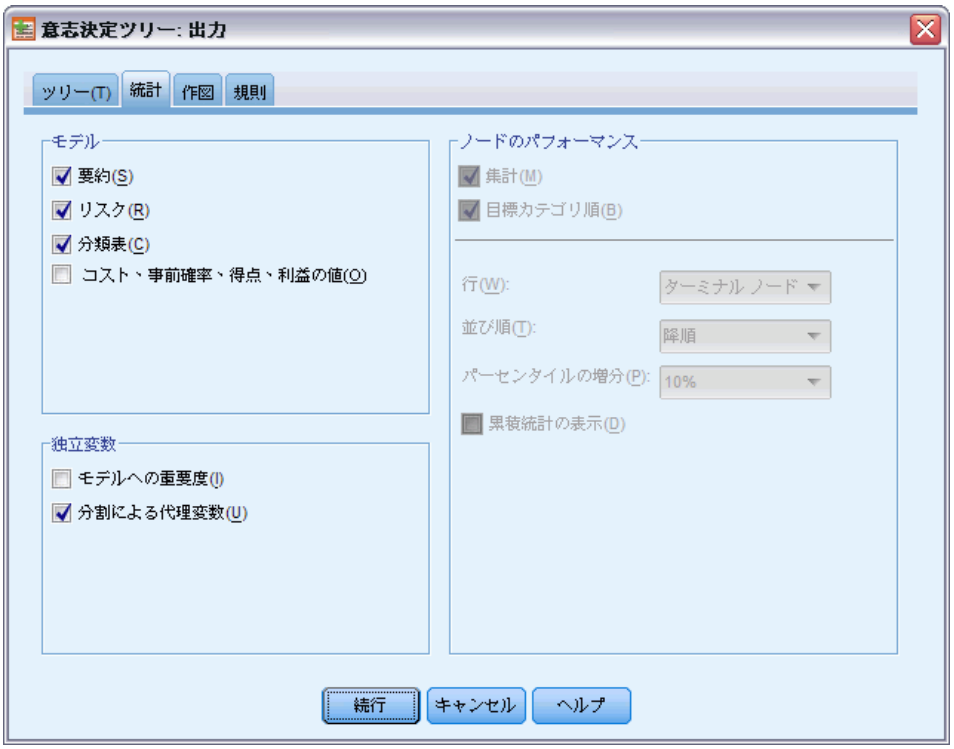
図 6-6
[基準] ダイアログ ボックスの [代理変数] タブ



各独立変数のノードの分割を [自動] に設定すると、モデル用に指定した他のすべての独立変数が使用可能な代理変数であると見なされます。今回の例では独立変数はそれほど多くないので、[自動] 設定が適しています。

- [続行] をクリックします。
- [ディシジョン ツリー] メイン ダイアログ ボックスの [出力] をクリックします。

図 6-7
[分類ツリー: 出力] ダイアログ ボックスの [統計] タブ



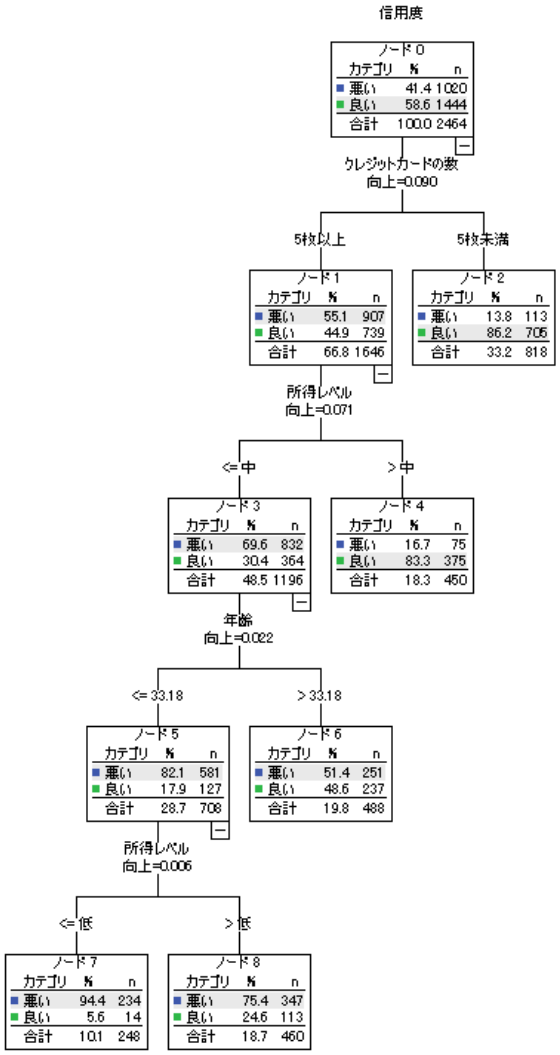
► [統計] タブをクリックします。

► [分割による代理変数] を選択します。

► [続行] をクリックし、[OK] をクリックして手続きを実行します。

CRT の結果

図 6-8
独立変数に欠損値のある CRT ツリー



今度のツリーが CHAID ツリーと似ていないことはすぐにわかります。そのこと自体には、必ずしも深い意味はありません。CRT ツリー モデルでは、すべての分割が 2 値です。つまり、各親ノードは 2 つの子ノードにしか分割されません。CHAID モデルでは、親ノードを多数の子ノードに分割できます。したがって、基礎となるモデルが同じであっても、それを表すツリーの外観は異なることが多いのです。

