IBM SPSS Data Preparation 20

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.160)の一般情報をお読みください。

本版は IBM® SPSS® Statistics 20 ,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。

はじめに

IBM® SPSS® Statistics は、データ分析の包括的システムです。Data Preparation 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。Data Preparation 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。

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。

追加の出版物

Marija Norušによる『SPSS Statistics: Guide to Data Analysis』、『SPSS Statistics: Statistical Procedures Companion』、『SPSS Statistics: Advanced Statistical Procedures Companion』が Prentice Hall から出版されました。補助的な資料としてご利用いただけます。これらの出版物には、SPSS Statistics Base モジュール、Advanced Statistics モジュール、Regression モジュールの統計的手続きについて記載されています。初めてデータ分析を行う場合、高度なアプリケーションを使用する場合に応じて、この本は IBM® SPSS® Statistics が提供している機能を効率よく使用するための手助けとなります。出版物の内容、サンプルの図表などの詳細は、作者の Web サイトを参照してください。
http://www.norusis.com

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

Data Preparation の概要

演算システムの処理能力が向上すると、それに比例して情報に対する需要も増大するため、データ収集がますます盛んになり、それに伴ってケースの個数、変数の個数、およびデータ入力エラーの件数も増加します。これらのエラーは、データ ウェアハウジングの究極の目標であるモデル予測における問題の原因となるため、データを「きれい」に保つ必要があります。ただし、貯蔵されたデータの量は、ケースを手動で確認する能力を遥かに超えているため、データを検証するために自動処理を実装することが不可欠です。

Data Preparation アドオン モジュールを使用すると、アクティブなデータセットの中にある異常なケースや、無効なケース、変数、およびデータ値を特定し、モデル作成のデータを準備できます。

Data Preparation の手続きの使用

Data Preparation の手続きの使用方法は、目的に応じて異なります。データのロード後の道筋は次のようになります。

- **メタデータの準備。** データ ファイル内の変数を確認し、有効な値、ラベル、および測定レベルを決定します。使用不可能でありながら誤ってコード化されることの多い変数値の組み合わせを特定します。この情報に基づいて検証規則を定義します。これは時間のかかる作業ですが、類似した属性を持つデータ ファイルを定期的に検証する必要がある場合は、その労力に見合う価値はあります。
- **データ検証。** 基本チェックを実行し、無効なケース、変数、およびデータ値を特定するために定義された検証規則に対するチェックを実行します。無効なデータが見つかると、原因を調べ、修正します。これには、メタデータの準備を通して別の手順が必要になることがあります。
- **モデルの準備。** 自動データ準備を使用して、モデル作成を改善する元のフィールドの変換を取得します。多くの予測モデルで問題を引き起こす潜在的な統計量の外れ値を特定します。一部の外れ値は、特定されていない無効な変数値の結果として発生します。これには、メタデータの準備を通して別の手順が必要になることがあります。

データ ファイルが「きれい」になったら、他の アドオン モジュールからモデルをビルドすることができます。

章
2

検証規則

規則は、ケースが有効かどうかを決定するために使われます。検証規則には次の 2 種類があります。

- **単一変数規則。**単一変数規則は、範囲外の値のチェックなど、1 つの変数に適用されるチェックの固定された集合によって構成されます。単一変数規則では、有効な値は値の範囲や許容可能な値のリストとして表現されます。
- **クロス変数規則。**クロス変数規則は 1 つの変数または変数の組み合わせに対して適用できるユーザー定義の規則です。クロス変数規則は、無効な値を示す論理式で定義されます。

検証規則は、データ ファイルのデータ辞書に保存されます。これによって、いったん規則を指定したらそれを再利用することができます。

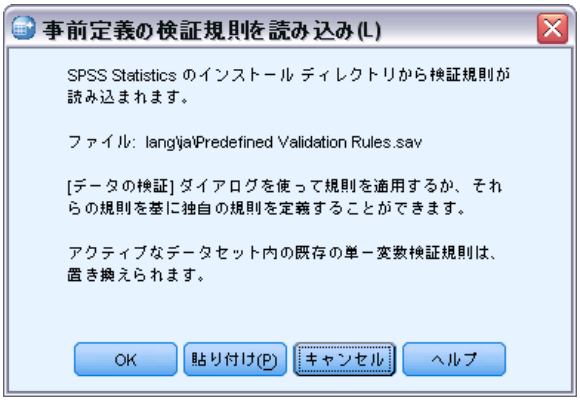
事前定義の検証規則のロード

インストレーション キットに付属している外部データ ファイルから既定義の規則を読み込むことによって、利用可能な検証規則のグループを取得することができます。

事前定義の検証規則をロードするには

▶ メニューから次の項目を選択します。
データ > 検証 > 事前定義の規則をロード...

図 2-1
事前定義の検証規則のロード



このプロセスによってアクティブなデータセット内の既存の単一変数規則が削除されることに注意してください。

また、データ プロパティのコピー ウィザードを使用して、データ ファイルから規則をロードすることもできます。

検証規則を定義

[検証規則を定義] ダイアログ ボックスを使って、単一変数規則とクロス変数規則を作成することができます。

検証規則を作成および表示するには

▶ メニューから次の項目を選択します。

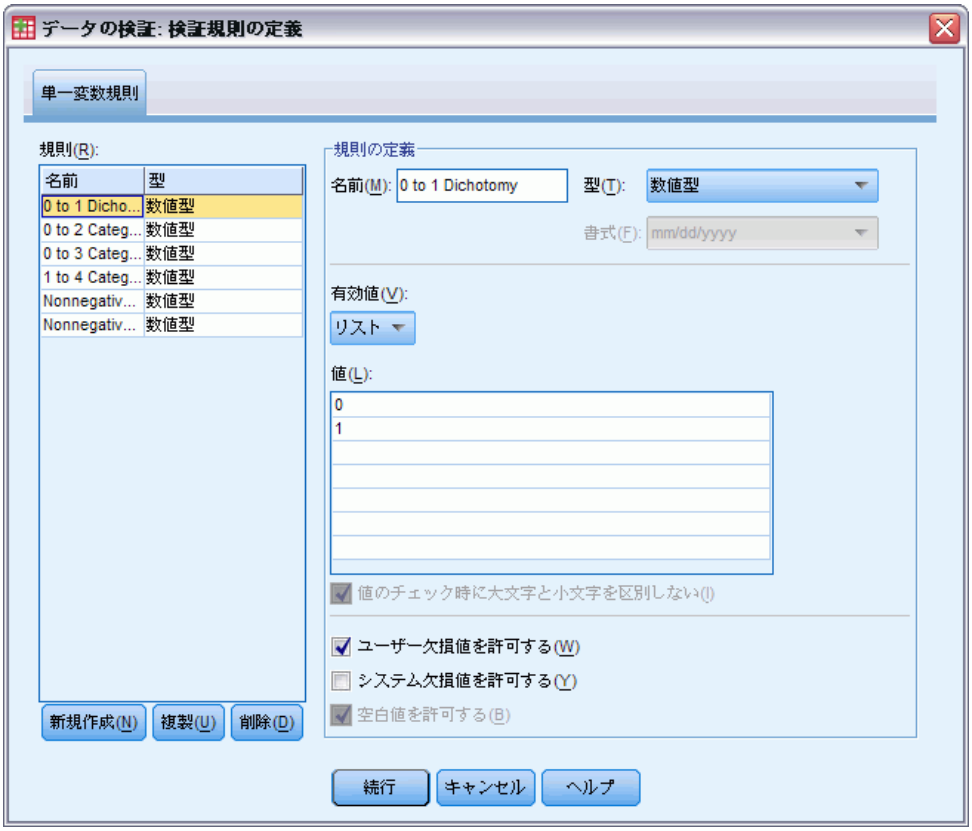
データ > 検証 > 規則の定義...

このダイアログ ボックスには、データ辞書から読み込まれた単一変数規則またはクロス変数規則が入力されます。規則がないときは、プレースホルダ規則が自動的に作成され、それを自分の目的に合うように変更することができます。

▶ [単一変数規則] タブと [クロス変数規則] タブで個々の規則を選択し、プロパティを表示および変更します。

単一変数規則を定義する

図 2-2
[検証規則の定義] ダイアログ ボックスの [単一変数規則] タブ



[単一変数規則] タブを使って、単一変数規則を作成、表示、および変更することができます。

規則。このリストは、単一変数検証規則を名前順で表示し、規則を適用できる変数の種類を表示します。このダイアログ ボックスが開かれると、データ辞書内で定義されている規則を表示します。定義されている規則がない場合は、「単一変数規則 1」という名前のプレースホルダ規則が表示されます。[規則] リストの下には、次のボタンが表示されます。

- **新規。**[規則] リストの一番下に新しい項目を追加します。その規則は選択され、「SingleVarRule n」という名前が付けられます。ここでの n は、新しい規則の名前が単一変数規則とクロス変数規則の中で一意となるような整数です。
- **複製。**[規則] リストの一番下に選択された項目のコピーを追加します。規則の名前は、単一変数規則とクロス変数規則の中で一意となるように修正されます。たとえば、「SingleVarRule 1」を複製すると、最初

の複製規則の名前は「SingleVarRule 1 のコピー」となり、2 番目は「SingleVarRule 1 のコピー (2)」となります。

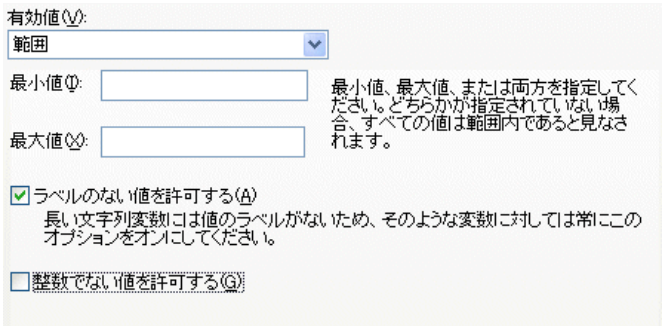
- **削除。**選択された規則を削除します。

規則の定義。これらのコントロールを使って、選択された規則のプロパティを表示および設定することができます。

- **名前。**規則の名前は、単一変数規則およびクロス変数規則の中で一意であることが必要です。
- **型。**規則を適用することができる変数の型です。[**数値**]、[**文字列**]、および［**日付**］のどれかを選択します。
- **書式。**日付変数に適用することができる規則の日付書式を選択することができます。
- **有効値。**有効値は、範囲と値のリストのいずれかで指定することができます。

［範囲の定義］では、有効な範囲を指定できます。範囲外の値は、無効として区別されます。

図 2-3
[単一変数規則] の [範囲の定義]



範囲を指定するには、最小値と最大値のどちらか、または両方を指定してください。チェック ボックスを使用すると、範囲内でラベルのない値または整数でない値を区別することができます。

［リストの定義］では、有効な値のリストを定義できます。リストに含まれない値は、無効として区別されます。

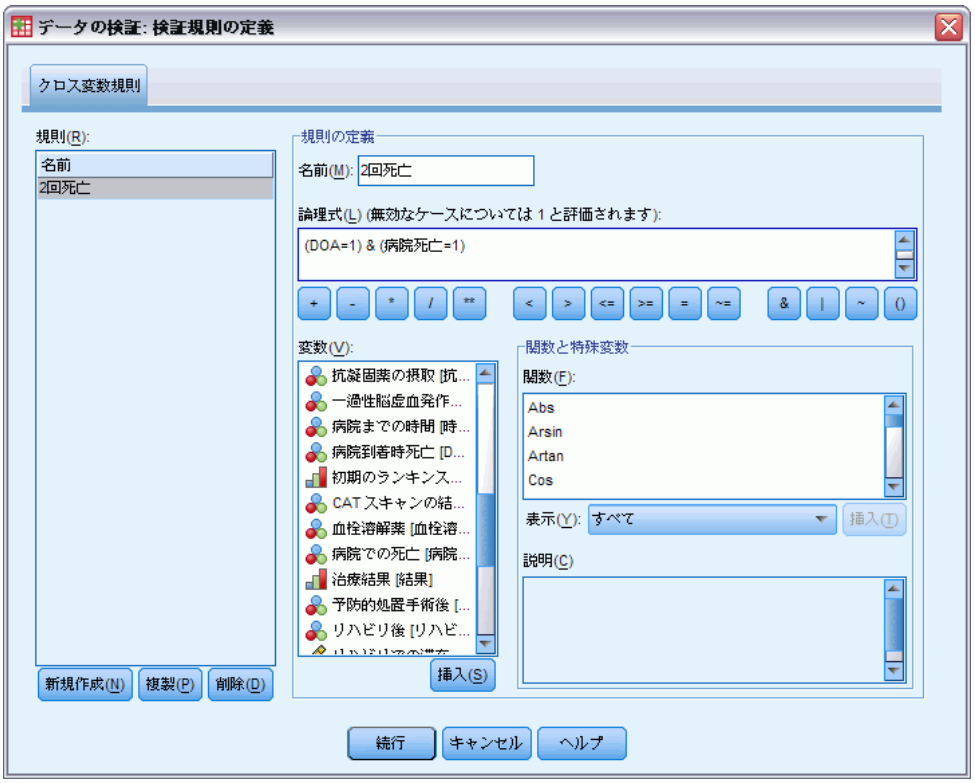
図 2-4
[単一変数規則] の [リストの定義]

格子内にリスト値を入力してください。チェック ボックスは、許容値のリストに対して文字列データ値がチェックされるときに大文字と小文字を区別するかどうかを指定します。

- **ユーザー欠損値を許可する。**ユーザー欠損値が無効として区別されるかどうかを制御します。
- **システム欠損値を許可する。**システム欠損値が無効として区別されるかどうかを制御します。文字列規則型には適用されません。
- **空白値を許可する。**空白（完全に空の値）が無効として区別されるかどうかを制御します。非文字列規則型には適用されません。

クロス変数規則を定義する

図 2-5
[検証規則を定義] ダイアログ ボックスの [クロス変数規則] タブ



[クロス変数規則] タブを使って、クロス変数規則を作成、表示、および変更することができます。

規則。このリストには、クロス変数検証規則の名前が表示されます。ダイアログ ボックスが開かれると、「CrossVarRule 1」という名前のプレースホルダ規則が表示されます。[規則] リストの下には、次のボタンが表示されます。

- **新規。**[規則] リストの一番下に新しい項目を追加します。その規則は選択され、「CrossVarRule n」という名前が付けられます。ここでの n は、新しい規則の名前が単一変数規則とクロス変数規則の中で一意となるような整数です。
- **複製。**[規則] リストの一番下に選択された項目のコピーを追加します。規則の名前は、単一変数規則とクロス変数規則の中で一意となるように修正されます。たとえば、「CrossVarRule 1」を複製すると、最初

の複製規則の名前は「CrossVarRule 1 のコピー」となり、2 番目は「CrossVarRule 1 のコピー (2)」となります。

- **削除。**選択された規則を削除します。

規則の定義。これらのコントロールを使って、選択された規則のプロパティを表示および設定することができます。

- **名前。**規則の名前は、単一変数規則およびクロス変数規則の中で一意であることが必要です。
- **論理式。**これは実質的に規則の定義です。無効なケースが 1 に評価されるように式をコード化してください。

式の作成

▶ 式を作成するには、[数式] ボックスに成分を貼り付けるか、直接入力します。

- [関数グループ] リストからグループを選択し、[関数と特殊変数] リストで関数または変数をダブルクリックする（または、関数や変数を選択し、[**挿入**] をクリックする）ことで、関数や通常使用するシステム変数を貼り付けることができます。次に、疑問符で示されたパラメータを入力します（関数のみに適用されます)。**[すべて]** というラベルの付いた関数グループには、使用可能な関数およびシステム変数がすべてリスト表示されます。現在選択している関数または変数の簡単な説明が、ダイアログ ボックスの予約領域に表示されます。
- 文字定数は、引用符またはアポストロフィで囲みます。
- 値に小数が含まれる場合、小数点には必ずピリオド (.) を使用してください。

章
3

データの検証

［データの検証］ダイアログ ボックスを使用すると、アクティブなデータセットの中にある疑わしいか無効なケース、変数、およびデータ値を特定することができます。

例: データ分析者が月次の顧客満足度レポートを依頼者に提供する必要があるとします。彼女が毎月受け取るデータは、不完全な顧客 ID、範囲外の変数値、および間違って入力されることの多い変数値の組み合わせがないかどうか品質チェックを行う必要があります。［データの検証］ダイアログ ボックスを使用して、分析者は、顧客を一意に特定する変数を指定したり、有効な変数の範囲を定める単一変数規則を定義したり、不可能な組み合わせを捕捉するためのクロス変数規則を定義したりすることができます。この手続きは、問題のケースと変数のレポートを返します。さらに、このデータには毎月同じデータ要素が含まれるため、分析者は翌月新しいデータ ファイルに規則を適用できます。

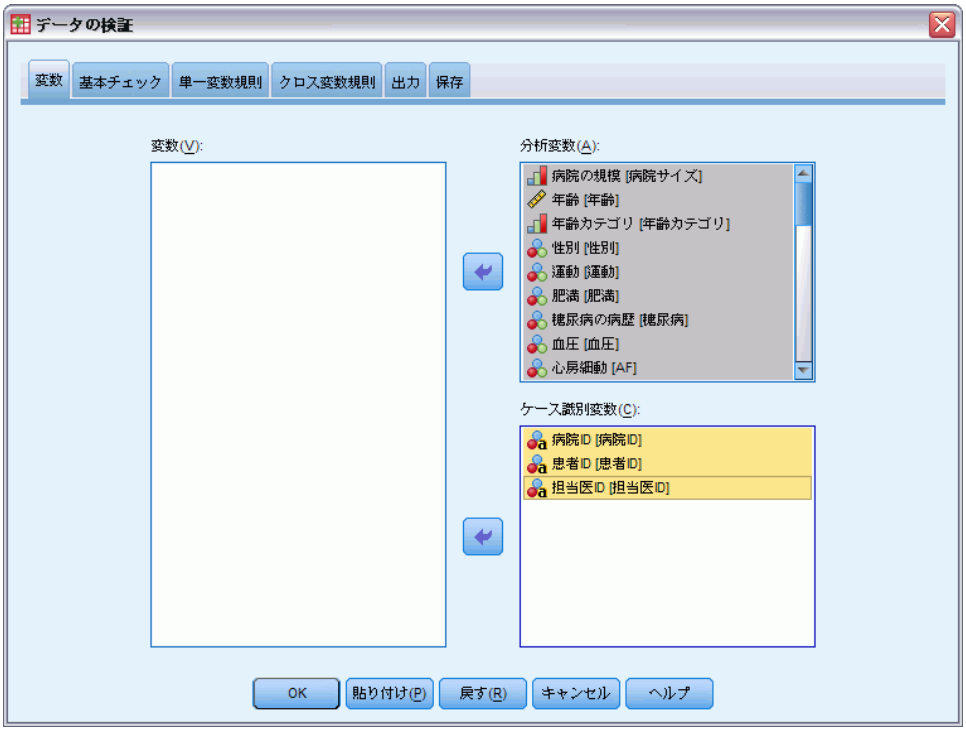
統計量。 この手続きは、さまざまなチェックを通らない変数、ケース、およびデータ値、単一変数規則およびクロス変数規則の違反数、および分析変数の簡単な記述要約のリストを作成します。

重み。 この手続きは、重み付け変数の指定を無視し、代わりに一般の分析変数として扱います。

データを検証するには

▶ メニューから次の項目を選択します。
データ > **検証(V)** > **データの検証(V)...**

図 3-1
[データの検証] ダイアログ ボックスの [変数] タブ



▶ 基本変数チェックまたは単一変数検証規則による検証のための分析変数を 1 つ以上選択します。

または、次を行うことができます。

▶ [**クロス変数規則**] タブをクリックし、1 つ以上のクロス変数規則を適用します。

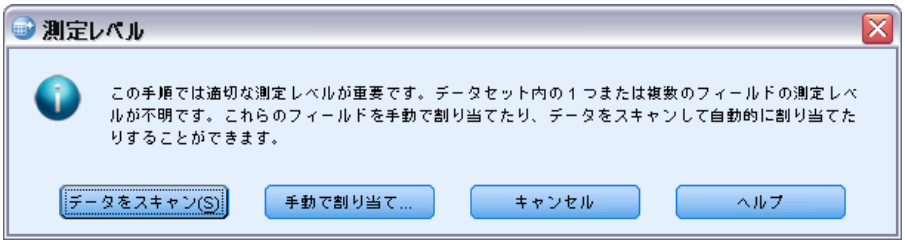
オプションとして、次の選択が可能です。

- 重複した ID や不完全な ID がないかチェックするためのケース識別変数を 1 つ以上選択します。ケース ID 変数は、ケースごとの出力にラベルを付けるためにも使用されます。2 つ以上のケース ID 変数が指定された場合は、それらの値の組み合わせがケース識別子として扱われます。

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 3-2
尺度の警告

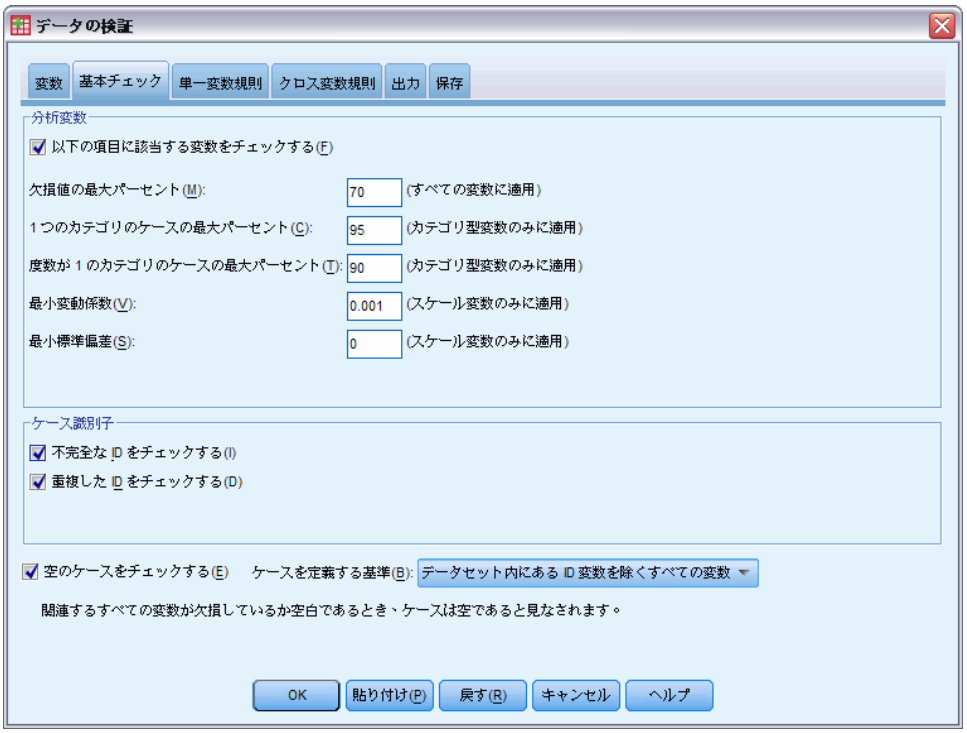


- **データをスキャン。** アクティブ データセットのデータを読み込み、デフォルトの尺度を尺度が現在不明なフィールドに割り当てます。データセットが大きい場合は時間がかかります。
- **手動で割り当てる。** 不明な尺度のフィールドをすべて表示するダイアログが開きます。このダイアログを使用して、尺度をこれらのフィールドに割り当てることができます。データ エディタの［変数ビュー］でも、尺度を割り当てることができます。

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。

[データの検証] の [基本チェック]

図 3-3
[データの検証] ダイアログ ボックスの [基本チェック] タブ



[基本チェック] タブでは、分析変数、ケース識別子、およびケース全体を選択することができます。

分析変数。 [変数] タブで分析変数を選択した場合、以下の有効性のチェックを選択することができます。チェック ボックスを使用して、チェックをオンまたはオフにできます。

- **欠損値の最大パーセント。** 欠損値の割合が指定された値より大きい分析変数を報告します。指定する値は、100 以下の正数である必要があります。
- **1 つのカテゴリのケースの最大パーセント。** 分析変数がカテゴリ型の場合、このオプションは、欠損していないカテゴリを表すケースの割合が指定された値より大きいカテゴリ分析変数を報告します。指定する値は、100 以下の正数である必要があります。パーセントは、変数の欠損値以外の値を持つケースに基づきます。
- **度数が 1 のカテゴリのケースの最大パーセント。** 分析変数がカテゴリ型の場合、このオプションでは、ケースを 1 つだけ含む変数のカテゴリの割合が、指定された値より大きいカテゴリ分析変数が報告されます。指定する値は、100 以下の正数である必要があります。

- **最小変動係数。** 分析変数がスケール型の場合、このオプションは、変動係数の絶対値が指定された値より小さいスケール分析変数を報告します。このオプションは、平均値が 0 でない変数に対してだけ適用されます。指定する値は、負でない数であることが必要です。0 を指定すると、変動チェックの係数がオフになります。
- **最小標準偏差。** 分析変数がスケール型の場合、このオプションは、標準偏差が指定された値より小さいスケール分析変数を報告します。指定する値は、負でない数であることが必要です。0 を指定すると、標準偏差チェックの係数がオフになります。

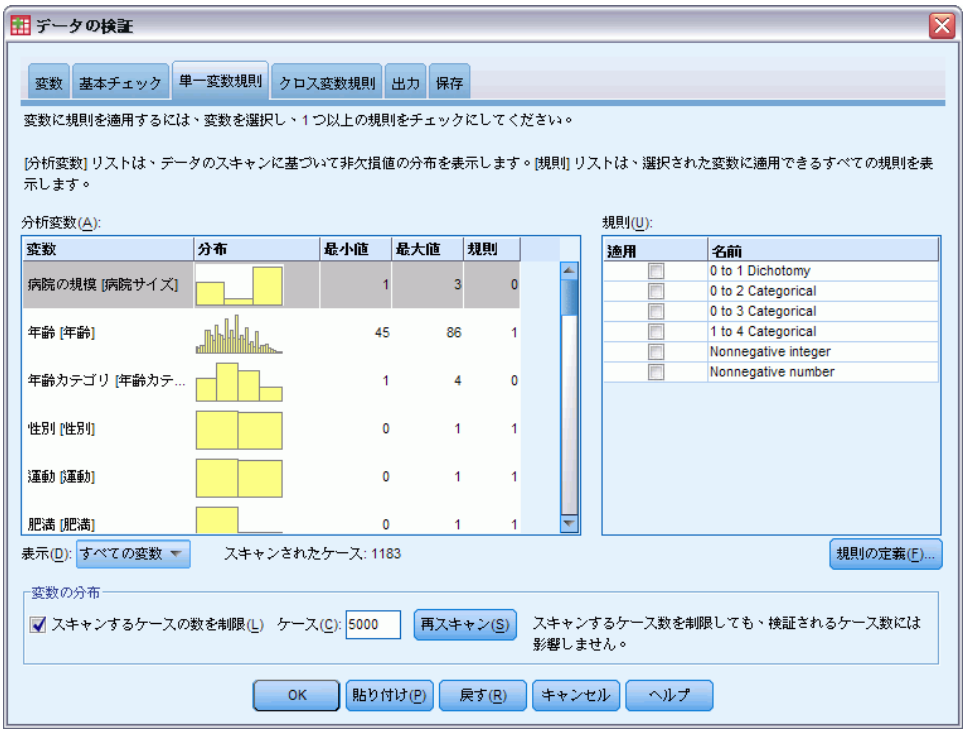
ケース識別子。［変数］タブでケース識別変数を選択した場合、以下の有効性のチェックを選択することができます。

- **不完全な ID をチェックする。** このオプションは、ケース識別子が不完全なケースを報告します。ある 1 つのケースで ID 変数が空か欠損値の場合、その識別子は不完全として扱われます。
- **重複した ID をチェックする。** このオプションは、ケース識別子が重複したケースを報告します。不完全な識別子は重複している可能性のある値のグループから除外されます。

空のケースをチェックする。 このオプションは、すべての変数が空か空白であるケースを報告します。空のケースを特定するために、ファイル内のすべての変数（ID 変数を除く）または［変数］タブに定義された分析変数だけを使用することができます。

[データの検証] の [単一変数規則]

図 3-4
[データの検証] ダイアログ ボックスの [単一変数規則] タブ



[単一変数規則] タブでは、使用可能な単一変数規則が表示され、それらの規則を分析変数に適用することができます。追加の単一変数規則を定義するには、[**規則の定義**] をクリックします。 詳細は、 2 章 p. 4 単一変数規則を定義する を参照してください。

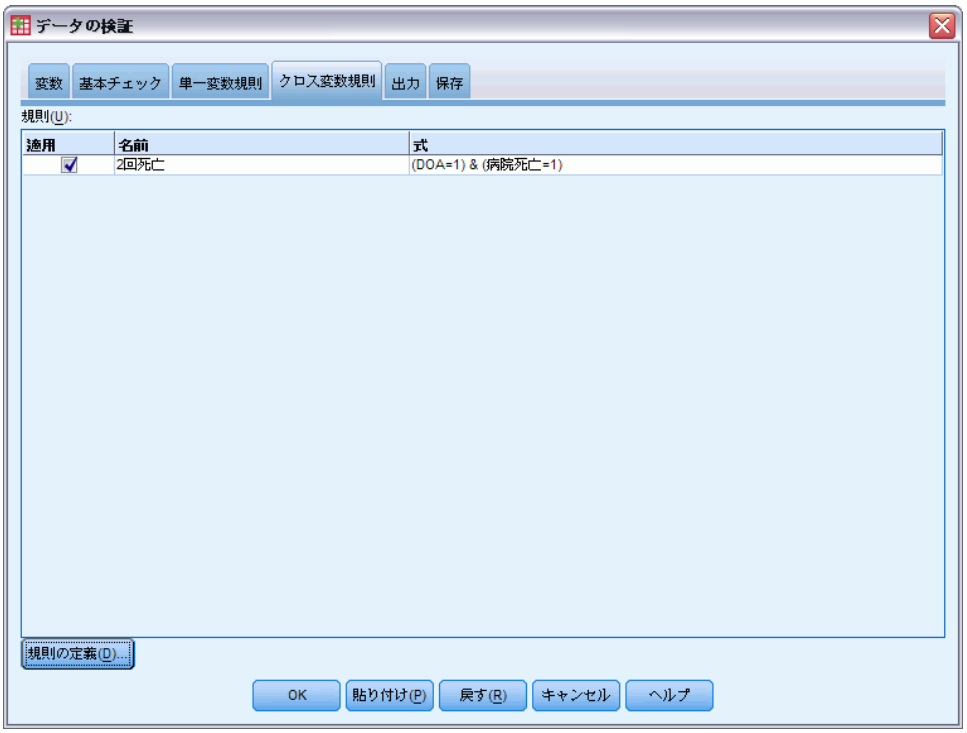
分析変数。 このリストは、分析変数を表示し、それらの分布を要約し、各変数に適用された規則の数を表示します。ユーザー欠損値とシステム欠損値が要約に含まれないことに注意してください。[表示] ドロップダウン リストは、どの変数が表示されるかを制御します。「**すべての変数**」、「**数値変数**」、「**文字列変数**、および」**日付変数**「のどれかを選択することができます。

規則。 分析変数に規則を適用するには、1 つ以上の変数を選択し、[規則] リストで適用したいすべての規則をオンにします。[規則] リストは、選択された分析変数に対して適切な規則だけを表示します。たとえば、数値変数が選択されている場合は数値規則だけが表示され、文字列変数が選択されている場合は文字列規則だけが表示されます。分析変数が選択されていないかデータ型が混在している場合、規則は表示されません。

変数の分布。［分析変数］リストに表示されている分布の要約は、すべてのケースを基にするか、［ケース］テキスト ボックスに指定して、最初の n 個のケースを基にすることができます。［**再スキャン**］をクリックすると、分布の要約が更新されます。

［データの検証］の［クロス変数規則］

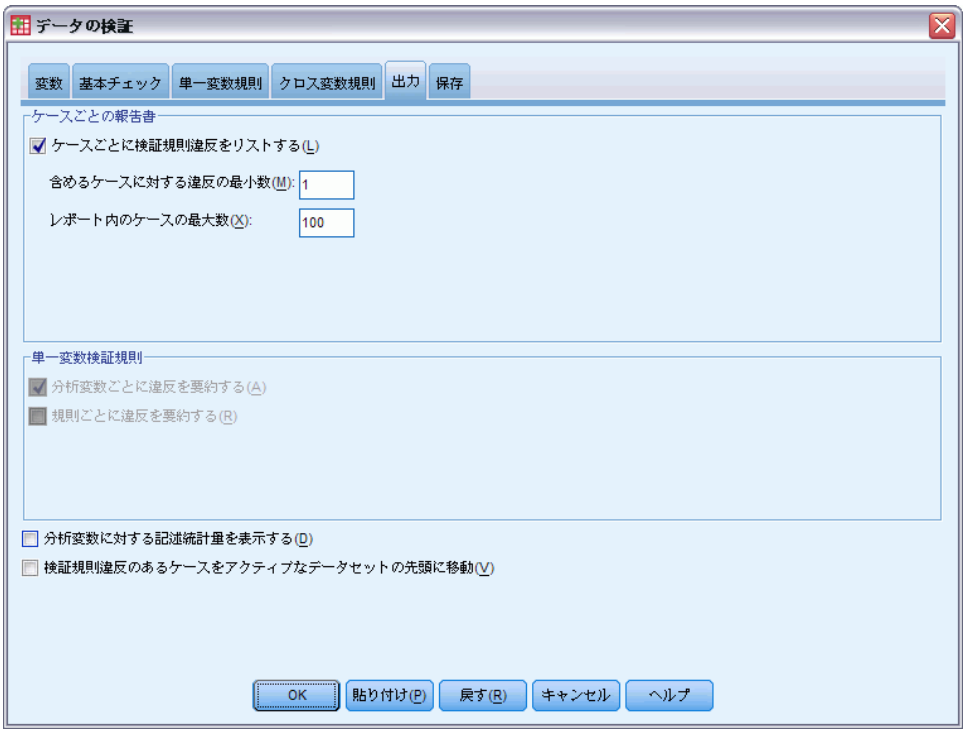
図 3-5
［データの検証］ダイアログ ボックスの［クロス変数規則］タブ



［クロス変数規則］タブでは、使用可能なクロス変数規則が表示され、それらの規則をデータに適用することができます。追加のクロス変数規則を定義するには、［**規則の定義**］をクリックします。 詳細は、 2 章 p. 7 クロス変数規則を定義する を参照してください。

[データの検証] の [出力]

図 3-6
[データの検証] ダイアログ ボックスの [出力] タブ



ケースごとの報告書。 単一変数規則またはクロス変数規則を適用した場合、ケースごとに検証規則違反を列挙するレポートを要求することができます。

- **違反の最小数。** このオプションは、レポートに含めるために必要な違反の最小数を指定します。正の整数を指定します。
- **ケースの最大数。** このオプションは、ケースのレポートに含まれるケースの最大数を指定します。1000 以下の正の整数を指定してください。

単一変数検証規則。 単一変数規則またはクロス変数規則を適用した場合、結果を表示するかどうかと、どのように表示するかを選択することができます。

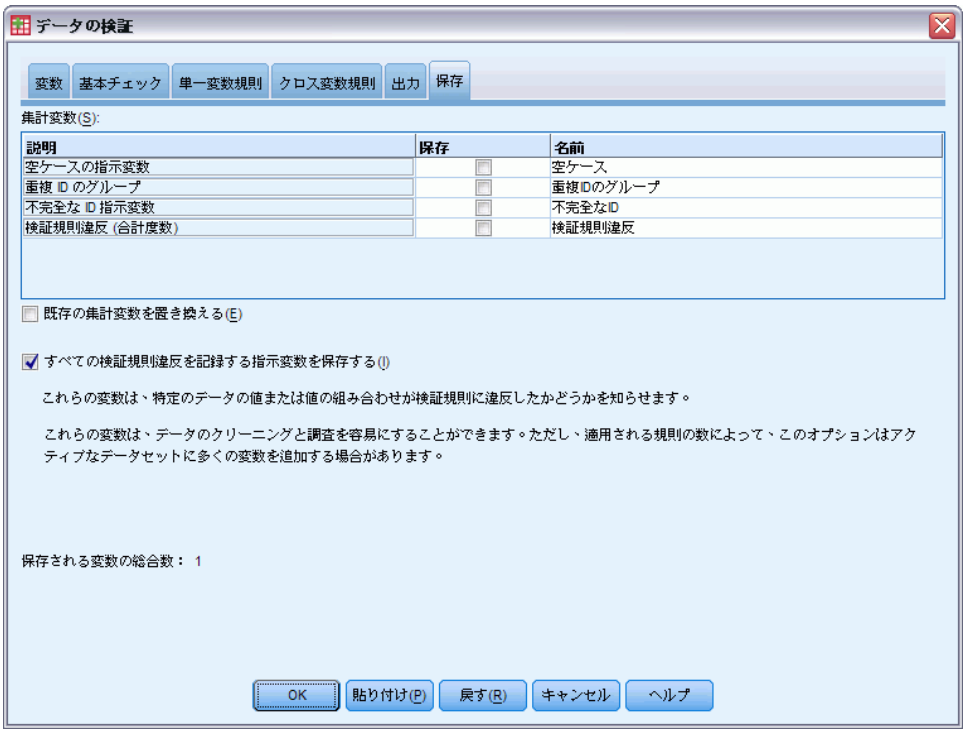
- **分析変数ごとに違反を要約する。** それぞれの分析変数について、このオプションは、違反したすべての単一変数検証規則と、それぞれの規則に違反した値の数を表示します。また、変数ごとに単一変数規則違反の総数を報告します。
- **規則ごとに違反を要約する。** それぞれの単一変数検証規則について、このオプションは、違反した規則と、それぞれの規則に対して無効な値の数を報告します。また、変数ごとに規則に違反した値の総数を報告します。

分析変数に対する記述統計量を表示。 このオプションを使用すると、分析変数の記述統計量を要求することができます。カテゴリ変数ごとに度数分布表が生成されます。スケール変数に対して、平均値、標準偏差、最小値、最大値を含む要約統計量の表が生成されます。

検証規則違反のあるケースをアクティブなデータセットの先頭に移動。 このオプションは、単一変数検証規則またはクロス変数検証規則を持つケースをアクティブなデータセットの先頭に移動します。

[データの検証] の [保存]

図 3-7
[データの検証] ダイアログ ボックスの [保存] タブ



[保存] タブでは、規則違反を記録する変数をアクティブなデータセットに保存することができます。

集計変数。 これらは、保存できる個々の変数です。保存する変数のチェック ボックスをオンにします。変数のデフォルトの名前が入力されますが、編集することができます。

- **空のケース指示変数。** 空のケースには1 の値が割り当てられます。他のすべてのケースは 0 にコード化されます。変数の値は、[基本チェック] タブで指定した範囲に反映されます。