ただし、次の点で、2 つのツリーには重要な違いがあります。

- CRT モデルでは、最も重要な独立（予測）変数は「クレジットカードの数」ですが、CHAID モデルでは、最も重要な予測変数は「所得レベル」です。
- クレジット カードの数が 5 枚未満のケースの場合、「クレジットカードの数」が信用度に関する唯一の有意な予測変数となり、ノード 2 がターミナル ノードとなります。
- CHAID モデルと同様に、このモデルには「所得レベル」と「年齢」も含まれていますが、「所得レベル」は第 1 予測変数ではなく第 2 予測変数となります。
- CRT モデルでは欠損値ではなく代理予測変数が使用されるため、[<欠損値>]> カテゴリを含むノードはありません。

図 6-9
CRT モデルの誤差テーブルと分類テーブル

相対リスク

推定値	標準誤差
.224	.008

成長方法: CRT
従属変数: 信用度

分類

観測	予測値		
	悪い	良い	正確なパーセント
悪い	832	188	81.6%
良い	364	1080	74.8%
全体のパーセント	48.5%	51.5%	77.6%

成長方法: CRT
従属変数: 信用度

- 誤差テーブルと分類テーブルは、全体の正分類率が約 78% であることを示しており、CHAID モデル（75%）よりわずかに増加しています。
- 「悪い」信用度を持つケースの正分類率は、CHAID モデルがわずか 64.3% であるのに対して、CRT モデルは 81.6% とかなり高くなっています。
- しかし、「良い」信用度を持つケースの正分類率は、CHAID が 82.8% であるのに対して、CRT は 74.8% と減少しています。

代理変数

CHAID モデルと CRT モデルで比率が異なっている原因の 1 つは、CRT モデルで代理変数を使用していることです。代理変数テーブルは、モデル内で代理変数がどのように使用されたかを示しています。

図 6-10
代理変数テーブル

親ノード	独立変数		改良	連関
0	1 次	クレジットカードの数	.090	
	Surrogate	車のローン	.052	.643
		年齢	.001	.004
1	1 次	所得レベル	.071	
	Surrogate	年齢	.001	.004
3	1 次	年齢	.022	
5	1 次	所得レベル	.006	
	Surrogate	年齢	3.93E-005	.009

- ルート ノード（ノード 0）では、最適な独立（予測）変数は「クレジットカードの数」です。
- 「クレジットカードの数」に欠損値があるケースについては、「車のローン」と「クレジットカードの数」の間に非常に強い連関（0.643）があるため、この変数が代理予測変数として使用されます。
- 「車のローン」にも欠損値のあるケースについては、「年齢」が代理変数として使用されます（ただし、連関値は非常に低くわずか 0.004 です）。
- 「年齢」はノード 1 からノード 5 の「所得レベル」の代理変数としても使用されます。

要約表

成長手法によって、欠損データの処理方法が異なります。モデルの作成に使用するデータに多くの欠損値がある場合（または多くの欠損値が含まれている他のデータ ファイルにモデルを適用する場合）、さまざまなモデルで欠損値の効果を評価する必要があります。モデルで代理変数を使用して欠損値を補う場合、CTR 手法または QUEST 手法を使用します。

付録

A

サンプル ファイル

製品とともにインストールされるサンプル ファイルは、インストール ディレクトリの Samples サブディレクトリにあります。［サンプル］サブディレクトリ内に次の各言語の別のフォルダがあります。英語、フランス語、ドイツ語、イタリア語、日本語、韓国語、ポーランド語、ロシア語、簡体字中国語、スペイン語、そして繁体中国語です。

すべてのサンプル ファイルが、すべての言語で使用できるわけではありません。サンプル ファイルがある言語で使用できない場合、その言語のフォルダには、サンプル ファイルの英語バージョンが含まれています。

説明

以下は、このドキュメントのさまざまな例で使用されているサンプル ファイルの簡単な説明です。

- **accidents.sav。** 与えられた地域での自動車事故の危険因子を年齢および性別ごとに調べている保険会社に関する架空のデータ ファイルです。各ケースが、年齢カテゴリと性別のクロス分類に対応します。
- **adl.sav。**脳卒中患者に提案される治療の効果を特定するための取り組みに関する架空のデータ ファイルです。医師団は、女性の脳卒中患者たちを、2 つのグループのいずれかにランダムに割り当てました。一方のグループは標準的な理学療法を受け、もう一方のグループは感情面の治療も追加で受けました。治療の 3 か月後に、各患者が日常生活の一般的な行動をどの程度とることができるかを、順序変数として得点付けしました。
- **advert.sav。**広告費とその売上成果の関係を調べるための小売業者の取り組みに関する架空のデータ ファイルです。この小売業者は、そのために、過去の売上と、それに関係する広告費のデータを収集しました。
- **aflatoxin.sav。**収穫物によって濃度が大きく異なる毒物であるアフラトキシンを、トウモロコシの収穫物に関して検定することに関する架空のデータ ファイルです。ある穀物加工業者は、8 つそれぞれの収穫物から 16 のサンプルを受け取って、10 億分の 1 単位でアフラトキシン レベルを測定しました。
- **anorectic.sav。**拒食行動または過食行動の標準的な症状の特定を目指して、調査員 が、摂食障害を持つ大人 55 人の調査を行いました。各患者が 4 年間で 4 回診察を受けたので、観測値は合計で 220 になりました。観測値ごとに、16 種類の症状に関して患者の得点が記録

© Copyright IBM Corporation 1989, 2012.