- **重複 ID のグループ。** 同じケース識別子を持つケース（不完全な識別子を持つケースを除く）には同じグループ番号を割り当てられます。一意または不完全な識別子を持つケースは 0 にコード化されます。
- **ID 指示変数が不完全。** 空のケースまたは不完全なケースの識別子には 1 の値が割り当てられます。その他すべてのケースは 0 にコード化されます。
- **検証規則違反。** これは、ケースごとの単一変数規則違反とクロス変数規則違反の合計数です。

既存の集計変数を置き換える。 データ ファイルに保存される変数が一意の名前でない場合、同じ名前の変数を置き換えます。

識別変数を保存する。 このオプションを使用すると、検証規則違反の完全な記録を保存することができます。それぞれの変数は、検証規則の応用例に対応し、ケースが規則に違反した場合に値が 1 になり、そうでない場合に値が 0 になります。

章
4

自動データ準備

分析に向けてデータを準備することは、プロジェクトにおいて最も重要な手順の 1 つですが、従来は最も時間を消費する手順の 1 つでもありました。自動データ準備（ADP）は、データ分析および修正の特定、問題となる、または有用でないと考えられるフィールドの除外、必要に応じた新しい属性の取得、高度なスクリーニング手法を用いたパフォーマンスの改善を行い、タスクを処理します。完全に**自動化**した方法でアルゴリズムを使用して、修正を選択または適用したり、**インタラクティブ**な方法を使用して、必要に応じて変更を実行、承認または拒否する前に変更をプレビューすることができます。

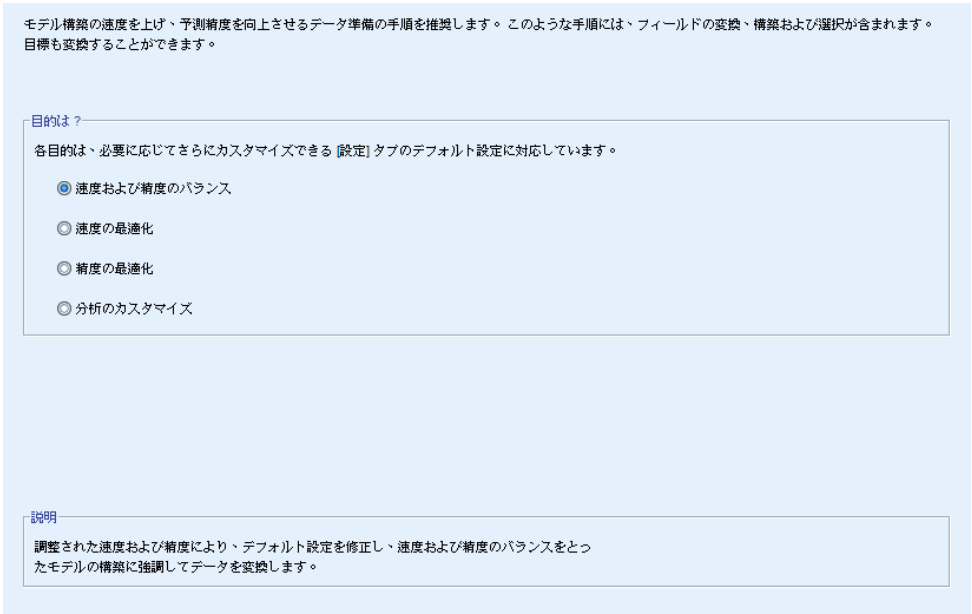
ADP を使用すると、実行する統計の概念の事前情報を必要とせず、モデルを迅速かつ用意に作成できるよう、データを準備することができます。モデルはより迅速に構築およびスコアリングするようになります。また、ADP を使用すると、、自動モデル作成プロセスの強固さをより向上させます。

注 :ADP で分析用のフィールドを準備する場合、古いフィールドの既存の値およびプロパティを置き換えるのではなく、調整または変換を含む新しいフィールドを作成します。古いフィールドは高度な分析には使用されません。役割は［なし］に設定されます。また、ユーザー欠損値情報は新たに作成されたフィールドには転送されません。新たに作成されたフィールドの欠損値はすべてシステム欠損値となります。

例: 世帯主の保険請求を調査するためのリソースが制限されている保険会社が、不正請求の恐れのある疑いを区別するためのモデルを作成したいと考えています。モデルを作成する前に、自動データ準備を使用して、モデル作成のためのデータを準備します。変換が適用される前に提案される変換を確認できる必要があるため、自動データ準備をインタラクティブ モードで使用します。 詳細は、 8 章 p.92 自動データ準備をインタラクティブに使用 を参照してください。

自動車産業グループは、さまざまな個人用自動車の売り上げを記録します。採算ベースを上回るモデルおよび下回るモデルを特定できるように、自動車の売り上げと自動車の特性との関係を確立したいと考えます。自動データ準備を使用して分析用のデータを準備し、準備「前」および準備「後」のデータを使用してモデルを作成し、結果がどのように異なるかを確認します。 詳細は、 8 章 p.103 自動データ準備を自動で使用 を参照してください。

図 4-1
自動データ準備の [目的] タブ



目的は？ 自動データ準備では、ほかのアルゴリズムがモデルを構築し、それらのモデルの予測精度を改善できる速度に影響を与えるような、データ準備の手順を推奨します。このような手順には、フィールドの変換、構築および選択が含まれます。目標も変換することができます。データ準備プロセスで重点を置く必要があるモデル作成の優先度を指定できます。

- **速度および精度のバランス:** このオプションでは、モデル作成アルゴリズムによってデータが処理される速度と、予測の精度の両方に同等の優先度を指定するよう、データを準備します。
- **速度の最適化:** このオプションでは、モデル作成アルゴリズムによってデータが処理される速度に優先度を与えるよう、データを準備します。大きいデータセットを処理する場合、または迅速な回答を求めている場合は、このオプションを選択します。
- **精度の最適化:** このオプションでは、モデル作成アルゴリズムによる予測生成の精度に優先度を与えるよう、データを準備します。
- **カスタム分析。**［設定］タブでアルゴリズムを手動で修正する場合、このオプションを選択します。継続して［設定］タブのオプションに変更を行うも、その他の目的と互換性がない場合、この設定が自動的に選択されます。

自動データ準備を取得するには

メニューから次の項目を選択します。
変換(T) > モデル作成のデータ準備 > 自動…

▶ [実行] をクリックします。

オプションとして、次の選択が可能です。

- ［目的］タブで目的を指定します。
- ［フィールド］タブでフィールドの割り当てを指定します。
- ［設定］タブでエキスパート設定を指定します。

インタラクティブ データ準備を取得するには

メニューから次の項目を選択します。
変換(T) > モデル作成のデータ準備 > インタラクティブ…

▶ ダイアログ ボックスの一番上のツールバーで [分析] をクリックします。

▶ ［分析］タブをクリックして、推奨されたデータ準備手順を確認します。

▶ 適切であれば、[実行] をクリックします。そうでない場合は、[分析のクリア] をクリックし、必要に応じて設定を変更し、[分析] をクリックします。

オプションとして、次の選択が可能です。

- ［目的］タブで目的を指定します。
- ［フィールド］タブでフィールドの割り当てを指定します。
- ［設定］タブでエキスパート設定を指定します。
- [XML の保存] をクリックして、推奨されたデータ準備の手順を XML ファイルに保存します。

[フィールド] タブ

図 4-2
自動データ準備の [フィールド] タブ

[フィールド] タブは、高度な分析に準備する必要のあるフィールドを指定します。

事前定義された役割を使用: このオプションを選択すると、既存のフィールド情報を使用します。役割が目標である単一フィールドがある場合、そのフィールドは目標として使用されます。そうでない場合、目標はありません。事前定義された役割が入力であるすべてのフィールドは、入力フィールドとして使用されます。入力フィールドは、少なくとも 1 つ必要です。

カスタム フィールド割り当ての使用: デフォルトのリストからフィールドを移動してフィールドの役割を上書きする場合、ダイアログは自動的にこのオプションに切り替わります。カスタム フィールドの割り当てを行う場合、次のフィールドを指定します。

- **目標 (省略可能)。** 目標が必要なモデルを作成する場合、目標フィールドを選択します。フィールドの役割を目標に設定する場合と類似しています。
- **入力:** 1 つ以上の入力フィールドを選択します。フィールドの役割を入力に設定する場合と類似しています。

[設定] タブ

[設定] タブは、アルゴリズムがデータをどのように処理するかを調整するために変更できる、複数グループの設定で構成されています。その他の目的と互換性のないデフォルト設定に変更を行うと、[目的] タブが自動的に更新され、[**分析のカスタマイズ**] オプションを選択します。

日付および時刻の準備

図 4-3
自動データ準備の日付および時刻の準備設定

多くのモデル作成アルゴリズムは、日付や時刻の詳細を直接処理することはできません。これらの設定を使用して、既存データの日付および時刻から、モデル入力として使用できる新しい期間データを取得できます。日付

および時刻を含むフィールドは、日付または時間のストレージ タイプで事前定義する必要があります。元の日付および時間フィールドは、自動データ準備に従うモデル入力としては推奨されません。

モデル作成の日付と時刻を準備: このオプションを選択解除すると、その他すべての［日付および時刻の準備］コントロールが無効になりますが選択は維持されます。

基準日までの経過時間を計算: 日付を含む各変数の基準日以降の年/月/日の数を生成します。

- **基準日:** 入力データの日付情報に関して、期間を計算する日付を指定します。[**今日の日付**] を選択すると、ADP が実行されている場合、現在のシステムの日付が常に使用されます。特定の日付を使用するには、[**固定日付**] を選択して、該当する日付を入力します。
- **期間(日数)の単位:** ADP が自動的に期間（日数）の単位を決定するかどうかを指定するか、年、月、または日付の [**固定単位**] を選択します。

基準時刻までの経過時間を計算: 時刻を含む各変数の基準日以降の時/分/秒の数を生成します。

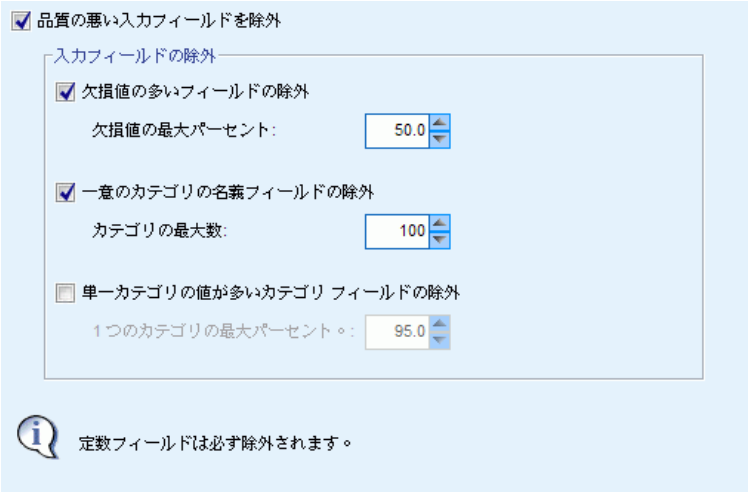
- **基準時刻:** 入力データの時間情報に関して、期間を計算する時刻を指定します。[**現在の時刻**] を選択すると、ADP が実行されている場合、現在のシステムの時刻が常に使用されます。特定の時刻を使用するには、[**固定時刻**] を選択して、該当する時刻を入力します。
- **期間(時間数)の単位:** ADP が自動的に期間（時間）の単位を決定するかどうかを指定するか、時間、分、または秒の [**固定単位**] を選択します。

周期的時間要素の取得: これらの設定を使用して、1 つの日付または時刻フィールドを 1 つまたは複数のフィールドに分割します。たとえば、3 つすべての日付チェックボックスをオンにすると、入力日付フィールド「1954-05-23」が、それぞれ [**フィールド名**] パネルで定義された接尾辞を使用する 1954、5、および 23 に分割され、元の日付フィールドは無視されます。

- **日付から取得:** 日付フィールドについて、年、月、日付またはそれらの組み合わせを取得するかどうかを指定します。
- **時刻から取得:** 時刻フィールドについて、時間、分、秒またはそれらの組み合わせを取得するかどうかを指定します。

フィールドの除外

図 4-4
自動データ準備のフィールドの除外設定



品質の悪いデータは、予測の精度に影響を与える場合があります。そのため、入力フィールドに適切な品質レベルを指定することができます。定数または 100% 欠損値であるすべてのフィールドは、自動的に除外されます。

品質の悪い入力フィールドを除外: このオプションを選択解除すると、その他すべての［フィールドを除外］コントロールが無効になりますが選択は維持されます。

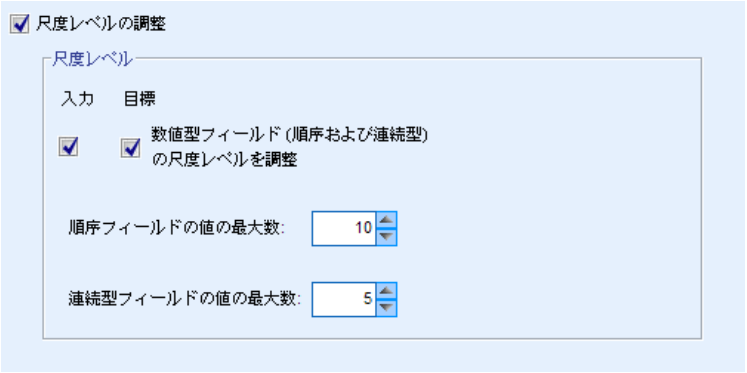
欠損値の多いフィールドの除外: 欠損値が指定された割合を超えて含まれるフィールドは、高度な分析から除外されます。0 以上 100 以下の値を指定しますが（0 はオプションの選択解除を示す）、すべての欠損値を含むフィールドは自動的に除外されます。デフォルトは 50 です。

一意のカテゴリの名義フィールドの除外: カテゴリ数が指定された数を超えて含まれるフィールドは、高度な分析から除外されます。正の整数を指定します。デフォルトは 100 です。ID、住所、名前などのモデル作成からレコード特有の情報を含むフィールドを自動的に削除する場合に役立ちます。

単一カテゴリの値が多いカテゴリ フィールドの除外: 指定された割合を超えるレコードが含まれるカテゴリを持つ順序型フィールドおよび名義型フィールドは、高度な分析から除外されます。0 以上 100 以下の値を指定しますが (0 はオプションの選択解除を示す)、定数フィールドは自動的に除外されます。デフォルトは 95 です。

尺度の調整

図 4-5
自動データ準備の尺度調整の設定



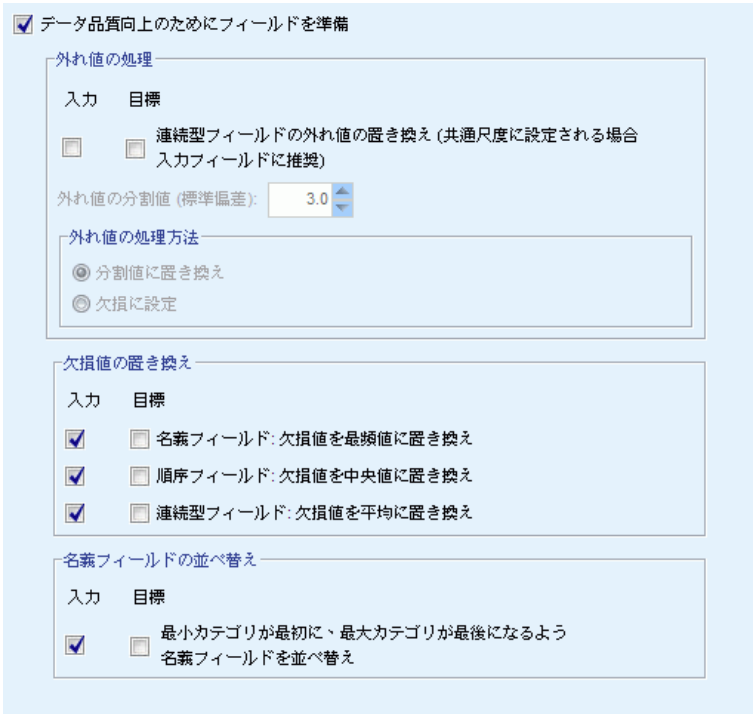
測定レベルの調整: このオプションを選択解除すると、その他すべての［測定の調整］コントロールが無効になりますが選択は維持されます。

測定レベル。 値が「少なすぎる」連続型フィールドの尺度レベルを順序型フィールドに調整するかどうか、値が「多すぎる」順序型フィールドを連続型フィールドの調整するかどうかを指定します。

- **順序フィールドの値の最大数:** 指定された数を超えたカテゴリを含む順序型フィールドは、連続型フィールドに変更されます。正の整数を指定します。デフォルトは 10 です。この値は、連続型フィールドの値の最小数以上でなければなりません。
- **連続型フィールドの値の最小数:** 一意の値が指定された数より少ない連続型フィールドは、順序型フィールドに変更されます。正の整数を指定します。デフォルトは 5 です。この値は、順序型フィールドの値の最大数以下でなければなりません。

データ品質の向上

図 4-6
自動データ準備のデータ品質向上の設定



データ品質向上のためにフィールドを準備: このオプションを選択解除すると、その他すべての［データ品質の向上］コントロールが無効になりますが選択は維持されます。

外れ値の処理: 入力フィールドおよび目標フィールドの外れ値を置き換えるかどうかを指定します。置き換える場合、標準偏差で測定した外れ値の分割値作成、および外れ値を置き換える方法を指定します。外れ値は、トリム化（分割値に設定）するか、欠損値として設定することによって置き換えることができます。欠損値に設定した外れ値は、次で選択された欠損値処理の設定にしたがって処理されます。

欠損値の置換: 連続型フィールド、名義型フィールド、または順序型フィールドの欠損値を置き換えるかどうかを指定します。

名義フィールドの並べ替え: 名義型（セット型）フィールドを最小カテゴリ(発生する頻度が最も少ない）から最大カテゴリ（発生する頻度が最も多い)の順番に並べ替えます。 新しいフィールド値は、頻度が最も少ないカテゴリの 0 から始まります。元のフィールドが文字列型である場合でも、新しいフィールドは数値型になります。たとえば、名義型フィールドの

データ値が「A」、「A」、「A」、「B」、「C」、「C」の場合、自動データ準備は「B」を 0 に、「C」を 1 に、「A」を 2 に再コード化します。

フィールドの尺度設定

図 4-7
自動データ準備のフィールドの尺度設定の設定

フィールドの尺度設定: このオプションを選択解除すると、その他すべての［フィールドの尺度設定］コントロールが無効になりますが選択は維持されます。

分析の重み付け: この変数には、分析（回帰または抽出）の重み付けが含まれます。分析の重み付けを使用して、目標フィールドのレベル間の分散における相違を処理します。連続型フィールドを選択します。

連続型入力フィールド:［z-スコア変換］または［min/max 変換］を使用して、連続型入力フィールドを正規化します。入力の尺度設定は、［選択および構築］設定で［フィールド構築の実行］を選択する場合に特に役立ちます。

- **z-スコア変換:** 観測された平均と標準偏差を母集団パラメータ推定として使用すると、フィールドは標準化され、z スコアは**最終平均値**および**最終標準偏差**が指定された正規分布の対応する値にマップされます。［**最終平均**