されました。患者 71（2 回目）、患者 76（2 回目）、患者 47（3 回目）の症状の得点が見つからなかったので、残っている 217 回分の観測値が有効です。

- **bankloan.sav。**債務不履行率を低減させるための銀行の取り組みに関する架空のデータ ファイルです。このファイルには、過去の顧客および見込み客 850 人に関する財務情報と人口統計情報が含まれています。最初の 700 ケースは、以前に貸付を行った顧客です。残りの 150 ケースは見込み顧客で、これらの顧客に関して銀行は信用リスクの良し悪しを分類する必要があります。
- **bankloan_binning.sav。**過去の顧客 5,000 人に関する財務情報と人口統計情報を含む架空のデータ ファイルです。
- **behavior.sav。**52 人の学生に 15 の状況と 15 の行動の組み合わせについて、0 = 「非常に適切」から 9 = 「非常に不適切」までの 10 段階でランク付けするよう依頼した研究があります 。個人間の平均を取ったため、値は非類似度としてみなされます。
- **behavior_ini.sav。**このデータ ファイルには、behavior.sav の 2 次元の解の初期布置が含まれています。
- **brakes.sav。**高性能自動車のディスク ブレーキを生産している工場での品質管理に関する架空のデータ ファイルです。このデータ ファイルには、8 台の機械で生産した 16 個のディスクの直径測定値が含まれています。ブレーキの目標の直径は 322 ミリメートルです。
- **breakfast.sav。**21 人の Wharton School MBA の学生およびその配偶者に、15 種類の朝食を好みの順に（1 = 「最も好き」から 15 = 「最も嫌い」まで）ランク付けするよう依頼した研究があります 。調査対象者の嗜好は、「すべて」から「スナックとドリンクのみ」まで、6 つの異なるシナリオに基づいて記録されました。
- **breakfast-overall.sav。**このデータ ファイルには、最初のシナリオ（「すべて」）のみの朝食の好みが含まれています。
- **broadband_1.sav。**全国規模のブロードバンド サービスの地域ごとの契約者数を含む架空のデータ ファイルです。このデータ ファイルには、85 地域の月々の契約者数が 4 年間分含まれています。
- **broadband_2.sav。**このデータ ファイルは broadband_1.sav と同じですが、データが 3 か月分追加されています。
- **car_insurance_claims.sav。**他の場所 で表示および分析される、自動車の損害請求に関するデータセットです。逆リンク関数を使用して従属変数の平均値を保険契約者の年齢、車種、製造年の線型結合と関連付けることにより、平均請求数はガンマ分布としてモデリングできます。申請された請求の数は、尺度重み付けとして使用できます。
- **car_sales.sav。**このデータ ファイルには、自動車のさまざまな車種やモデルの架空の売上推定値、定価、仕様が含まれています。定価と仕様はそれぞれ、edmunds.com と製造元のサイトから入手しました。

- **car_sales_uprepared.sav。** 変換したバージョンのフィールドを含まない car_sales.sav の修正したバージョンです。
- **carpet.sav。**一般的な例 としては、新しいカーペット専用洗剤を市販することに関心のある企業が消費者の嗜好に関する 5 種類の因子（パッケージのデザイン、ブランド名、価格、サービスシール、料金の払い戻し）の影響について調べたい場合があります。パッケージのデザインには、3 つの因子レベルがあります。それぞれ塗布用ブラシの位置が異なります。また、3 つのブランド名（K2R、Glory、および Bissell）、3 つの価格水準があり、最後の 2 つの因子のそれぞれに対しては 2 つのレベル（「なし」または「あり」）があります。10 人の消費者が、これらの因子により定義された 22 個のプロファイルに順位を付けます。変数「嗜好」には、各プロファイルの平均順位の序列が含まれています。順位が低いほど、嗜好度は高くなります。この変数には、各プロファイルの嗜好測定値がすべて反映されます。
- **carpet_prefs.sav。**このデータ ファイルは carpet.sav と同じ例に基づいていますが、10 人の消費者それぞれから収集した実際のランキングが含まれています。消費者は、22 種類の製品プロファイルを、一番好きなものから一番嫌いなものまで順位付けすることを依頼されています。変数 PREF1 から PREF22 には、carpet_plan.sav で定義されている、関連するプロファイルの ID が含まれています。
- **catalog.sav。**このデータ ファイルには、あるカタログ会社が販売した 3 つの製品の、架空の月間売上高が含まれています。5 つの予測変数のデータも含まれています。
- **catalog_seasfac.sav。**このデータ ファイルは catalog.sav と同じですが、季節性の分解手続きとそれに付随する日付変数から計算した一連の季節因子が追加されています。
- **cellular.sav。**解約率を削減するための携帯電話会社の取り組みに関する架空のデータ ファイルです。解約の傾向スコアは、0 ～ 100 の範囲でアカウントに適用されます。スコアリングが 50 以上のアカウントはプロバイダの変更を考えている場合があります。
- **ceramics.sav。**新しい上質の合金に標準的な合金より高い耐熱性があるかどうかを特定するための、ある製造業者の取り組みに関する架空のデータ ファイルです。各ケースが 1 つの合金の別々のテストを表し、軸受けの耐熱温度が記録されます。
- **cereal.sav。**880 人を対象に、朝食の好みについて、年齢、性別、婚姻状況、ライフスタイルが活動的かどうか（週 2 回以上運動するか）を含めて調査した、架空のデータ ファイルです。各ケースが別々の回答者を表します。
- **clothing_defects.sav。**ある衣料品工場での品質管理工程に関する架空のデータ ファイルです。工場で生産される各ロットから、調査員が衣料品のサンプルを取り出し、不良品の数を数えます。

- **coffee.sav。**このデータ ファイルは、6 つのアイスコーヒー ブランド について受けた印象に関連しています。回答者は、アイス コーヒーに対する 23 の各印象属性に対して、その属性が言い表していると思われるすべてのブランドを選択しました。機密保持のため、6 つのブランドを AA、BB、CC、DD、EE、および FF で表しています。
- **contacts.sav。**企業のコンピュータ営業グループの担当者リストに関する架空のデータ ファイルです。各担当者は、所属する会社の部門および会社のランクによって分類されています。また、最新の販売金額、最後の販売以降の経過時間、担当者の会社の規模も記録されています。
- **creditpromo.sav。**最近のクレジット カード プロモーションの有効性を評価するための、あるデパートの取り組みに関する架空のデータ ファイルです。このために、500 人のカード所有者がランダムに選択されました。そのうち半分には、今後 3 か月間の買い物に関して利率を下げることをプロモーションする広告を送付しました。残り半分には、通常どおりの定期的な広告を送付しました。
- **customer_dbase.sav。**自社のデータ ウェアハウスにある情報を使用して、反応がありそうな顧客に対して特典を提供するための、ある会社の取り組みに関する架空のデータ ファイルです。顧客ベースのサブセットをランダムに選択して特典を提供し、顧客の反応が記録されています。
- **customer_information.sav。** 名前や住所など、顧客の連絡先情報を含む架空のデータ ファイルです。
- **customer_subset.sav。**customer_dbase.sav の 80 件のケースのサブセット。
- **debate.sav。**政治討論の出席者に対して行った調査の、討論の前後それぞれの回答に関する架空のデータ ファイルです。各ケースが別々の回答者に対応します。
- **debate_aggregate.sav。**debate.sav 内の回答を集計する、架空のデータ ファイルです。各ケースが、討論前後の好みのクロス分類に対応しています。
- **demo.sav。**月々の特典を送付することを目的とした、購入顧客のデータベースに関する架空のデータ ファイルです。顧客が特典に反応したかどうかが、さまざまな人口統計情報と共に記録されています。
- **demo_cs_1.sav。**調査情報のデータベースをコンパイルするための、ある会社の取り組みの最初のステップに関する架空のデータ ファイルです。各ケースが別々の都市に対応し、地域、地方、地区、および都市の ID が記録されています。
- **demo_cs_2.sav。**調査情報のデータベースをコンパイルするための、ある会社の取り組みの第 2 のステップに関する架空のデータ ファイルです。各ケースが、最初のステップで選択した都市の別々の世帯単位に対応し、地域、地方、地区、都市、区画、および単位の ID が記録されます。計画の最初の 2 つの段階からの抽出情報も含まれています。

- **demo_cs.sav。**コンプレックス サンプル計画を使用して収集された調査情報を含む架空のデータ ファイルです。各ケースが別々の世帯単位に対応し、さまざまな人口統計情報および抽出情報が記録されています。
- **dmdata.sav。** これは、ダイレクト マーケティング企業の人口統計情報および購入情報を含む架空のデータです。dmdata2.sav には、テスト メールを受け取った連絡先のサブセットの情報を含み、dmdata3.sav には、テスト メールを受け取らなかった残りの連絡先に関する情報を含みます。
- **dietstudy.sav。**この架空のデータ ファイルには、”Stillman diet” の研究結果が含まれています。各ケースが別々の被験者に対応し、被験者のダイエット前後の体重（ポンド単位）と、トリグルセリド レベル（mg/100 ml 単位）が記録されています。
- **dvdplayer.sav。**新しい DVD プレーヤーの開発に関する架空のデータ ファイルです。プロトタイプを使用して、マーケティング チームはフォーカス グループ データを収集しました。各ケースが別々の調査対象ユーザーに対応し、ユーザーの人口統計情報と、プロトタイプに関する質問への回答が記録されています。
- **german_credit.sav。**このデータ ファイルは、カリフォルニア大学アーバイン校の Repository of Machine Learning Databases にある “German credit” データセットから取ったものです。
- **grocery_1month.sav。**この架空のデータ ファイルは、grocery_coupons.sav データ ファイルの週ごとの購入を「ロールアップ」して、各ケースが別々の顧客に対応するようにしたものです。その結果、週ごとに変わっていた変数の一部が表示されなくなり、買物の総額が、調査を行った 4 週間の買物額の合計になっています。
- **grocery_coupons.sav。**顧客の購買習慣に関心を持っている食料雑貨店チェーンが収集した調査データを含む架空のデータ ファイルです。各顧客を 4 週間に渡って追跡し、各ケースが別々の顧客の週に対応しています。その週に食料品に費やした金額も含め、顧客がいつどこで買物をするかに関する情報が記録されています。
- **guttman.sav。**Bell は、予想される社会グループを示す表を作成しました。Guttman は、この表の一部を使用しました。この表では、社会交互作用、グループへの帰属感、メンバとの物理的な近接性、関係の形式化などを表す 5 個の変数が、理論上の 7 つの社会グループと交差しています。このグループには、観衆（例、フットボールの試合の観戦者）、視聴者（例、映画館または授業の参加者）、公衆（例、新聞やテレビの視聴者）、暴徒（観衆に似ているが、より強い交互作用がある）、第一次集団（親密な関係）、第二次集団（自発的な集団）、および近代コミュニティ（物理的により密接した近接性と特化されたサービスの必要性によるゆるい同盟関係）があります。