値] に数値を、そして [最終標準偏差] に正の数を指定します。標準化された尺度設定に対応し、デフォルトはそれぞれ 0 および 1 となります。

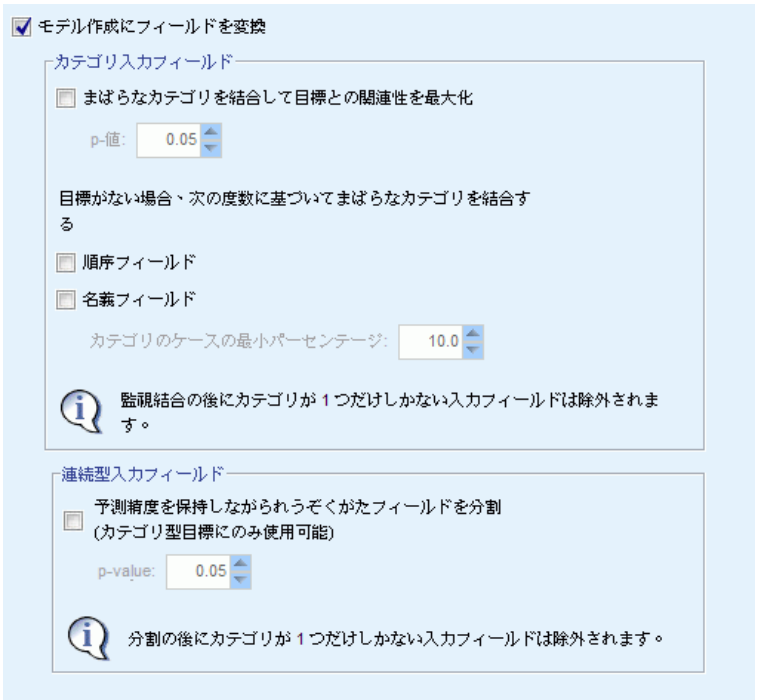
- **min/max 変換:** 観測された平均と標準偏差を母集団パラメータ推定として使用すると、フィールドは、最小値および最大値が指定された一様分布の対応する値にマップされます。[最大値] は [最小値] より大きく、値を指定します。

連続型目標: Box-Cox 変換を使用して、連続型目標を、指定された [最終平均値] および [最終標準偏差] である近似正規分布のフィールドに変換します。[最終平均値] に数値を、そして [最終標準偏差] に正の数を指定します。デフォルトはそれぞれ 0 および 1 となります。

注 :目標が ADP によって変換されている場合、変換された目標を使用して作成された後続のモデルは、変換された単位をスコアリングします。結果を解釈して使用するために、予測値を元の尺度に変換する必要があります。 詳細は、 p. 49 スコアの後方変換 を参照してください。

フィールドの変換

図 4-8
自動データ準備のフィールドの変換設定



データの予測精度を向上させるために、入力フィールドを変換することができます。

モデル作成にフィールドを変換: このオプションを選択解除すると、その他すべての［フィールドの変換］コントロールが無効になりますが選択は維持されます。

カテゴリ入力フィールド

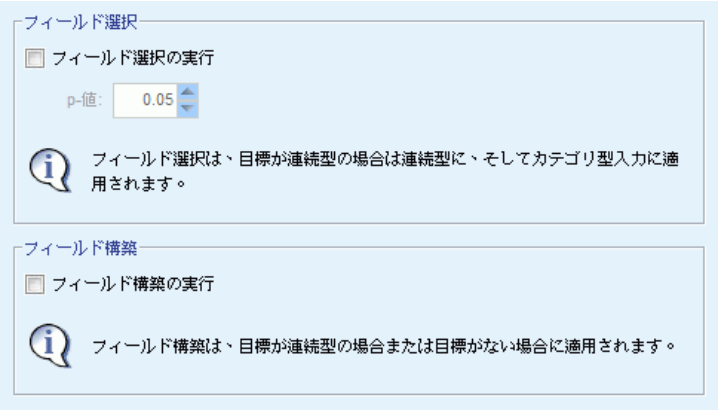
- **まばらなカテゴリを結合して目標との関連性を最大化:** 目標と関連して処理するフィールドの数を減らして、より節約的なモデルを作成します。同様のカテゴリが、入力フィールドと目標フィールド間の関係に基づいて特定されます。それほど重要でないカテゴリ、つまり p-値が指定された値より大きいカテゴリは、結合されます。0 より大きく、1 より小さい値を指定します。すべてのカテゴリが 1 つのカテゴリに結合されると、元のバージョンのフィールドおよび派生したバージョンのフィールドは、予測値がないため、高度な分析からは除外されます。
- **目標がない場合、度数に基づいてまばらなカテゴリを結合する:** データセットに目標がない場合、順序型フィールドおよび名義型フィールドのまばらなカテゴリを結合できます。等度数法を使用して、レコード数合計のパーセントが指定された最小値よりも小さいカテゴリは結合されます。0 ～ 100 の値を指定します。デフォルトは 10 です。ケース数が指定された最小パーセントに満たないカテゴリがない場合、または 2 つのカテゴリしかない場合、結合が停止します。

連続型入力フィールド: データセットにカテゴリ型目標が含まれている場合、強い関連を持つ連続型入力フィールドを分割して、処理のパフォーマンスを向上させることができます。ビンが「等質なサブグループ」に基づいて作成され、指定したp-値を等質なサブグループを決める基準値のアルファとして使用する Scheffe 手法で特定されます。0 より大きく、1 以下の値を指定します。デフォルトは 0.05 です。カテゴリ化操作によって特定フィールドに単一ビンが生成される場合、予測値としての値がないため、元のバージョンのフィールドおよびカテゴリ化されたフィールドは除外されます。

注 :ADP のカテゴリ化は最適カテゴリ化とは異なります。最適カテゴリ化では、エントロピー情報を使用して、連続型フィールドをカテゴリ フィールドに変換します。最適カテゴリ化では、データを並べ替え、メモリ内にすべて保存する必要があります。ADP では、等質サブグループを使用して、連続型フィールドを分割します。ADP カテゴリ化では、データを並べ替え、メモリ内にすべて保存する必要はありません。等質サブグループの方法を使用して連続型フィールドをカテゴリ化すると、カテゴリ化したあとのカテゴリ数は、常に目標内のカテゴリ数と等しいか少なくなります。

選択と構築

図 4-9
自動データ準備の選択と構築設定



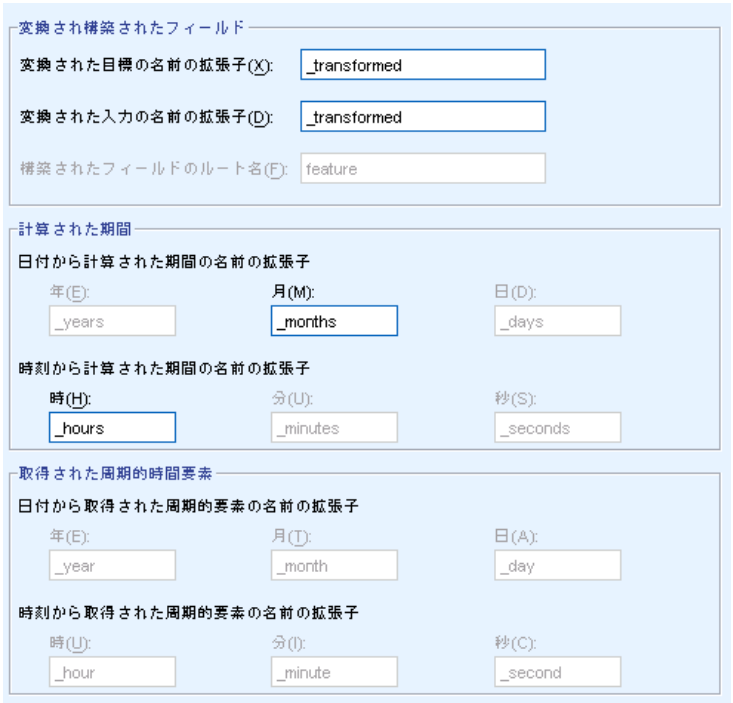
データの予測精度を向上させるために、既存フィールドに基づいて新しいフィールドを構築できます。

フィールド選択を実行: 目標フィールドを持つ相関の p-値が指定された p-値より大きい場合、連続型入力フィールド分析から削除されます。

フィールド構築の実行: 複数の既存フィールドの組み合わせから新しいフィールドを取得します。古いフィールドは、高度な分析には使用されません。このオプションは、目標が連続型の場合または目標がない場合にのみ、連続型入力フィールドに適用されます。

フィールドの名前付け

図 4-10
自動データ準備のフィールドの名前付け設定



新しいフィールドや変換されたフィールドを用意に特定できるようにするために、ADP は新しい基本名、接頭辞または接尾辞を作成し、適用します。それらの名前を修正して、ニーズおよびデータにより関連付けることができます。

変換され構築されたフィールド。 変換された目標フィールドおよび入力フィールドの適用する名前の拡張子を指定します。

さらに、[選択および構築] 設定を使用して、構築されるフィールドに適用する接頭辞名を指定します。数値の接尾辞をこの接頭辞のルート名に追加して、新しい名前を作成します。番号の形式は、次のように、取得された新しいフィールドの数によって異なります。

- 構築フィールド数が 1 ～ 9 の場合、feature1 ～ feature9 となります。
- 構築フィールド数が 10 ～ 99 の場合、feature01 ～ feature99 となります。
- 構築フィールド数が 100 ～ 999 の場合、feature001 ～ feature999 となります。

これにより、構築されたフィールドは、フィールド数に関係なく、合理的な順序で並べ替えられます。

日付および時刻から算出した期間。 日付および時刻から算出した期間に適用する名前の拡張子を指定します。

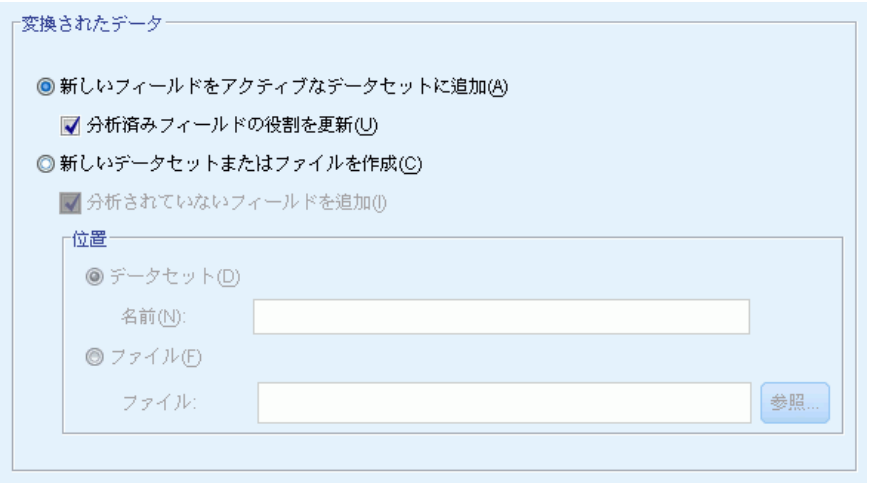
日付および時刻から算出した周期的要素。 日付および時刻から算出した周期的要素に適用する名前の拡張子を指定します。

変換の適用と保存

インタラクティブ データ準備または自動データ準備のどちらのダイアログを使用しているかによって、変換の適用および保存の設定が若干異なります。

インタラクティブ データ準備の変換の適用設定

図 4-11
インタラクティブ データ準備の変換の適用設定



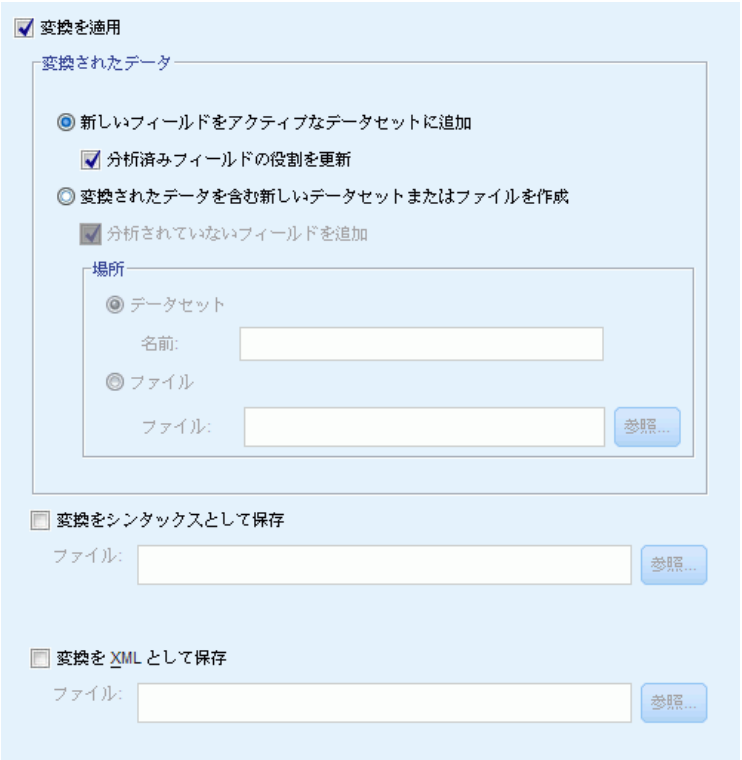
変換されたデータ。変換されたデータを保存する場所を指定します。

- **新しいフィールドをアクティブなデータセットに追加。**自動データ準備で作成されたフィールドは、新規フィールドとしてアクティブなデータセットに追加されます。[**分析済みフィールドの役割を更新**] で、自動データ準備で高度な分析から除外されたフィールドの役割を［なし］に設定します。
- **変換されたデータを含む新しいデータセットまたはファイルを作成。**自動データ準備で推奨されたフィールドは、新規データセットまたはファイルに追加されます。[**分析されていないフィールドを追加**] を選択すると、［フィールド］タブで指定されていない元のデータセットのフィールドを新しいデータセットに追加します。 ID、住所、名前などのモデル

作成で使用される情報を含むフィールドを新しいデータセットに伝送する場合に役立ちます。

自動データ準備の適用および保存の設定

図 4-12
自動データ準備の適用および保存の設定



[変換データ] グループは、インタラクティブ データ準備と同じです。自動データ準備では、次の追加オプションを使用できます。

変換を適用。[自動データ準備] ダイアログで、このオプションを選択解除すると、その他すべての [適用して保存] コントロールが無効になりますが選択は維持されます。

変換をシンタックスとして保存。推奨された変換をコマンド シンタックスとして外部ファイルに保存します。**[貼り付け]** をクリックすると変換をコマンド シンタックスとしてシンタックス ウィンドウに貼り付けるため、[インタラクティブ データ準備] ダイアログに、このコントロールはありません。

変換を XML として保存。推奨された変換を XML 形式で外部ファイルに保存します。TMS MERGE を使用してモデル PMML と結合したり、TMS IMPORT を使用して別のデータセットに適用できます。ダイアログの一番上にあ

るツールバーの［XML を保存］をクリックすると、変換を XML として保存するため、［インタラクティブ データ準備］ダイアログに、このコントロールはありません。

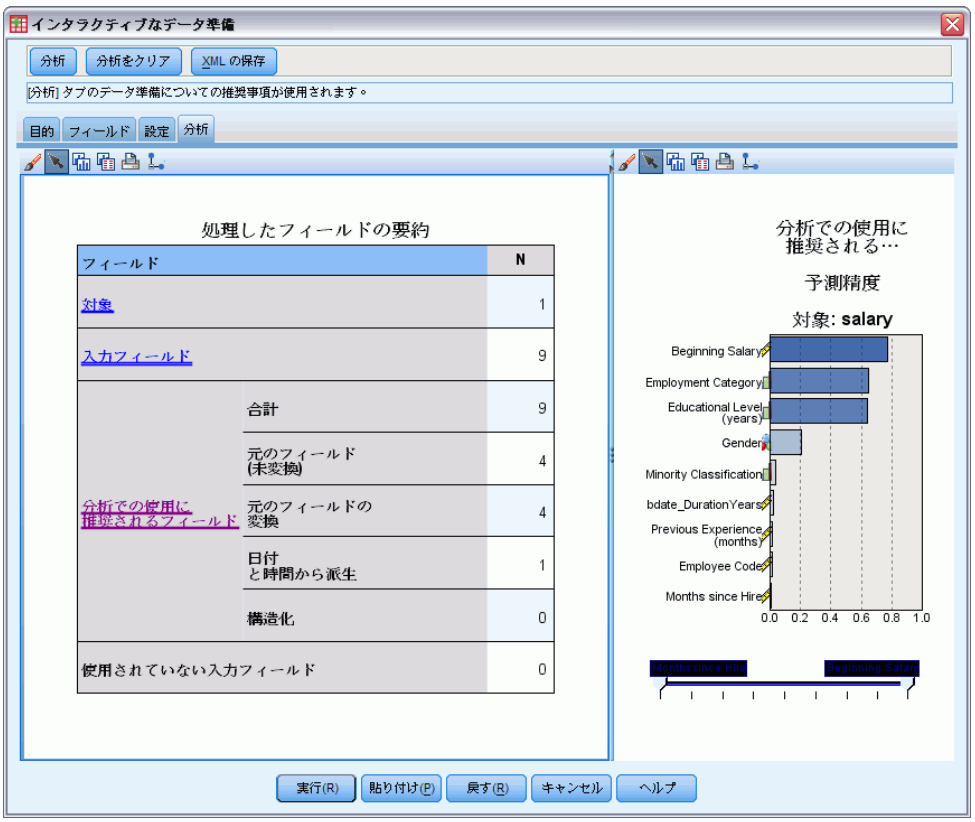
［分析］タブ

注：［インタラクティブ データ準備］ダイアログの［分析］タブを使用して、推奨された変換を確認することができます。［自動データ準備］ダイアログに、このステップはありません。

- ［目的］タブ、［フィールド］タブ、［設定］タブで行った変更など、ADP 設定に問題がない場合、**［データを分析］**をクリックしてください。アルゴリズムにより設定がデータ入力に適用され、［分析］タブに結果が表示されます。

［分析］タブには、データの処理の概要を示すテーブル形式の出力およびグラフィック出力が含まれ、スコアリング用のデータをどのように修正または改善するかについての推奨事項が表示されます。これらの推奨事項を確認し、承認したり拒否したりすることができます。

図 4-13
自動データ準備の［分析］タブ



［分析］タブは 2 つのパネルで構成されています。左側はメイン ビュー、右側はリンク ビューまたは補助ビューです。メイン ビューには、次の 3 種類があります。

- フィールド処理の要約（デフォルト）。 詳細は、 p. 37 フィールド処理の要約 を参照してください。
- フィールド。 詳細は、 p. 38 フィールド を参照してください。
- アクションの概要。 詳細は、 p. 40 アクションの概要 を参照してください。

リンク/補助ビューには、次の 4 種類あります。

- 予測の精度（デフォルト）。 詳細は、 p. 41 予測精度 を参照してください。
- フィールド テーブル。 詳細は、 p. 42［フィールド］テーブル を参照してください。
- フィールド詳細。 詳細は、 p. 43 フィールド詳細 を参照してください。
- アクションの詳細。 詳細は、 p. 46 アクションの詳細 を参照してください。

ビュー間のリンク

メイン ビューで、表内の下線付きテキストは、リンク ビューの表示を制御します。テキストをクリックすると、特定のフィールド、一連のフィールドまたは処理中のステップに関する詳細を取得できます。最後に選択したリンクは濃い色で表示されます。これにより、2 つのビュー パネルのコンテンツ間の接続を特定できます。

ビューのリセット

元の分析に関する推奨事項を再度表示し、［分析］ビューに行った変更を取り消す場合、メイン ビュー パネルの一番下にある **[リセット]** をクリックしてください。

フィールド処理の要約

図 4-14
フィールド処理の要約

フィールド処理の要約

フィールド		N
目標		1
入力フィールド		9
分析での使用が推奨される特徴	合計	8
	元のフィールド (未変換)	1
	元のフィールドの 変換	7
	日付と時刻から 派生	0
	構築済み	0
未使用の入力フィールド		1

[フィールド処理の要約] 表には、フィールドの状態や構築フィールド数への変更など、処理に対する全体の影響の射影したスナップショットが表示されます。

モデルは実際に構築されていないため、データ準備の前後に予測精度船体の変更に対する測定またはグラフはありません。その代わり、推奨された各予測の予測精度についてのグラフを表示できます。