- **health_funding.sav。**医療用資金（人口 100 人あたりの金額）、罹患率（人口 10,000 人あたりの人数）、医療サービス機関への訪問率（人口 10,000 人あたりの人数）のデータを含む、架空のデータ ファイルです。各ケースが別々の都市を表します。
- **hivassay.sav。**HIV 感染を発見する迅速な分析方法を開発するための、ある製薬研究所の取り組みに関する架空のデータ ファイルです。分析の結果は、8 段階の濃さの赤で表現され、色が濃いほど感染の可能性が高くなります。研究所では 2,000 件の血液サンプルに関して試験を行い、その半数が HIV に感染しており、半分は感染していませんでした。
- **hourlywagedata.sav。**管理職から現場担当まで、またさまざまな経験レベルの看護師の時給に関する架空のデータ ファイルです。
- **insurance_claims.sav。**不正請求の恐れがある、疑いを区別するためにモデルを作成する必要がある保険会社の仮説データ ファイルです。各ケースがそれぞれの請求を表します。
- **insure.sav。**10 年満期の生命保険契約に対し、顧客が請求を行うかどうかを示す危険因子を調査している保険会社に関する架空のデータ ファイルです。データ ファイルの各ケースは、年齢と性別が一致する、請求を行った契約と行わなかった契約のペアを表します。
- **judges.sav。**訓練を受けた審判（および 1 人のファン）が 300 件の体操の演技に対して付けた得点に関する架空のデータ ファイルです。各行が別々の演技を表し、審判たちは同じ演技を見ました。
- **kinship_dat.sav。**Rosenberg と Kim は、15 種類の親族関係用語（祖父、祖母、父、母、叔父、叔母、兄弟、姉妹、いとこ、息子、娘、甥、姪、孫息子、孫娘）の分析を行いました。Rosenberg と Kim は、大学生の 4 つのグループ（女性 2 組、男性 2 組）に、類似性に基づいて上記の用語を並べ替えるよう依頼しました。2 つのグループ（女性 1 組、男性 1 組）には、1 回目と違う条件に基づいて、2 回目の並べ替えをするように頼みました。このようにして、合計で 6 つの「ソース」が取得できました。各ソースは、15×15 の近接行列に対応します。この近接行列のセルの数は、ソースの人数から、ソース内でオブジェクトを分割した回数を引いたものです。
- **kinship_ini.sav。**このデータ ファイルには、kinship_dat.sav の 3 次元の解の初期布置が含まれています。
- **kinship_var.sav。**このデータ ファイルには、kinship_dat.sav の解の次元の解釈に使用できる独立変数である性別、世代、および(ation), and 親等が含まれています。特に、解の空間をこれらの変数の線型結合に制限するために使用できます。
- **marketvalues.sav。**1999 ～ 2000 年の間の、イリノイ州アルゴンキンの新興住宅地での住宅売上に関するデータ ファイルです。これらの売上は、公開レコードの問題となります。

- **nhis2000_subset.sav。**National Health Interview Survey (NHIS) は、米国国民を対象とした人口ベースの大規模な調査です。全国の代表的な世帯サンプルについて対面式で調査が行われます。各世帯のメンバーに関して、人口統計情報、健康に関する行動および状態の観測値が得られます。このデータ ファイルには、2000 年の調査から得られた情報のサブセットが含まれています。National Center for Health Statistics。National Health Interview Survey, 2000。一般使用データおよびドキュメント。ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2000/。2003 年にアクセス。
- **ozone.sav。**データには、残りの変数からオゾン濃度を予測するための、6個の気象変数に対する 330 個の観測値が含まれています。それまでの研究者 、が、他の研究者と共に、これらの変数間に非線型性を確認しています。この場合、標準的な回帰アプローチは使用できません。
- **pain_medication.sav。**この架空のデータ ファイルには、慢性関節炎を治療する抗炎症薬の臨床試験の結果が含まれています。特に興味深いことは、薬の効果が出るまでの時間と、既存の薬剤との比較です。
- **patient_los.sav。**この架空のデータ ファイルには、心筋梗塞 (MI、または「心臓発作」) の疑いで入院した患者の治療記録が含まれています。各ケースが別々の患者に対応し、入院に関連する多くの変数が記録されています。
- **patlos_sample.sav。**この架空のデータ ファイルには、心筋梗塞 (MI、または「心臓発作」) の治療中に血栓溶解剤を投薬された患者のサンプルの治療記録が含まれています。各ケースが別々の患者に対応し、入院に関連する多くの変数が記録されています。
- **poll_cs.sav。**市民の法案支持率を議会開会前に特定するための、世論調査員の取り組みに関する架空のデータ ファイルです。各ケースは登録有権者に対応しています。ケースごとに、有権者が居住している郡、町、区域が記録されています。
- **poll_cs_sample.sav。**この架空のデータ ファイルには、poll_cs.sav の有権者のサンプルが含まれています。サンプルは、poll.csplan 計画ファイルで指定されている計画に従って抽出され、このデータ ファイルには包含確率およびサンプル重み付けが記録されています。ただし、抽出計画では確率比例 (PPS) 法を使用するため、結合選択確率を含むファイル (poll_jointprob.sav) もあります。サンプル抽出後、有権者の人口統計および法案に関する意見に対応する追加の変数が収集され、データ ファイルに追加されました。
- **property_assess.sav。**限られたリソースで資産価値評価を最新に保つための、郡の評価担当者の取り組みに関する架空のデータ ファイルです。各ケースは、前年に郡内で売却された資産に対応します。データ ファイル内の各ケースでは、資産が存在する町、最後に訪問した評価担当者、その評価からの経過時間、当時行われた評価、および資産の売却価値が記録されています。

- **property_assess_cs.sav。**限られたリソースで資産価値評価を最新に保つための、州の評価担当者の取り組みに関する架空のデータ ファイルです。各ケースは州内の資産に対応します。データ ファイル内の各ケースでは、資産が存在する郡、町、および区域、最後の評価からの経過時間、および当時行われた評価が記録されています。
- **property_assess_cs_sample.sav。**この架空のデータ ファイルには、property_assess_cs.sav の資産のサンプルが含まれています。サンプルは、property_assess.csplan 計画ファイルで指定されている計画に従って抽出され、このデータ ファイルには包含確率およびサンプル重み付けが記録されています。サンプル抽出後、現在の価値変数が収集され、データ ファイルに追加されました。
- **recidivism.sav。**管轄地域での累犯率を把握するための、政府の法執行機関の取り組みに関する架空のデータ ファイルです。各ケースは元犯罪者に対応し、人口統計情報、最初の犯罪の詳細、初犯から 2 年以内の場合は 2 回目の逮捕までの期間が記録されています。
- **recidivism_cs_sample.sav。**管轄地域での累犯率を把握するための、政府の法執行機関の取り組みに関する架空のデータ ファイルです。各ケースは 2003 年の 7 月に最初の逮捕から釈放された元犯罪者に対応し、人口統計情報、最初の犯罪の詳細、2006 年 7 月までの 2 回目の逮捕のデータが記録されています。犯罪者は recidivism_cs.csplan で指定された抽出計画に従って抽出された部門から選択されます。 調査では確率比例（PPS）法を採用したため、結合選択確率を保持したファイル（recidivism_cs_jointprob.sav）も用意されています。
- **rfm_transactions.sav。**購入日、購入品目、各取引のマネタリー量など、購買取引データを含む架空のデータ ファイルです。
- **salesperformance.sav。**2 つの新しい販売トレーニング コースの評価に関する架空のデータ ファイルです。60 人の従業員が 3 つのグループに分けられ、全員が標準のトレーニングを受けます。さらに、グループ 2 は技術トレーニングを、グループ 3 は実践的なチュートリアルを受けます。トレーニング コースの最後に各従業員がテストを受け、得点が記録されました。データ ファイルの各ケースは別々の訓練生を表し、割り当てられたグループと、テストの得点が記録されています。
- **satisf.sav。**ある小売業者が 4 箇所の店舗で行った満足度調査に関する架空のデータ ファイルです。合計で 582 人の顧客を調査し、各ケースは 1 人の顧客からの回答を表します。
- **screws.sav。**このデータ ファイルには、ねじ、ボルト、ナット、鋲（びょう）の特性に関する情報が含まれています。
- **shampoo_ph.sav。**あるヘアケア製品工場での品質管理に関する架空のデータ ファイルです。定期的に、6 つの異なる製品が測定され、pH が記録されます。目標範囲は 4.5 〜 5.5 です。

- **ships.sav。**他の場所 で表示および分析される、波による貨物船への損害に関するデータセットです。件数は、船舶の種類、建造期間、およびサービス期間によって、ポワゾン率で発生するものとしてモデリングできます。因子のクロス分類によって形成されたテーブルの各セルのサービス月数の集計によって、危険にさらされる確率の値が得られます。
- **site.sav。**業務拡大に向けて新たな用地を選択するための、ある会社の取り組みに関する架空のデータ ファイルです。2 人のコンサルタントを雇って、用地を別々に評価させました。広範囲のレポートに加えて、各用地を「良い」、「普通」、「悪い」のいずれかで集計しました。
- **smokers.sav。**このデータ ファイルは、1998 年の National Household Survey of Drug Abuse から抜粋したものであり、アメリカの世帯の確率サンプルです。(http://dx.doi.org/10.3886/ICPSR02934) したがって、このデータ ファイルを分析する場合は、まず人口の傾向を反映させてデータを重み付けする必要があります。
- **stocks.sav** このデータ ファイルには、1 年あたりの在庫価格、量が含まれています。
- **stroke_clean.sav。**この架空のデータ ファイルには、[データの準備] オプションの手続きを使用して整理した後の、医療データベースの状態が含まれています。
- **stroke_invalid.sav。**この架空のデータ ファイルには、医療データベースの初期状態が含まれており、データ入力にいくつかエラーがあります。
- **stroke_survival。** この架空のデータ ファイルは、虚血性脳卒中で数回の困難に直面した後リハビリ プログラムを終えた患者の生存時間に関するものです。脳卒中後、心筋梗塞の発生、虚血性脳卒中、または出血性脳卒中が注意され、イベントの時間が記録されます。脳卒中後に実施されたリハビリ プログラムの最後まで生存した患者のみが含まれるため、サンプルは左側が切り捨てられます。
- **stroke_valid.sav。**この架空のデータ ファイルには、[データの検証] 手続きを使用して確認した後の、医療データベースの状態が含まれています。異常である可能性のあるケースが含まれています。
- **survey_sample.sav。** このデータ ファイルには、人口統計データおよびさまざまな態度指標などの調査データが含まれています。これは「1998 NORC General Social Survey」の変数のサブセットに基づいていますが、いくつかのデータ値が変更され、追加の架空変数がデモの目的で追加されています。
- **telco.sav。**顧客ベースにおける解約率を削減するための電気通信会社の取り組みに関する架空のデータ ファイルです。各ケースが別々の顧客に対応し、人口統計やサービス利用状況などのさまざまな情報が記録されています。
- **telco_extra.sav。**このデータ ファイルは telco.sav データ ファイルに似ていますが、「期間」および対数変換された顧客支出の属性が削除され、標準化された対数変換顧客支出の変数に置き換えられています。