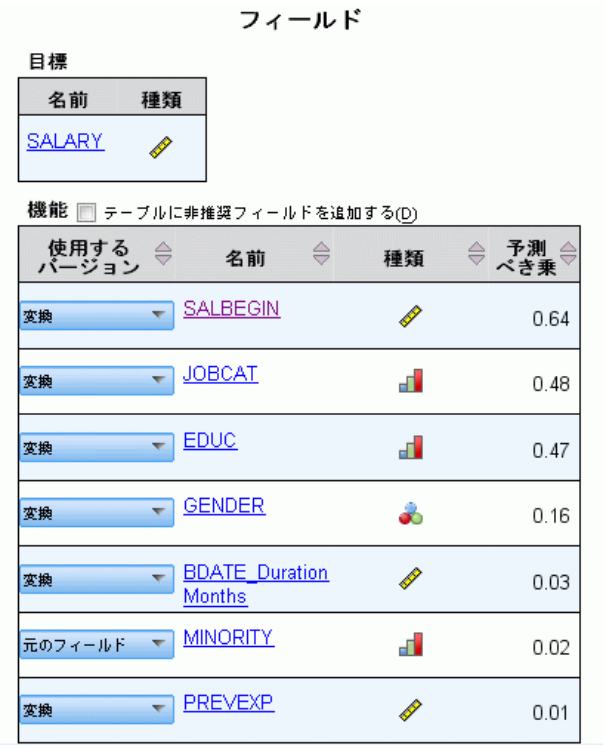
表には、次の情報が表示されます。

- 目標フィールド数。
- 元の入力予測値数。
- 分析およびモデリングでの使用が推奨される予測値。これには、推奨されるフィールド数の合計、推奨される元の変換されていないフィールド数、推奨される変換されたフィールド数（中間バージョンのフィールド、日付/時刻予測値から算出したフィールド、構築済み予測値を除く）、推奨される日付/時刻フィールドから算出したフィールド数、推奨される構築された予測値数が含まれます。
- 元の形式でも、派生フィールドとしても、あるいは構築された予測値に対する入力としても、いかなる形式でも使用が推奨されない入力予測値の数。

[フィールド] 情報に下線がある場合、クリックするとリンク ビューに詳細が表示されます。[目標]、[入力フィールド]、および [未使用の入力フィールド] の詳細は、[フィールド テーブル] リンク ビューに表示されます。詳細は、p. 42 [フィールド] テーブル を参照してください。[分析の使用が推奨されるフィールド] は、[予測精度] リンク ビューに表示されます。 詳細は、p. 41 予測精度 を参照してください。

フィールド

図 4-15
フィールド



[フィールド] メイン ビューには、処理済みフィールドと、ADP が下流モデルにそれらのフィールドの使用を推奨するかどうかを表示します。任意のフィールドについての推奨事項を上書きできます。たとえば、構築済みフィールドを除外する、または ADP が除外を推奨するフィールドを追加するなどです。フィールドが変換された場合、推奨された変換を受け入れるか、元のバージョンを使用するかを決定できます。

[フィールド] ビューは、2 つのテーブルで構成されています。1 つは目標フィールドについてのテーブル、もう 1 つは処理されたまたは作成された予測値についてのテーブルです。

[目標] テーブル

[目標] テーブルには、目標がデータに定義されているかどうかだけが表示されます。

テーブルには、次の 2 つの列があります。

- **名前。** 目標フィールドの名前またはラベルです。フィールドが変換された場合でも、元の名前が常に使用されます。
- **測定レベル。** 測定レベルを示すアイコンが表示されます。マウス ポインタをアイコンの上に停止させると、データについて説明するラベル（連続型、順序型、名義型など）が表示されます。

  目標が変換されると、[測定レベル] 列には、最終的な変換バージョンが反映されます。注 ：目標の変換をオフにすることはできません。

[予測変数] テーブル

[予測変数] テーブルは常に表示されます。テーブルの各行は、フィールドを示します。デフォルトでは、行は予測精度の高い順に並んでいます。

通常のフィールドの場合、元の名前は常に行の名前として使用されます。元のバージョンおよび派生バージョンの日付/時刻フィールドがテーブルの各行に表示されます。また、テーブルには構築済み予測値も表示されます。

テーブルに表示される変換されたバージョンのフィールドは、常に最終バージョンを示します。

デフォルトでは、推奨されたフィールドのみが、[予測変数] テーブルに表示されます。残りのフィールドを表示するには、テーブルの上にある [テーブルに非推奨フィールドを追加する] ボックスを選択します。これらのフィールドは、テーブルの一番下に表示されます。

テーブルには、次の列が表示されます。

- **使用バージョン。** フィールドを下流で使用するかどうか、推奨された変換を使用するかどうかを制御するドロップダウン リストが表示されます。デフォルトでは、ドロップダウン リストには推奨事項が反映されます。

  変換された通常の予測値の場合、[変換済み]、[変換前]、[使用しない] の 3 つの選択肢があります。

  変換されていない通常の予測値の場合、選択肢は [変換前] と [使用しない] です。

  派生した日付/時刻フィールドおよび構築済み予測値の場合、選択肢は [変換済み] と [使用しない] です。

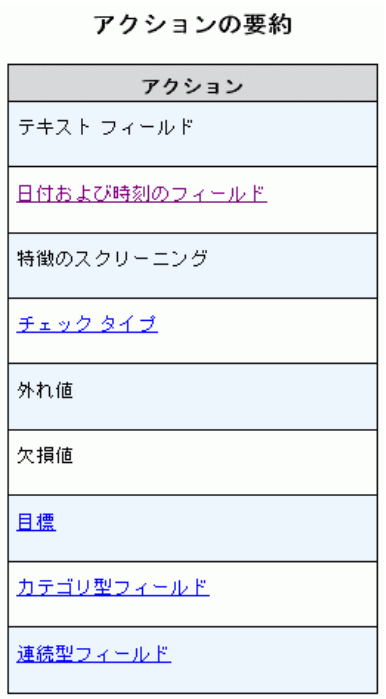
  元の日付フィールドの場合、ドロップダウン リストは無効となり、[使用しない] に設定されます。

注 :変換前バージョンと変換済みバージョンの両方の予測値の場合、[変換前] と [変換済み] でバージョンを変更すると、自動的にそれらのフィールドの [測定レベル] および [予測精度] の設定が更新されます。

- **名前。** 各フィールドの名前はリンクになっています。名前をクリックすると、フィールドに関する詳細情報がリンク ビューに表示されます。詳細は、 p. 43 フィールド詳細 を参照してください。
- **測定レベル。** データ型を示すアイコンが表示されます。マウス ポインタをアイコンの上に停止させると、データについて説明するラベル（連続型、順序型、名義型など）が表示されます。
- **予測精度。** ADP が推奨するフィールドについての予測精度のみが表示されます。この列は、目標が定義されている場合に表示されます。予測精度は 0 ～ 1 で、値が大きいほど、予測精度が「良い」ことを示します。一般的に、予測精度は ADP 分析の予測を比較するのに役立ちますが、予測精度の値を分析間で比較することはできません。

アクションの概要

図 4-16
アクションの概要



自動データ準備で実行された各アクションについて、入力予測値は変換および/または除外されます。ステップを通過したフィールドは、次のステップで使用されます。最後のステップまで通過したフィールドがモデ

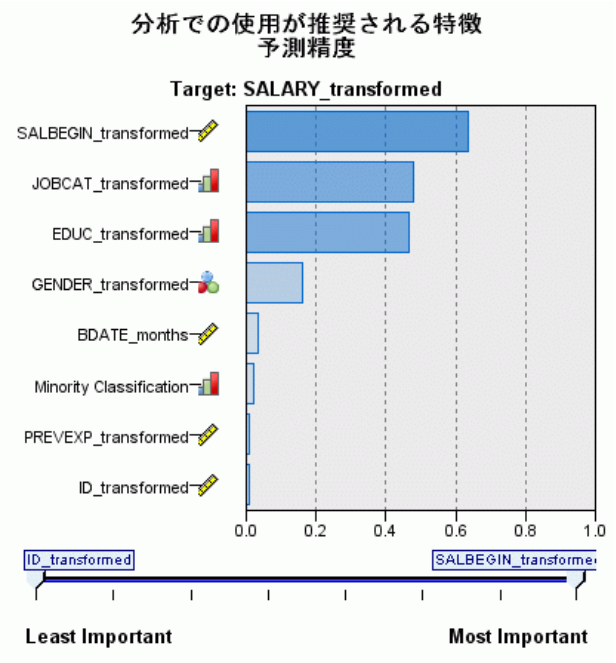
ル作成に推奨されます。変換された入力予測値および構築された予測値は除外されます。

アクションの概要は、ADP で実行された処理のアクションが表示された、単純な表です。[**アクション**] に下線がある場合、クリックすると実行された操作の詳細がリンク ビューに表示されます。 詳細は、 p. 46 アクションの詳細 を参照してください。

注 ：元のバージョンおよび最終変換されたバージョンのフィールドのみが表示され、分析中に使用された中間バージョンのフィールドは表示されません。

予測精度

図 4-17
予測精度



デフォルトでは、分析が初めて実行された場合に、または［ファイル処理の要約］ビューで［**分析およびモデリングでの使用が推奨される予測値**］を選択した場合に表示され、図用には推奨予測値の予測精度が表示されます。フィールドは、予測精度によって並べ替えられ、値が最も大きいフィールドが最上位に表示されます。

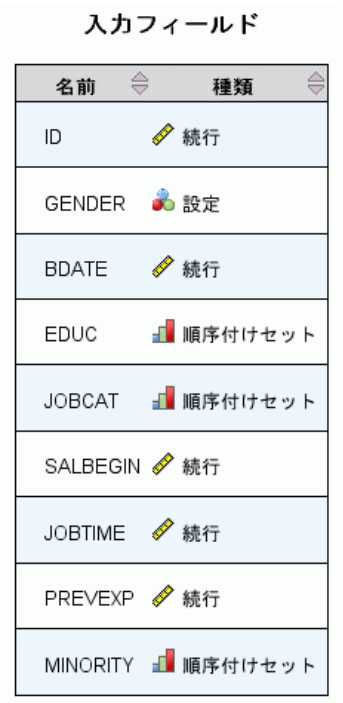
変換されたバージョンの通常の予測値の場合、フィールド名には、[設定] タブの [フィールド名] パネルで選択した接尾辞が反映されます (例: _transformed)。

各フィールド名の後に、測定レベルを示すアイコンが表示されます。

各推奨予測値の予測精度は、目標が連続型かカテゴリかに応じて、線型回帰、または naïve Bayes から算出されます。

[フィールド] テーブル

図 4-18
フィールド テーブル



[フィールド処理の要約] メイン ビューで **[目標]**、**[予測変数]**、**[未使用の予測変数]** をクリックすると表示され、[フィールド テーブル] ビューには関連するフィールドを示す単純なテーブルが表示されます。

テーブルには、次の 2 つの列があります。

- **名前。**予測値の名前。

  目標フィールドの場合、目標が変換されている場合でも、フィールドの元の名前またはラベルが使用されます。

変換されたバージョンの通常の予測値の場合、フィールド名には、[設定] タブの [フィールド名] パネルで選択した接尾辞が反映されます（例： _transformed）。

日付および時刻から派生したフィールドの場合、 最終的に変換されたバージョンの名前が使用されます（例： bdate_years）。

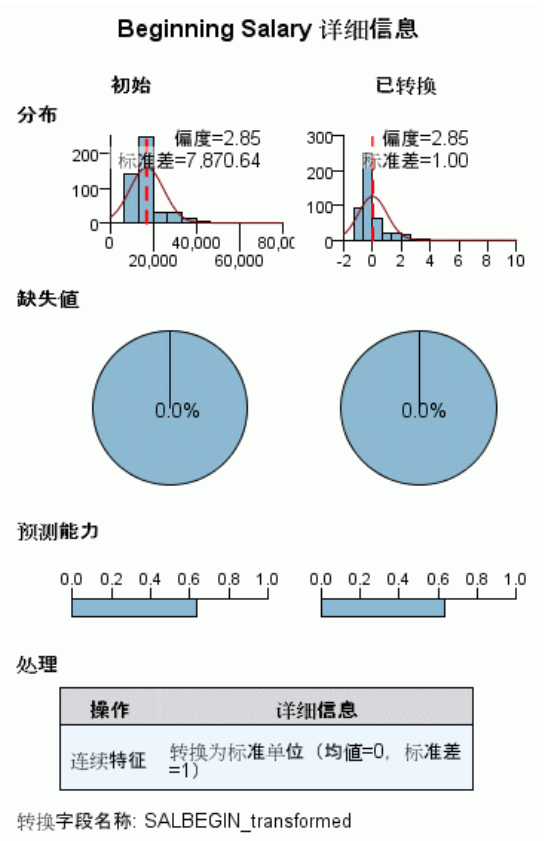
構築された予測値の場合、構築された予測値の名前が使用されます（例： Predictor1）。

- **測定レベル。** データ型を示すアイコンが表示されます。

  目標フィールドの場合、[**測定レベル**] は常に変換されたバージョンが反映されます（目標フィールドが変換されている場合）。たとえば、順序型（順序セット型）から連続型（範囲型、スケール）への変更、またはその逆も同様です。

フィールド詳細

図 4-19
フィールド詳細



[フィールド] メイン ビューで [名前] をクリックすると表示され、[フィールド詳細] ビューには選択したフィールドの分布、欠損値、予測精度グラフ（該当する場合）が表示されます。また、必要に応じて、フィールドの処理履歴や変換フィールドの名前も表示されます。

各図表セットについて、2 つのバージョンが並んで表示され、変換が適用されたフィールドと適用されていないフィールドを比較します。変換されたベージョンのフィールドがない場合、元のバージョンの図表のみが表示されます。派生した日付/時刻フィールドおよび構築済み予測値の場合、新しい予測値の図表のみ表示されます。

注 :カテゴリ数が多すぎるためにフィールドが除外された場合、処理の履歴のみが表示されます。

分布図

連続型フィールドの分布は、正規曲線が重なり、平均値を表す垂直参照線を使用したヒストグラムで表示されます。カテゴリ フィールドは棒グラフで表示されます。

ヒストグラムには、標準偏差や歪度を示すラベルがつけられています。ただし、値の数が 2 以下の場合、または元のフィールドの分散が 10 ～ 20 より小さい場合、歪度は表示されません。

図表の上にマウスポインタを停止させると、ヒストグラムの平均値、またはカテゴリのレコード数合計の度数またはパーセンテージを棒グラフで表示します。

欠損値のグラフ

円グラフは、変換が適用された場合、変換が適用されていない場合の欠損値の割合を比較します。グラフのラベルはパーセンテージを示します。

ADP が欠損値の処理を実行した場合、変換後の円グラフには置換値、つまり欠損値の変わりに使用される値もラベルで表示します。

グラフにマウスポインタを停止させると、全体のレコード数の欠損値数と全体の割合が表示されます。

予測精度グラフ

推奨フィールドについて、棒グラフに変換前後の予測精度が表示されます。目標フィールドが変換されると、予測精度は変換後の目標フィールドについて計算されます。

注 :目標が定義されていない場合、またはメイン ビュー パネルで目標をクリックした場合、予測精度のグラフは表示されません。

グラフの上のマウス ポインタを停止させると、予測精度の値が表示されます。

処理履歴表

表には、変換されたバージョンのフィールドがどのように取得されたかを示されます。ADP によって行われた処理が、実行順に表示されます。ただし、特定のステップにおいては、特定のフィールドに対して複数の処理が実行されている場合があります。

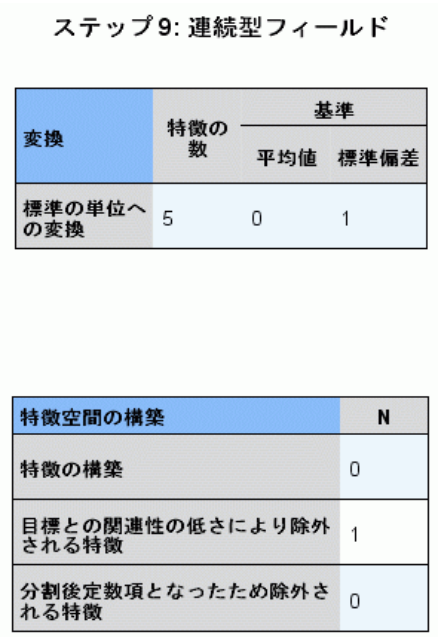
注 :この表は、変換されていないフィールドには表示されません。

表内の情報は、2 列または 3 列に分けて表示されます。

- **アクション。** アクションの名前。 (例: 連続型予測値)。 詳細は、p.46 アクションの詳細 を参照してください。
- **詳細。** 実行された処理のリスト(例: 標準単位への変換)。
- **関数。** 構築された予測値にのみ表示され、「.06*age + 1.21*height」など、入力フィールドの線型結合が表示されます。

アクションの詳細

図 4-20
ADP 分析 – アクションの詳細



ステップ 9: 連続型フィールド

変換	特徴の数	基準	
		平均値	標準偏差
標準の単位への変換	5	0	1

特徴空間の構築	N
特徴の構築	0
目標との関連性の低さにより除外される特徴	1
分割後定数項となったため除外される特徴	0

［アクションの概要］メイン ビューで下線の付いた［**アクション**］を選択した場合に表示されます。［アクションの詳細］リンク ビューには、実行された各アクションのアクション固有の情報およびおよび共通情報が表示されます。アクション固有の詳細情報が最初に表示されます。

各アクションについて、説明が、リンク ビューの一番上にタイトルとして表示されます。アクション固有の詳細がタイトルの下に表示され、派生予測値数、フィールドの再計算、目標の変換、結合または並べ替えられたカテゴリ、構築または除外された予測値の詳細が含まれる場合があります。

各アクションが処理されるごとに、予測値が除外されたり結合されたりするなどの処理中に使用される予測値数が変わる場合があります。

注 :アクションが無効になった場合、または指定された目標がなかった場合、［アクションの概要］メイン ビューでアクションがクリックされた場合、アクションの詳細の代わりにエラー メッセージが表示されます。

アクション数は 9 つですが、すべての分析で、すべての処理が行われるわけではありません。

テキスト フィールド テーブル

テーブルには、次の数が表示されます。

- 分析から除外された予測値

日付および時刻の予測値テーブル

テーブルには、次の数が表示されます。

- 日付および時刻予測値から算出した期間
- 日付および時刻の要素
- 派生した日付および時刻の予測値の合計

期間（日数）が計算された場合、基準日または基準時刻が脚注として表示されます。

予測値のスクリーニング テーブル

テーブルには、処理から除外された次の予測値の数が表示されます。

- 定数
- 欠損値の多い予測値
- 単一カテゴリのケース数が多い予測値
- カテゴリ数の多い名義型フィールド（セット）
- 除外された予測値の合計

測定レベルの確認テーブル

テーブルには再計算されたフィールド数を、次の項目に分けて表示します。

- 連続型として計算された順序型フィールド（順序セット型）
- 順序型フィールドとして計算された連続型フィールド
- 再計算の合計

連続型または順序型である入力フィールド（目標または予測値）がない場合、脚注に表示されます。

外れ値テーブル

テーブルには、外れ値の処理方法の数が表示されます。

- ［設定］タブの［入力と目標の準備］パネルの設定に応じて、外れ値が検出されトリム化された連続型フィールドの数、または外れ値が検出され欠損値に設定された外れ値の連続型フィールドの数。
- 外れ値を処理した後定数項となったために除外される連続型フィールドの数。

1 つの脚注には外れ値の分割値、連続型である入力フィールド（目標または予測値）がない場合、別の脚注が表示されます。

欠損値テーブル

テーブルには欠損値を置換したフィールド数を、次の項目に分けて表示します。

- 目標。目標が指定されていない場合、この行は表示されません。
- 予測値。名義型（セット型）、順序型（順序セット型）、連続型に分割して表示されます。
- 置換された欠損値の合計数。

目標テーブル

テーブルには、目標が変換されたかどうかについて、次のように表示されます。

- 正規性への Box-Cox 変換。指定の基準（平均および標準偏差）およびラムダを示す列に分割されます。
- 安定性を向上させるために並べ替えられた目標カテゴリ。

カテゴリ型予測値 テーブル

テーブルには、次のようなカテゴリ型予測値の数が表示されます。

- 安定性を向上させるためにカテゴリが最小から最大に並べ替えられている。
- 目標との関連性を最大化するためにカテゴリが結合されている。
- まばらなカテゴリを処理するためにカテゴリが結合されている。
- 目標との関連性の低さにより除外されている。
- 結合後定数項となったため除外されている。

カテゴリ型予測値がない場合、脚注が表示されます。

連続型予測値 テーブル

テーブルには、2 種類があります。一方のテーブルには、次のような変換フィールドの数からいずれかが表示されます。

- 標準の単位に変換された予測値。また、変換された予測値の数、指定された平均値、標準偏差が表示されます。
- 共通範囲にマッピングされた予測値。また、指定された最小値や最大値のほか、min-max 変換を使用して変換された予測値数も表示されます。
- 分割された予測値と分割された予測値数。

もう一方のテーブルには、予測値スペース構築の詳細が、次のような予測値数で表示されます。

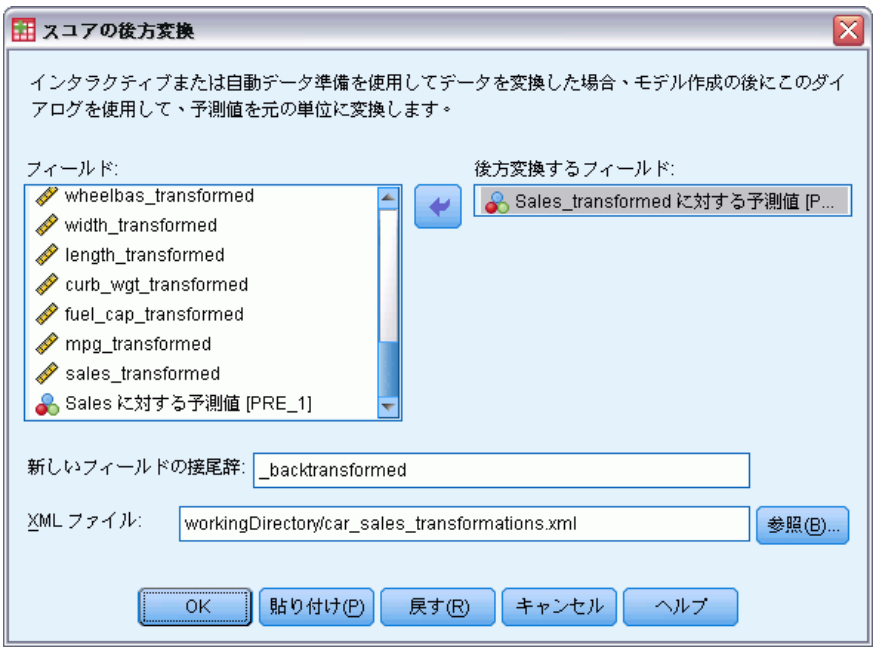
- 構築済み。
- 目標との関連性の低さにより除外されている。
- 分割後定数項となったため除外されている。
- 構築後定数項となったため除外されている。

入力となっている連続型予測値がない場合、脚注が表示されます。

スコアの後方変換

目標が ADP によって変換されている場合、変換された目標を使用して作成された後続のモデルは、変換された単位をスコアリングします。結果を解釈して使用するために、予測値を元の尺度に変換する必要があります。

図 4-21
スコアの後方変換



スコアを後方変換するには、メニューから次の項目を選択します。

変換(T) > モデル作成のデータ準備 > スコアの後方変換...

- 後方変換するフィールドを選択してください。このフィールドには、変換された目標のモデル予測値が入力されている必要があります。
- 新規変数の接尾辞を指定します。この新しいフィールドには、変換前の目標の元の尺度でモデル予測値が入力されている必要があります。
- ADP 変換を含む XML ファイルの場所を指定します。インタラクティブ データ準備または自動データ準備のダイアログで保存したファイルでなければなりません。 詳細は、 p.33 変換の適用と保存 を参照してください。

章 5

例外ケースの特定

異常検知手続きは、クラスタ グループのノルムからの偏差に基づいて異常ケースを検索します。この手続きは、推論的データ分析の前に、探索的データ分析手順において、データ監査の目的で異常ケースをすばやく検索するように設計されています。このアルゴリズムは、汎用的な異常検知用に設計されています。つまり、この異常ケースの定義は、医療産業における異常な支払いパターンの検知や、金融業会におけるマネー ロンダリングの検知など、異常の定義を正確に定義できる特定の応用例に固有のものではありません。

例: 脳卒中の治療結果に関する予測モデルは、異常な観測値の影響を受けやすいため、モデルを作成するデータ分析の担当者はデータの品質に気を使います。こうした異常な観測値の中には、非常に特異なケースを表しているため予測に使用するのは適当でないものがあります。また、技術的には「正しい」値であっても、誤って入力されたために、データ検証の手続きでは検出できない観測値もあります。[例外ケースの特定] 手続きは、分析者が外れ値の取り扱いを決めることができるように、それらの外れ値を見つけて報告します。

統計量。 この手続きは、同位グループ、連続変数とカテゴリ変数の同位グループ ノルム、同位グループ ノルムの偏差に基づく異常指数、および異常と見なされるケースに最も寄与している変数の変数影響値を作成します。

データの考慮事項

データ。この手続きは、連続変数とカテゴリ変数の両方に使用できます。それぞれの行は異なる観測値を表し、それぞれの列は同位グループの基となる異なる変数を表します。出力に印を付けるためにケース識別変数をデータ ファイル内で使用することができますが、分析では使用されません。欠損値は許可されます。重み付け変数が指定されている場合は無視されます。

検知モデルは、新しい検定データ ファイルに適用することができます。検定データの要素は、学習データの要素と同じである必要があります。また、アルゴリズム設定によっては、得点付けの前にモデルを作成するために使用される欠損値の処理が検定データ ファイルに適用される場合があります。

ケースの並び順。ケースの並び順によって解が異なる可能性があることに注意してください。並び順の影響を最小限に抑えるには、ケースを無作為に並べます。特定の解の安定性を確認するには、異なる無作為な順序で並べ替えられたケースを使用していくつかの異なる解を得てください。ファイ

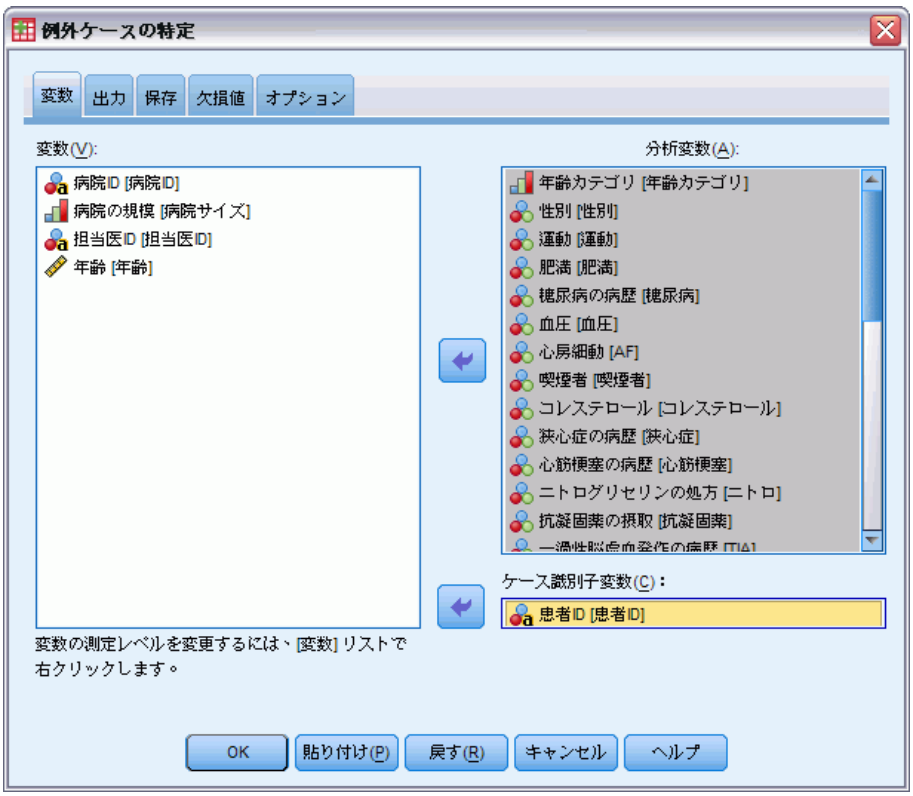
ル サイズが非常に大きい場合は、異なる無作為な順序で並べ替えられたケースのサンプルを使用し、複数回に分けて実行することができます。

仮定。このアルゴリズムは、すべての変数が一定でなく独立していることを仮定し、すべての入力変数について欠損値を持つケースがないことを仮定します。各連続変数は正規分布であると仮定し、各カテゴリ変数は多項分布であると仮定します。経験的内部検定は、この手続きが独立および分布仮定の違反に対して堅牢であることを示していますが、これらの仮定がどの程度満たされているか把握するようにしてください。

例外ケースを特定するには

▶ メニューから次の項目を選択します。
データ > 例外ケースの特定(I)...

図 5-1
[例外ケースの特定] ダイアログ ボックスの [変数] タブ



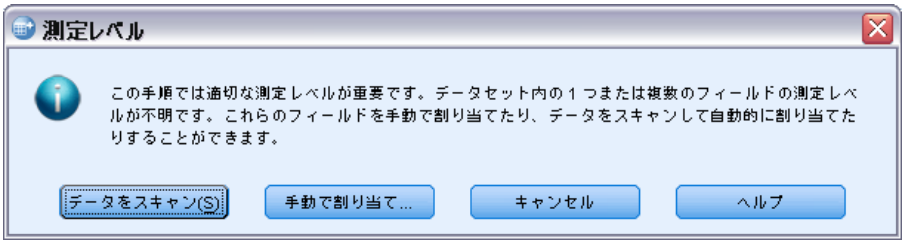
▶ 最低 1 つの分析変数を選択します。

▶ オプションで、出力のラベル付けに使用するケース識別変数も選択できます。

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 5-2
尺度の警告

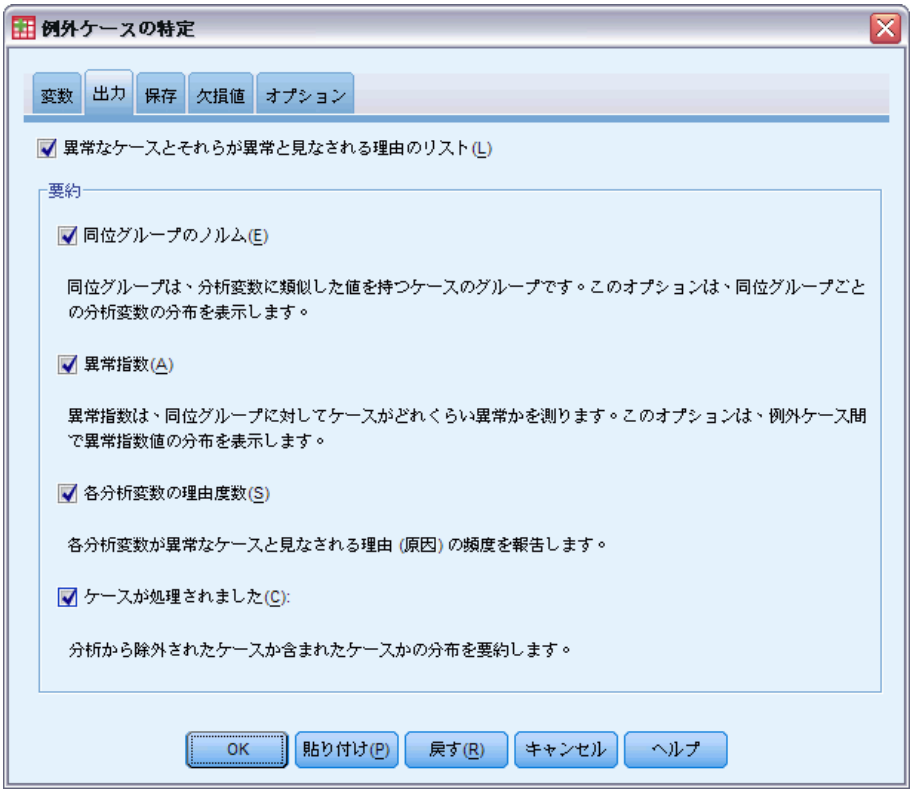


- **データをスキャン。** アクティブ データセットのデータを読み込み、デフォルトの尺度を尺度が現在不明なフィールドに割り当てます。データセットが大きい場合は時間がかかります。
- **手動で割り当てる。** 不明な尺度のフィールドをすべて表示するダイアログが開きます。このダイアログを使用して、尺度をこれらのフィールドに割り当てることができます。データ エディタの［変数ビュー］でも、尺度を割り当てることができます。

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。

[例外ケースの特定] の [出力]

図 5-3
[例外ケースの特定] ダイアログ ボックスの [出力] タブ



異常なケースとそれらが異常と見なされる理由のリスト。このオプションは次の 3 つの表を作成します。

- 異常ケースの指数リストは、異常と見なされたケースとその異常指数値を表示します。
- 異常ケース同位 ID リストは、例外ケースとどの同位グループに関する情報を表示します。
- 異常理由リストは、ケース番号、理由変数、変数影響値、変数の値、および理由ごとの変数のノルムを表示します。

すべての表は、異常指数で降順に並べ替えられます。さらに、[変数] タブでケース識別変数が指定されている場合は、ケースの ID が表示されます。

要約。このグループのコントロールは分布の要約を作成します。

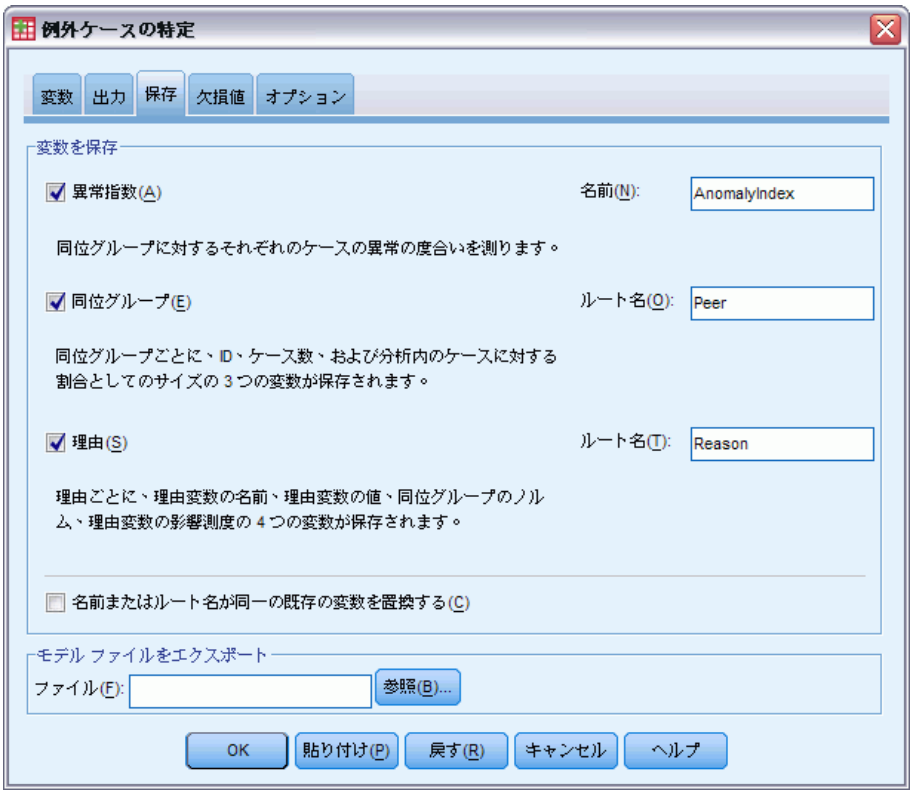
- **同位グループのノルム。**このオプションを選択すると、[連続変数ノルム] 表（分析で連続変数が使用されている場合）または [カテゴリ変数ノルム] 表（分析でカテゴリ変数が使用されている場合）を表示できます。

［連続変数ノルム］表には、同位グループごとに、各連続変数の平均偏差および標準偏差が表示されます。また［カテゴリ変数ノルム］表には、同位グループごとに、各カテゴリ変数の最頻値（度数が最も大きいカテゴリ）、度数、および度数パーセントが表示されます。連続変数の平均とカテゴリ変数の最頻値は、分析のノルム値として使用されます。

- **異常指数。**異常指数の要約には、異常度が最も高いと判定されたケースの異常指数の記述統計量が表示されます。
- **各分析変数の理由度数。**それぞれの理由に対し、各変数が理由として出現する頻度およびその割合（パーセント）がこの表に表示されます。また、この表は、それぞれの変数の影響の記述統計量を報告します。［オプション］タブで理由の最大数が 0 に設定されている場合、このオプションは使用できません。
- **処理されたケース。**処理されたケースの要約には、アクティブなデータセットにおけるすべてのケースの回数と回数パーセント、分析に組み込まれたケースと除外されたケース、および各同位グループのケースが表示されます。

[例外ケースの特定] の [保存]

図 5-4
[例外ケースの特定] ダイアログ ボックスの [保存] タブ



変数を保存。このグループにあるオプションを選択することにより、モデル変数をアクティブなデータセットに保存できます。また、保存する変数と同じ名前の既存の変数を置き換えることもできます。

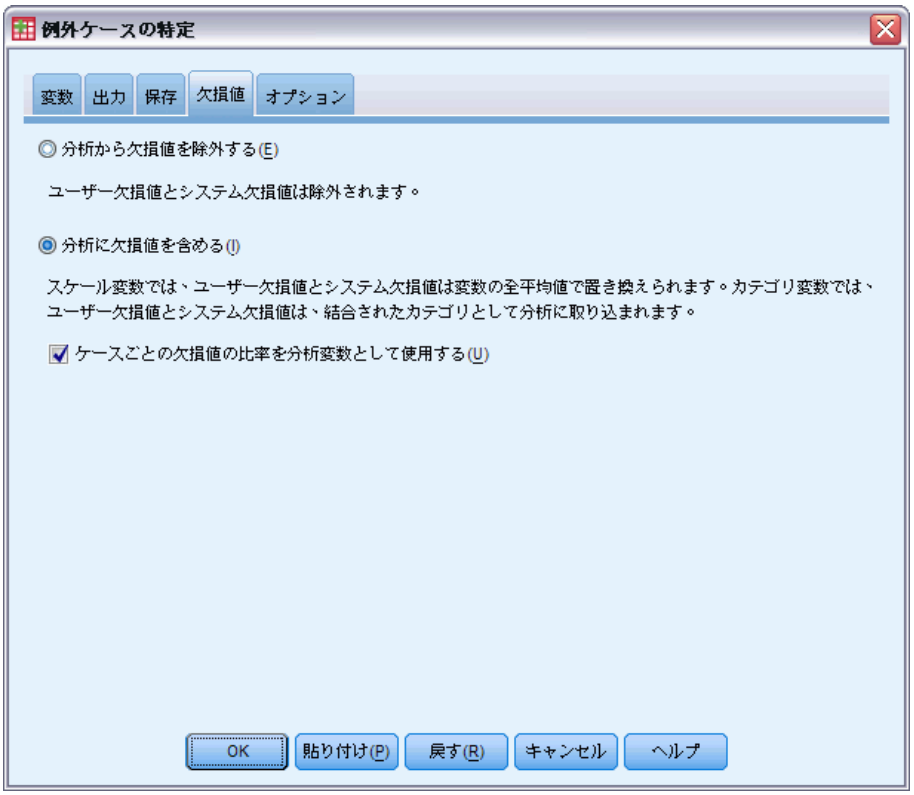
- **異常指数。**各ケースについて異常指数値を指定された名前の変数に保存します。
- **同位グループ。**ケースごとに、同位グループの ID、ケース度数、および割合（パーセント）として表されたサイズを、指定されたルート名の変数に保存します。たとえば、ルート名 Peer が指定された場合、Peerid、PeerSize、および PeerPctSize の各変数が生成されます。Peerid はケースの同位グループ ID、PeerSize はグループのサイズ、PeerPctSize はグループのサイズの割合です。
- **理由。**理由変数のグループを指定されたルート名で保存します。理由変数のグループは、理由となる変数の名前、変数の影響測度、変数の値、およびノルム値で構成されます。グループの数は、[オプション] タブで要求された理由の数によって変わります。たとえば、ルート名 Reason が指定された場合、ReasonVar_k、ReasonMeasure_k、