- **telco_missing.sav。**このデータ ファイルは telco.sav データ ファイルのサブセットですが、一部の人口統計データ値が欠損値に置き換えられています。
- **testmarket.sav。**この架空のデータ ファイルは、新しいメニューを追加しようというファースト フード チェーンの計画に関連しています。新製品をプロモーションするためのキャンペーンには 3 つの候補があるため、新メニューはいくつかのランダムに選択した市場にある場所で紹介されます。場所ごとに別々のプロモーションを使用し、最初の 4 週間の新メニューの週間売上高が記録されます。各ケースが場所と週に対応します。
- **testmarket_1month.sav。**この架空のデータ ファイルは、testmarket.sav データ ファイルの週ごとの売上を「ロールアップ」して、各ケースが別々の場所に対応するようにしたものです。その結果、週ごとに変わっていた変数の一部が表示されなくなり、売上高が、調査を行った 4 週間の売上高の合計になっています。
- **tree_car.sav。**これは、人口統計および自動車購入価格のデータを含む架空のデータ ファイルです。
- **tree_credit.sav。**これは、人口統計および銀行ローン履歴のデータを含む架空のデータ ファイルです。
- **tree_missing_data.sav。**これは、人口統計および銀行ローン履歴のデータと、多数の欠損値を含む架空のデータ ファイルです。
- **tree_score_car.sav。**これは、人口統計および自動車購入価格のデータを含む架空のデータ ファイルです。
- **tree_textdata.sav。**尺度および値ラベルを割り当てる前の、変数のデフォルトの状態を示すことを主な目的とする、変数を 2 つだけ含む単純なデータ ファイルです。
- **tv-survey.sav。**テレビ スタジオで実施された、ヒットした番組の放送期間を延長するかどうかを検討する調査に関する架空のデータ ファイルです。906 人の回答者に、さまざまな条件下でこの番組を視聴するかどうかを質問しました。各行は別々の回答者を表し、各列は別々の条件を表します。
- **ulcer_recurrence.sav。**このファイルには、潰瘍の再発を防ぐための 2 つの治療の有効性を比較するように計画された調査の情報の一部が含まれています。これは区間調査の良い例であり、他の場所 で表示および分析されています。
- **ulcer_recurrence_recoded.sav。**このファイルでは、ulcer_recurrence.sav の情報が、単に調査終了時のイベント確率ではなく調査の区間ごとのイベント確率をモデリングできるように再編成されています。これは他の場所 で表示および分析されています。
- **verd1985.sav。**このデータ ファイルは調査 に関連しています。8 つの変数に対する 15 人の被験者の回答を記録しました。対象となる変数が 3 つのグループに分類されます。グループ 1 には「年齢」と「婚姻」、

グループ 2 には「ペット」と「新聞」、グループ 3 には「音楽」と「居住地域」がそれぞれ含まれます。「ペット」は多重名義として尺度化され、「年齢」は順序として尺度化されます。また、その他のすべての変数は単一名義として尺度化されます。

- **virus.sav。**自社のネットワーク上のウィルスの影響を特定するための、インターネット サービス プロバイダ（ISP）の取り組みに関する架空のデータ ファイルです。この ISP は、ネットワーク上の感染した E メール トラフィックの（およその）パーセンテージを、発見の瞬間から脅威が阻止されるまで追跡しました。
- **wheeze_steubenville.sav。**これは、子供 に対する大気汚染の健康上の影響の長期調査から得られたサブセットです。このデータには、オハイオ州スビューベンビルの 7 歳、8 歳、9 歳、10 歳の子供を対象に行った、喘鳴の状態の反復 2 値測定と、調査の初年に母親が喫煙していたかどうかの固定記録が含まれています。
- **workprog.sav。**体の不自由な人をより良い仕事に就かせようとする政府の事業プログラムに関する架空のデータ ファイルです。プログラムの参加者候補のサンプルが追跡されました。その中には、ランダムに選ばれてプログラムに登録された人と、そうでない人がいました。各ケースが別々のプログラム参加者を表します。
- **worldsales.sav** このデータ ファイルには、大陸および製品ごとの販売収益が含まれています。

付録

B

注意事項

この情報は、世界各国で提供される製品およびサービス向けに作成されています。

IBM はこのドキュメントで説明する製品、サービス、機能は他の国では提供していない場合があります。現在お住まいの地域で利用可能な製品、サービス、および、情報については、お近くの IBM の担当者にお問い合わせください。IBM 製品、プログラム、またはサービスに対する参照は、IBM 製品、プログラム、またはサービスのみが使用することができることを説明したり意味するものではありません。IBM の知的所有権を侵害しない機能的に同等の製品、プログラム、またはサービスを代わりに使用することができます。ただし、IBM 以外の製品、プログラム、またはサービスの動作を評価および確認するのはユーザーの責任によるものです。

IBMは、本ドキュメントに記載されている内容に関し、特許または特許出願中の可能性があります。本ドキュメントの提供によって、これらの特許に関するいかなる権利も使用者に付与するものではありません。ライセンスのお問い合わせは、書面にて、下記住所に送ることができます。

IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785, U.S.A.

2 バイト文字セット (DBCS) 情報についてのライセンスに関するお問い合わせは、お住まいの国の IBM Intellectual Property Department に連絡するか、書面にて下記宛先にお送りください。

神奈川県大和市下鶴間1623番14号 日本アイ・ビー・エム株式会社 法務・知的財産 知的財産権ライセンス渉外

以下の条項はは、イギリスまたはこのような条項が法律に反する他の国では適用されません。 International Business Machines は、明示的または黙示的に関わらず、第三者の権利の侵害しない、商品性または特定の目的に対する適合性の暗黙の保証を含むがこれに限定されない、いかなる保証なく、本出版物を「そのまま」提供します一部の州では、特定の取引の明示的または暗示的な保証の免責を許可していないため、この文が適用されない場合があります。

この情報には、技術的に不適切な記述や誤植を含む場合があります。情報については変更が定期的に行われます。これらの変更は本書の新版に追加されます。IBM は、本書に記載されている製品およびプログラムについて、事前の告知なくいつでも改善および変更を行う場合があります。

© Copyright IBM Corporation 1989, 2012.

IBM 以外の Web サイトに対するこの情報内のすべての参照は、便宜上提供されているものであり、決してそれらの Web サイトを推奨するものではありません。これらの Web サイトの資料はこの IBM 製品の資料に含まれるものではなく、これらの Web サイトの使用はお客様の責任によるものとします。

IBM はお客様に対する一切の義務を負うことなく、自ら適切と考える方法で、情報を使用または配布することができるものとします。

本プログラムのライセンス取得者が (i) 別途作成されたプログラムと他のプログラム (本プログラムを含む) との間の情報交換および (ii) 交換された情報の相互利用を目的とした本プログラムに関する情報の所有を希望する場合、下記住所にお問い合わせください。

IBM Software Group, Attention:Licensing, 233 S. Wacker Dr., Chicago, IL 60606, USA.

上記のような情報は、該当する条項および条件に従い、有料で利用できるものとします。

本ドキュメントに記載されている許可されたプログラムおよびそのプログラムに使用できるすべてのライセンス認証された資料は、IBM Customer Agreement、IBM International Program License Agreement、および当社とかわした同等の契約の条件に基づき、IBM によって提供されます。

IBM 以外の製品に関する情報は、それらの製品の供給業者、公開済みの発表、または公開で使用できるソースから取得しています。IBM は、それらの製品のテストは行っておらず、IBM 以外の製品に関連する性能、互換性、またはその他の要求については確証できません。IBM 以外の製品の性能に関する質問は、それらの製品の供給業者に通知する必要があります。

この情報には、日常の業務処理で用いられるデータや報告書の例が含まれています。できる限り詳細に説明するため、例には、個人、企業、ブランド、製品などの名前が使用されています。これらの名称はすべて架空のものであり、実際の企業で使用される名称および住所とは一切関係ありません。

この情報をソフトコピーでご覧になっている場合は、写真やカラーのイラストが表示されない場合があります。

商標

IBM、IBM ロゴ、および ibm.com、SPSS は、世界の多くの国で登録された IBM Corporation の商標です。IBM の商標の現在のリストは、http://www.ibm.com/legal/copytrade.shtml を参照してください。

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

Intel、Intel のロゴ、Intel Inside、Intel Inside のロゴ、Intel Centrino、Intel Centrino のロゴ、Celeron、Intel Xeon、Intel SpeedStep、Itanium、および Pentium は、米国およびその他の国の Intel Corporation または関連会社の商標または登録商標です。

Java およびすべての Java ベースの商標およびロゴは、米国およびその他の国の Sun Microsystems, Inc. の商標です。

Linux は、米国およびその他の国における Linus Torvalds の登録商標です。

Microsoft、Windows、Windows NT、および Windows のロゴは、米国およびその他の国における Microsoft 社の商標です。

UNIX は、米国およびその他の国における The Open Group の登録商標です。

この製品は、WinWrap Basic (Copyright 1993-2007, Polar Engineering and Consulting, http://www.winwrap.com) を使用します。

その他の製品名およびサービス名等は、IBM または他の会社の商標です。

Adobe 製品のスクリーンショットは Adobe Systems Incorporated の許可を得て転載しています。

Microsoft 製品のスクリーンショットは Microsoft 社の許可を得て転載しています。

索引