ReasonValue_k、および ReasonNorm_k の各変数が生成されます。ここで、k は k 番目の理由であることを表します。理由の数が 0 に設定されている場合は、このオプションを使用できません。

モデル ファイルをエクスポート。モデルを XML 形式で保存します。

[例外ケースの特定] の [欠損値]

図 5-5
[例外ケースの特定] ダイアログ ボックスの [欠損値] タブ

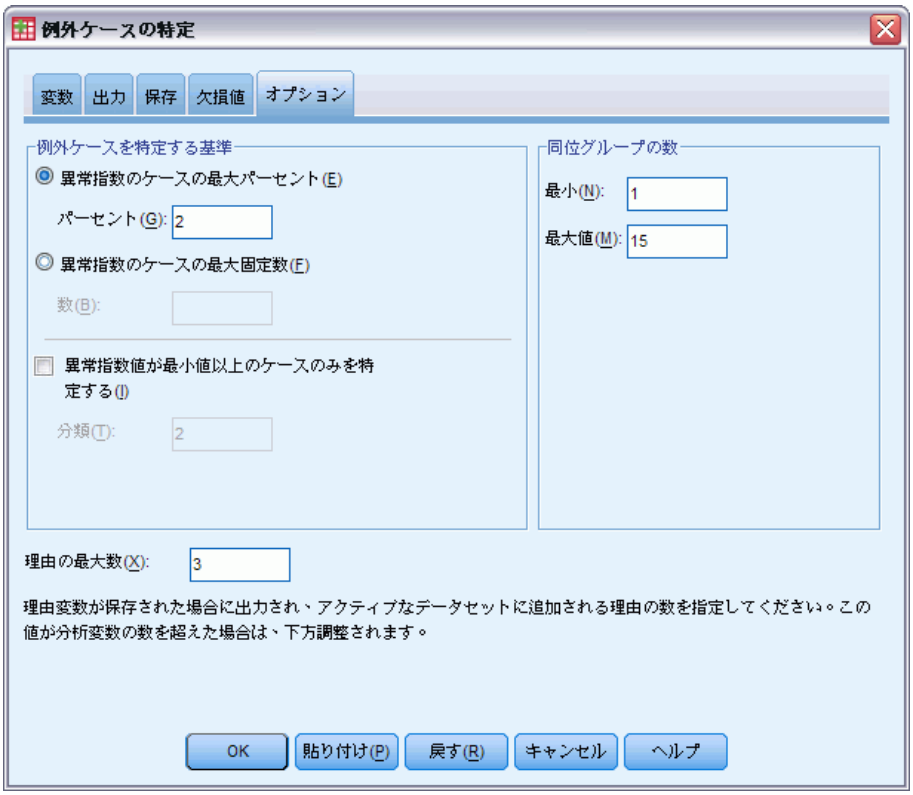


[欠損値] タブは、ユーザー欠損値とシステム欠損値の処理方法を制御するために使用します。

- **分析から欠損値を除外する。**欠損値を持つケースは分析から除外されます。
- **分析に欠損値を含める。**連続変数の欠損値には対応する全平均が代入されます。また、カテゴリ変数の欠損カテゴリはグループ化されて有効なカテゴリとして扱われます。そして処理された変数は分析で使用されます。必要であれば、ケースごとの欠損値の比率を表す追加の変数の作成を要求し、その変数を分析で使用することもできます。

[例外ケースの特定] オプション

図 5-6
[例外ケースの特定] ダイアログ ボックスの [オプション] タブ



例外ケースを特定する基準。これらの選択項目によって異常リストに含まれるケースの数が決まります。

- **異常指数が最高のケースのパーセント。**100 以下の正数を指定します。
- **異常指数のケースの最大固定数。**アクティブなデータセット内のケースのうち、分析で使用されるケースの総数を超えない正の整数を指定します。
- **異常指数値が最小値以上のケースのみを特定する。**負でない整数を指定します。ケースの異常指数値が指定された打ち切り点以上の場合、そのケースは異常と見なされます。このオプションは、[**ケースのパーセント**] オプションおよび [**ケースの固定数**] オプションと共に使用されます。たとえば、ケースの固定数として 50 を指定し、打ち切り値として 2 を指定した場合、異常リストは最大で 50 個の異常指数値が 2 以上のケースによって構成されます。

同位グループの数。手続きは、指定された最小値と最大値の間の数のグループを検索します。これらの値は正の整数である必要があり、最小値は最大値以下の値である必要があります。指定された値が等しいとき、手続きは固定数の同位グループを仮定します。

注 :データ内の変動の量によっては、データがサポートできる同位グループの数が、指定された最小値より小さくなる場合もあります。そのような状況では、手続きが作成する同位グループが少なくなる場合があります。

理由の最大数。理由は、変数の影響測度、この理由の変数の名前、変数の値、および対応する同位グループの値で構成されます。負でない整数を指定してください。この値が、分析で使用し処理された変数の数以上である場合、すべての変数が表示されます。

DETECTANOMALY コマンドの追加機能

コマンド シンタックス言語を使用して、次のことも実行できます。

- すべての分析変数を明示的に指定しないでアクティブなデータセット内のいくつかの変数を除外する（EXCEPT サブコマンドを使用）。
- 連続変数とカテゴリ変数の影響を均衡させるための調整値を指定する（CRITERIA サブコマンドで MLWEIGHT キーワードを使用）。

複雑なシンタックス情報については、「コマンド シンタックス リファレンス」を参照してください。

章
6

最適カテゴリ化

［最適カテゴリ化］手続きは、各スケール変数の値をビンに分配して、1 つ以上のスケール変数（以下 **ビン（分割）入力変数**と呼びます）を離散化するためのものです。ビンの構成は、ビン分割プロセスを「監視」するカテゴリ ガイド変数に基づいて最適化されます。元のデータ値の代わりにビンを使用することで、より詳しい分析ができます。

例。次に示すように、1 つの変数が取りうる値の個数を減らすことには、有用な点が数多くあります。

- 他の手続きに必要なデータ要件を満たすことができます。離散化された変数は、カテゴリ型として扱うことができるため、カテゴリ変数を必要とする手続きに使用できます。たとえば［クロス集計表］手続きでは、すべての変数がカテゴリ型であることが必要です。
- データの内容を秘匿することができます。値をレポートする際、実際の値の代わりにビンに分割された値を使用することで、データ ソースの内容を秘匿できます。最適カテゴリ化の手続きでは、基準に従ってビンを選択できます。
- パフォーマンスが向上します。手続きの中には、値の個数を減らすことでより効率的に処理できるものもあります。たとえば多項ロジスティック回帰は、離散化された変数を使用することにより、処理速度を向上させることができます。
- データの完全な区切りまたは準完全な区切りが明確になります。

［最適カテゴリ化］と［連続変数のカテゴリ化］との違い。［連続変数のカテゴリ化］ダイアログ ボックスでは、いくつかの方法で、ガイド変数を使わずにビンを自動作成できます。これら「監視なし」の規則は、度数分布表などの記述統計量を生成する際には有効ですが、最終的に予測モデルを構成することが目的である場合は、最適カテゴリ化の方が方法として優れています。

出力。この手続きを使用すると、ビンの分割点および各ビン（分割）入力変数の記述統計量をまとめた表を作成できます。この他にも、ビン（分割）入力変数のビン分割された値を含むアクティブなデータセットに新しい変数を保存したり、離散化する新しいデータで使用できるように、ビン規則をコマンド シンタックスとして保存したりできます。

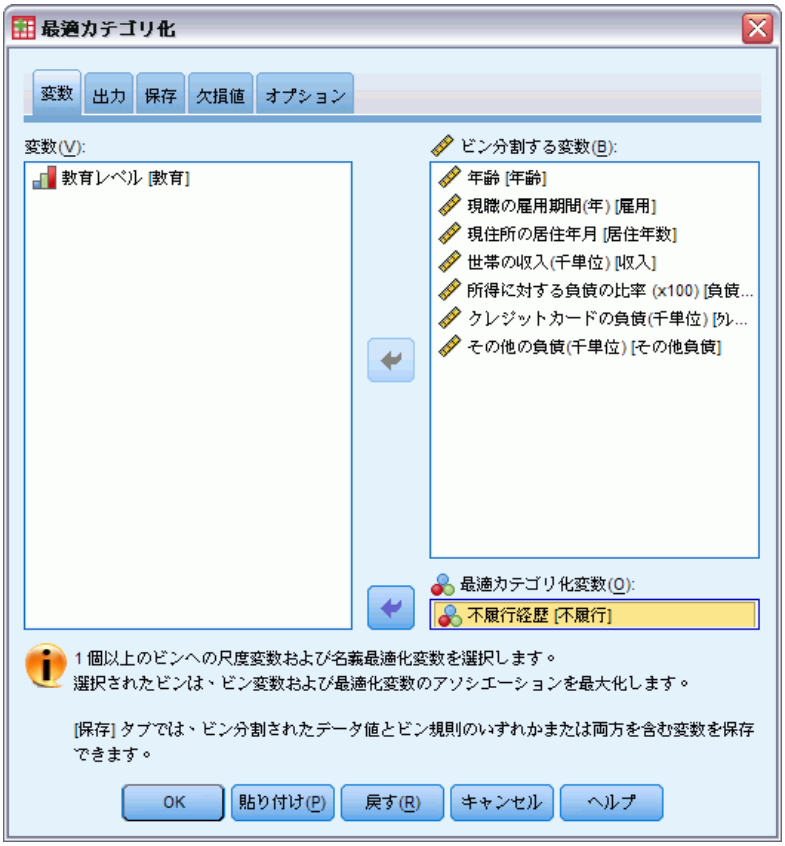
データ。この手続きでは、ビン（分割）入力変数は、数値型スケール変数であることが必要です。またガイド変数は、カテゴリ変数でなければなりませんが、数値型か文字型かは問いません。

最適カテゴリ化を行うには

メニューから次の項目を選択します。

変換 > 最適カテゴリ化...

図 6-1
[最適カテゴリ化] ダイアログ ボックスの [変数] タブ



- ビン（分割）入力変数を 1 つ以上選択します。
- ガイド変数を選択します。

ビン分割されたデータ値を含む変数は、デフォルトでは生成されません。[保存] タブで、これらの変数を保存します。

最適カテゴリ化の出力

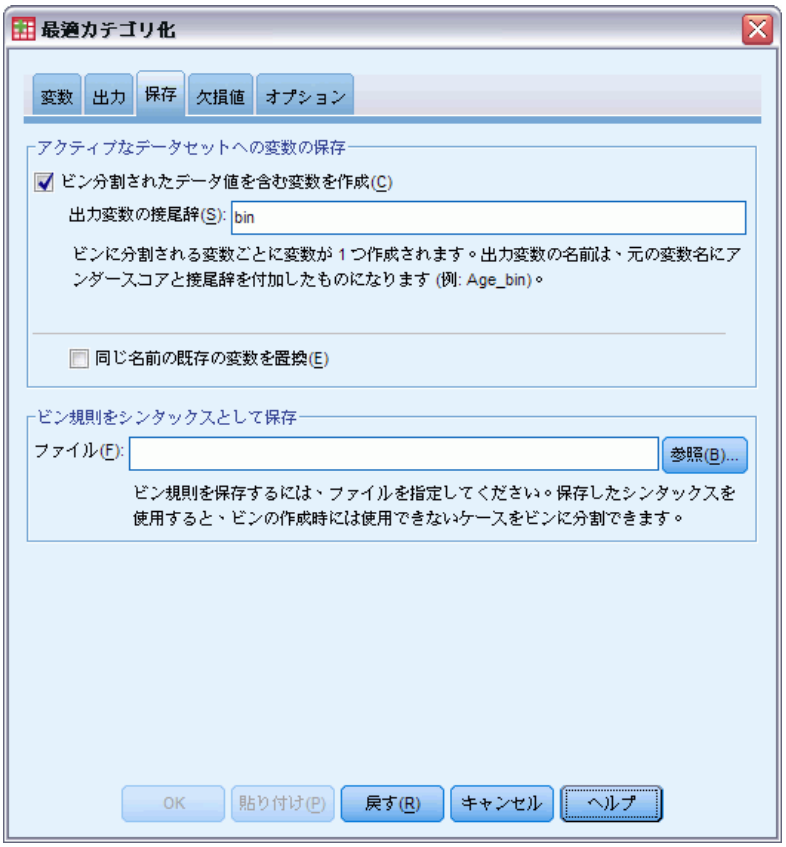
図 6-2
[最適カテゴリ化] ダイアログ ボックスの [出力] タブ

[出力] タブでは、さまざまな結果の表示を制御できます。

- **ビンの終点。**各ビン（分割）入力変数の終点を表示します。
- **ビン分割される変数の記述統計量。**各ビン（分割）入力変数に対して、有効値を持つケースの数、欠損値を持つケースの数、異なる有効値の個数、および最小値/最大値が表示されます。またガイド変数に対して、関連するビン（分割）入力変数ごとのクラス分布が表示されます。
- **ビン分割される変数のモデル エントロピー。**各ビン（分割）入力変数に対して、ガイド変数を基にした変数の予測精度の尺度が表示されます。

最適カテゴリ化の保存

図 6-3
[最適カテゴリ化] ダイアログ ボックスの [保存] タブ

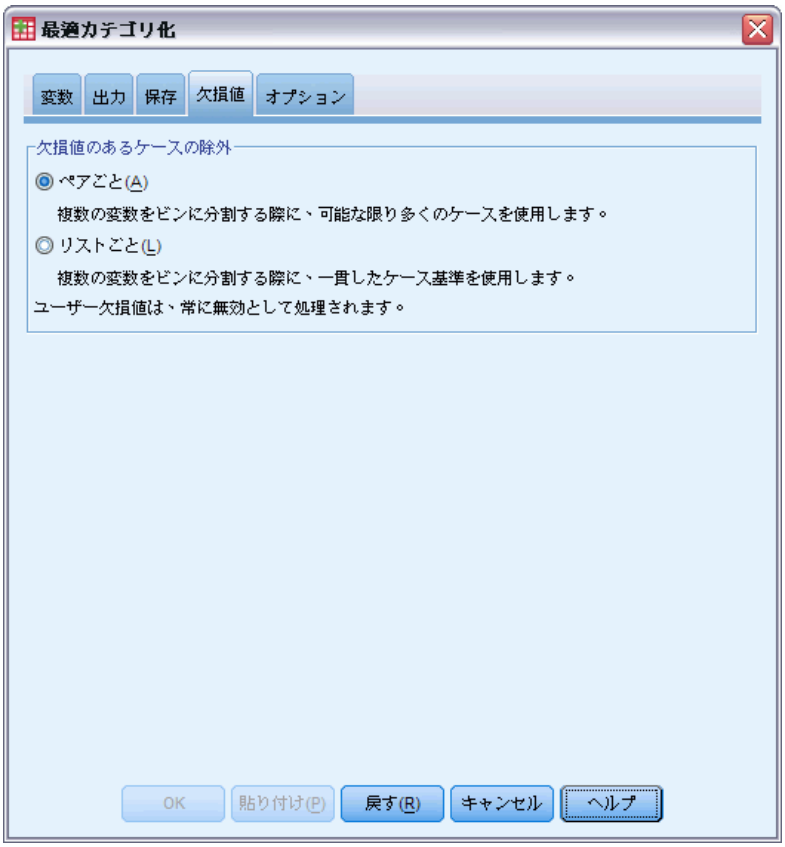


アクティブなデータセットへの変数の保存。元の変数の代わりにビン分割されたデータ値を使用することで、さらなる分析ができます。

ビン規則をシンタックスとして保存。他のデータセットをビン分割する場合に使用できるコマンド シンタックスが生成されます。再割り当て規則は、ビン分割アルゴリズムによって決定される分割点に基づきます。

最適カテゴリ化の欠損値

図 6-4
[最適カテゴリ化] ダイアログ ボックスの [欠損値] タブ

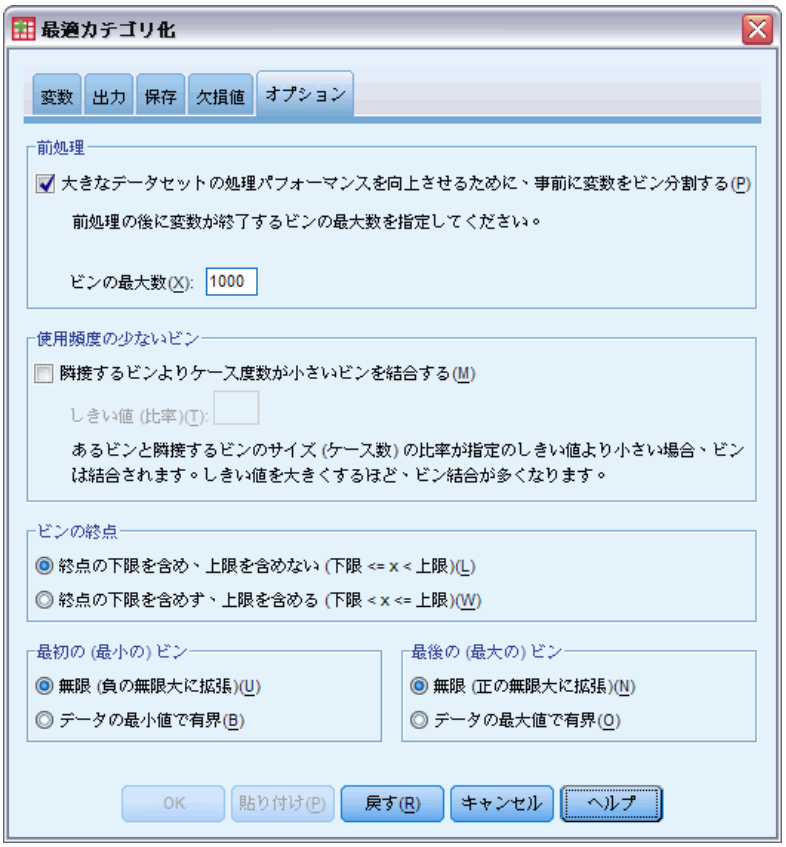


[欠損値] タブでは、欠損値を処理する場合、リストごとの削除を行うかペアごとの削除を行うかを指定できます。ユーザー欠損値は常に、無効な値として処理されます。元の変数値を新しい変数に再割り当てする場合、ユーザー欠損値はシステム欠損値に変換されます。

- **ペアごと。**このオプションは、ガイド変数とビン（分割）入力変数のペアに対して適用されます。手続きでは、ガイド変数およびビン（分割）入力変数が非欠損値であるすべてのケースが使用されます。
- **リストごと。**このオプションは、[変数] タブで指定されたすべての変数に適用されます。欠損値を持つ変数が 1 つでもあるケースは除外されます。

最適カテゴリ化のオプション

図 6-5
[最適カテゴリ化] ダイアログ ボックスの [オプション] タブ



前処理。ビン（分割）入力変数を多数の異なる値に「事前ビン分割」することにより、最終的なビンの質を大きく損なうことなく、処理時間を短縮できます。作成されるビンの数に関する上限は、ビンの最大数によって指定されます。したがって、最大数を 1000 と指定した場合、ビン（分割）入力変数の持つ異なる値の個数が 1000 未満であれば、そのビン（分割）入力変数に対して作成される前処理済みのビンの数は、ビン（分割）入力変数が持つ異なる値の個数に等しくなります。

使用頻度の少ないビン。場合によっては、手続きを通して作成されるビンのケース数が極端に少ないことがあります。本質的ではないこうした分割点は、次の方法により削除できます。

- ある変数に対して、アルゴリズムにより n_{final} 個の分割点が検出された（つまり $n_{final}+1$ 個のビンが検出された）とします。このとき、i =

2、...、n_{final} に対応するビン（値が 2 番目に小さいビンから最も大きいビンまで）に対して、次の計算を実行します。

$$\frac{sizeof(b_i)}{\min(sizeof(b_{i-1}),sizeof(b_{i+1}))}$$

sizeof(b) はビンに含まれるケースの数です。

- この値が、指定した結合しきい値未満の場合、b_i は使用頻度が低いとみなされ、b_{i-1} または b_{i+1} のうち、クラスの情報エントロピーが小さい方に結合されます。

この手続きでは、すべてのビンについて上記の一連の処理が行われます。

ビンの終点。このオプションでは、区間の下限をどのように定義するかを指定できます。分割点の値は手続きによって自動的に決定されるため、このオプションは、必要に応じて使用してください。

最初の（最小の）ビン/最後の（最大の）ビン。これらのオプションでは、各ビン（分割）入力変数に対する分割点の最小点および最大点をどのように定義するかを指定できます。手続きでは通常、ビン（分割）入力変数は実数直線上の値を取ると想定されますが、理論上または実用上の理由から範囲を制限する場合は、最小値/最大値によってその範囲を定めます。

OPTIMAL BINNING コマンドの追加機能

コマンド シンタックスを使用すると、次の作業も実行できます。

- 等度数法による監視なしカテゴリ化の実行（CRITERIA サブコマンドを使用）。

シンタックスの詳細は、『Command Syntax Reference』を参照してください。

パート II:
例

章
7

データの検証

［データの検証］手続きは、無効の疑いがあるかまたは実際に無効なケース、変数、およびデータ値を特定するためのものです。

医療データベースの検証

医療組織からデータ分析の依頼を受けた担当者は、システム内の情報の品質を管理しなければなりません。この管理では、値や変数をチェックし、データ入力チームの責任者向けのレポート作成も行います。

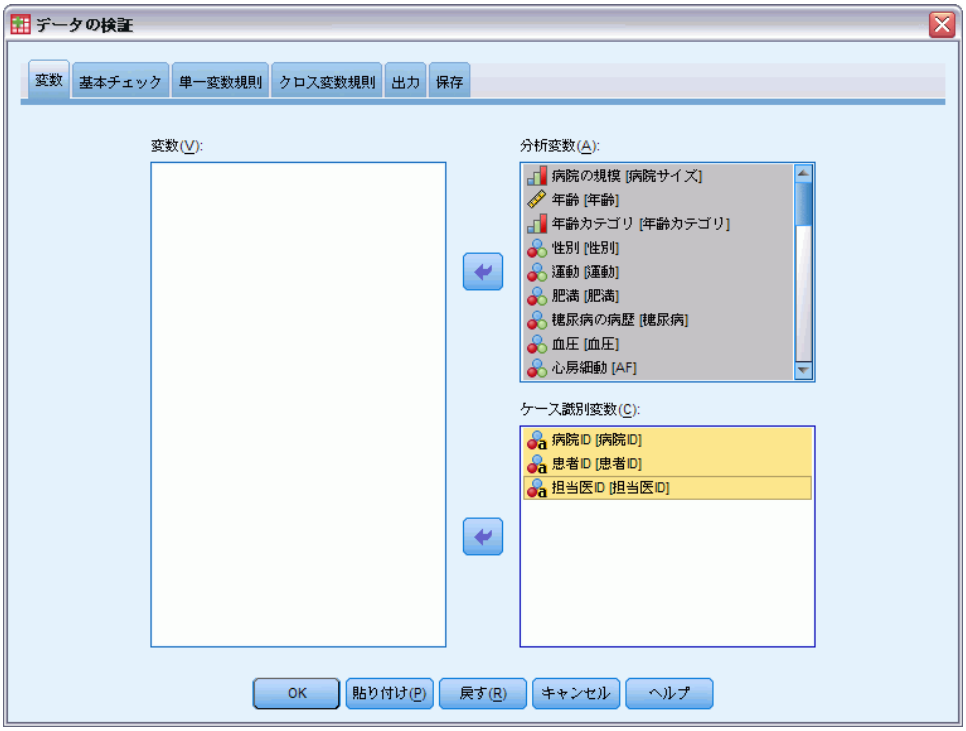
データベースの最新の状態は、stroke_invalid.sav に収集されています。詳細は、 A 付録 p.149 サンプル ファイル を参照してください。 データの検証手続きを使用すると、レポートの作成に必要な情報を取得できます。これらの分析結果を生成するためのシンタックスは、validatedata_stroke.sps にあります。

基本チェックの実行

▶ ［データの検証］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。

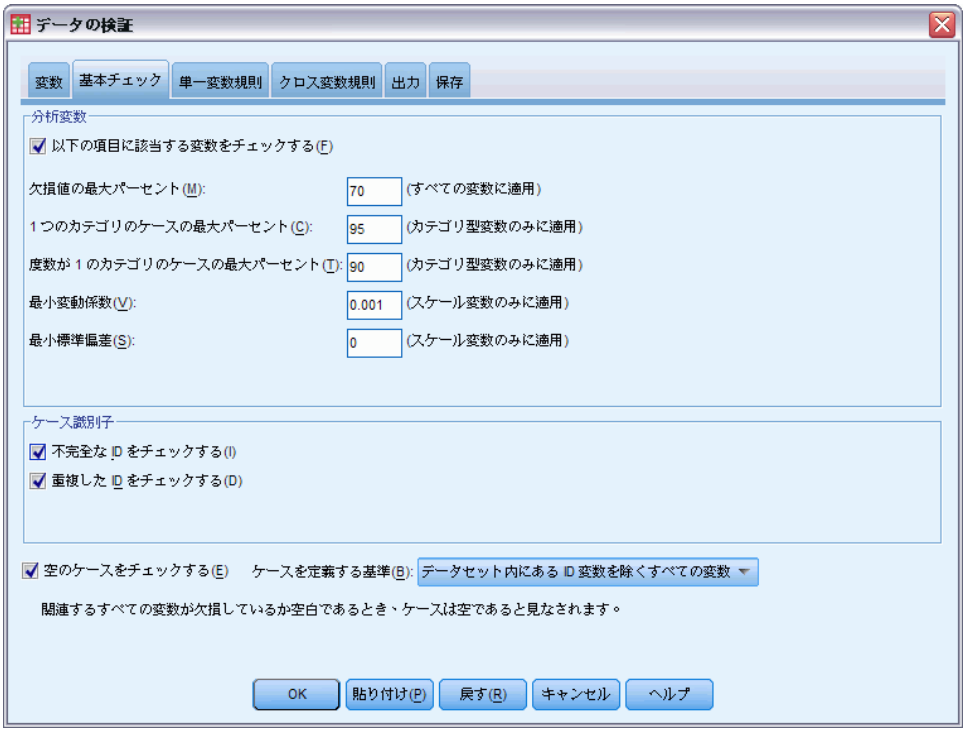
データ > **検証(V)** > **データの検証(V)...**

図 7-1
[データの検証] ダイアログ ボックスの [変数] タブ



- 分析変数として、「病院の規模」、および「年齢」から「6 か月後のレコードバーセルインデックス」までの変数を選択します。
- またケース識別変数として、「病院 ID」、「患者 ID」、および「担当医 ID」を選択します。
- [**基本チェック**] タブをクリックします。

図 7-2
[データの検証] ダイアログ ボックスの [基本チェック] タブ



デフォルトの設定は、実行に必要な内容になっています。

▶ [OK] をクリックします。

警告

図 7-3
警告

警告

すべてのケース変数またはデータ値が要求されたチェックを通ったため、要求された出力の一部またはすべては表示されません。

分析変数が基本チェックを無事通過し、空のケースも存在しない場合は、その結果としてこれらのチェックに関する出力は行われない旨の警告が表示されます。

不完全な識別子

図 7-4
不完全なケース識別子

不完全な識別子

	識別子		
ケース	病院ID	患者ID	担当医ID
288	OZN		125304
573		6137798782	790697
774		2322241867	176466

ケース識別変数に欠損値が含まれている場合、そのケースは正しく識別されません。このデータ ファイルの場合、ケース 288 では 患者 ID が欠損しており、ケース 573 および 774 では 病院 ID が欠損しています。

重複した識別子

図 7-5
重複したケース識別子 (先頭の 11 ケース)

識別子の重複

識別子グループの重複	重複の数	重複した識別子のあるケース	識別子		
			病院ID	患者ID	担当医ID
1	2	10, 11	PBW	1406462419	355184
2	2	14, 15	PBW	2191527525	355184
3	2	21, 22	PBW	7237535360	616528
4	2	28, 29	NHV	4592215163	942982
5	2	30, 31	NHV	7628592330	371884
6	2	64, 65	NHV	0300750006	371884
7	2	83, 84	QWS	4590625286	215041
8	2	86, 87	QWS	6272818258	817329
9	2	96, 97	QWS	1959349605	215041
10	3	100, 101, 102	QWS	5856145337	817329
11	3	104, 105, 106	QWS	1543897849	817329
12	2	122, 123	QWS	9535631975	215041

ケースは、識別変数の値の組み合わせにより一意に識別されることが必要です。重複した識別子の表には、先頭の 11 エントリが表示されています。こうした重複は、複数のイベントを持つ患者が、そのイベントごとに別々

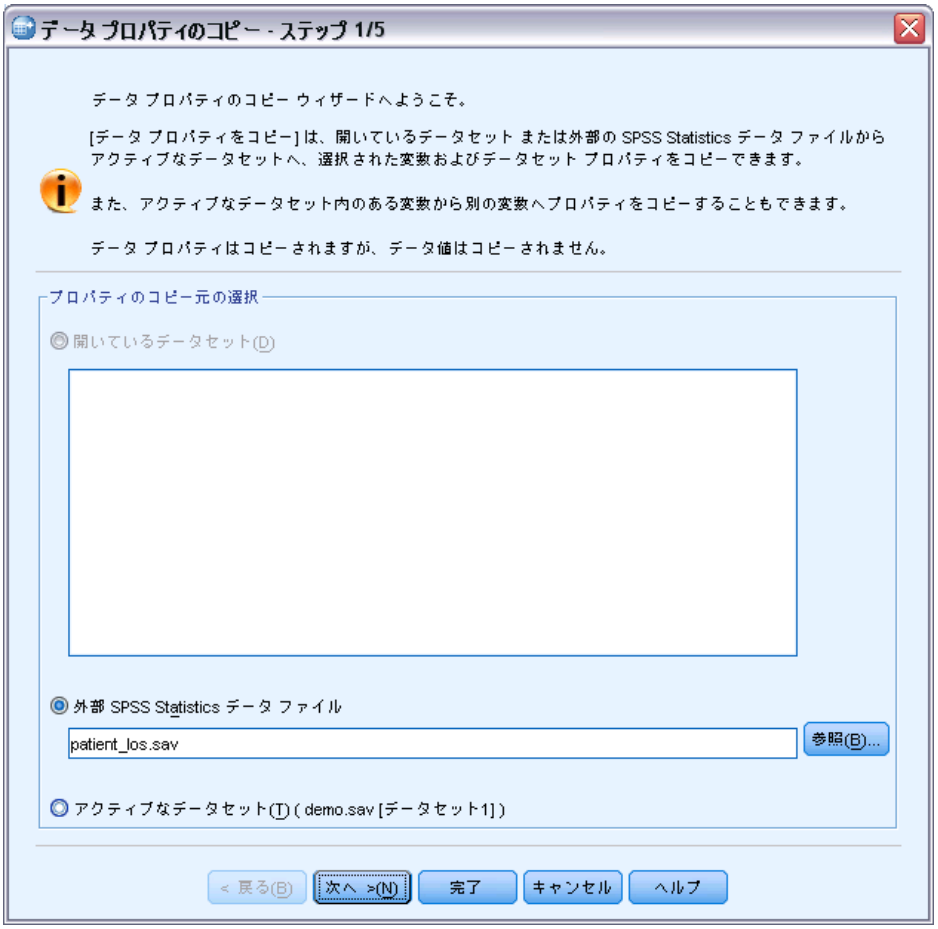
のケースとして入力されたことが原因となります。この情報は 1 つの行にまとめることができるので、こうしたケースは整理するようにします。

別のファイルにある規則をコピーして使用

現在扱っているデータ ファイル内の変数とほぼ同じ変数を持つプロジェクトがその他に見つかったとします。そのプロジェクトに対して定義されている検証規則は、関連するデータ ファイルのプロパティとして保存されているため、そのファイルのデータ プロパティをコピーすることにより、現在扱っているデータ ファイルに適用できます。

▶ 別のファイルから規則をコピーするには、メニューから次の項目を選択します。
データ > データ プロパティのコピー(C)

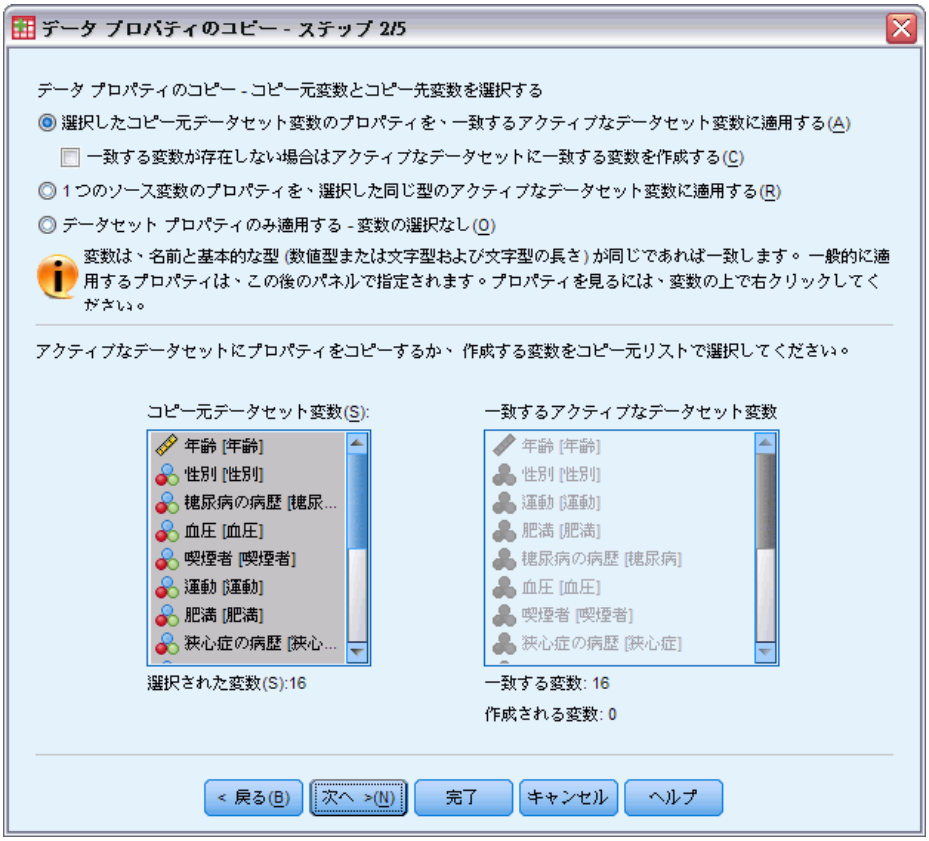
図 7-6
[データ プロパティのコピー] - ステップ 1 (ようこそ)



▶ 外部の IBM® SPSS® Statistics データ ファイル patient_los.sav からプロパティをコピーするように選択します。詳細は、 A 付録 p.149 サンプル ファイル を参照してください。

▶ [次へ] をクリックします。

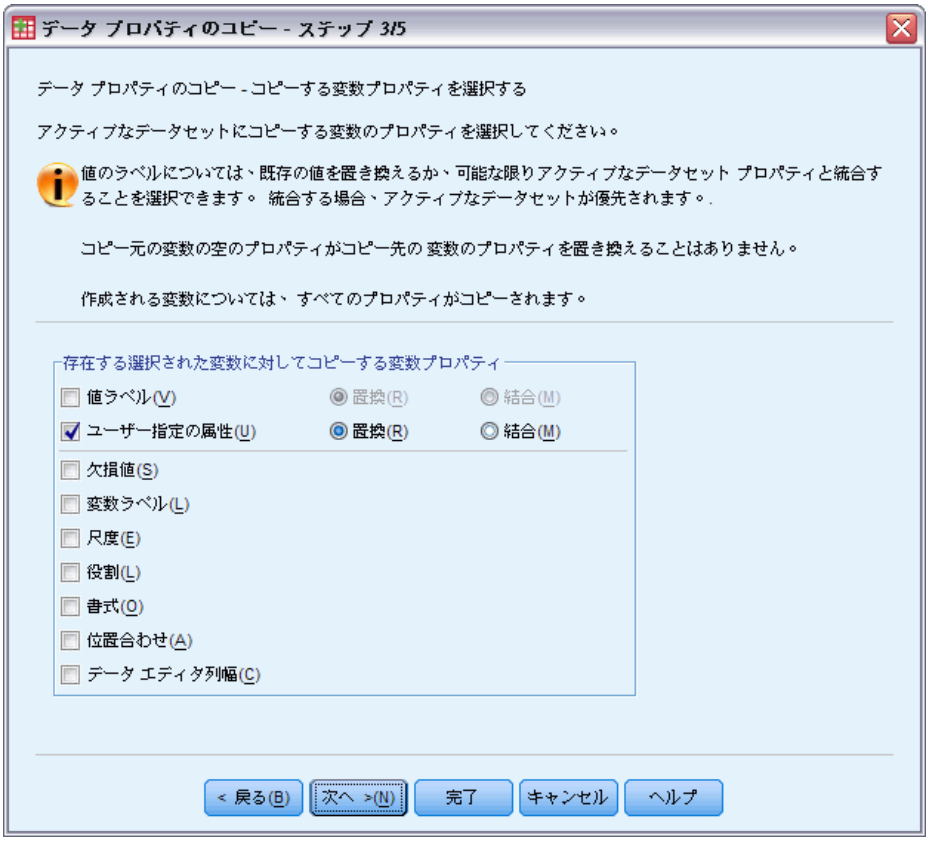
図 7-7
[データ プロパティのコピー] - ステップ 2 (変数の選択)



これらの変数を、プロパティのコピー元となる patient_los.sav から stroke_invalid.sav 内の対応する変数にコピーします。

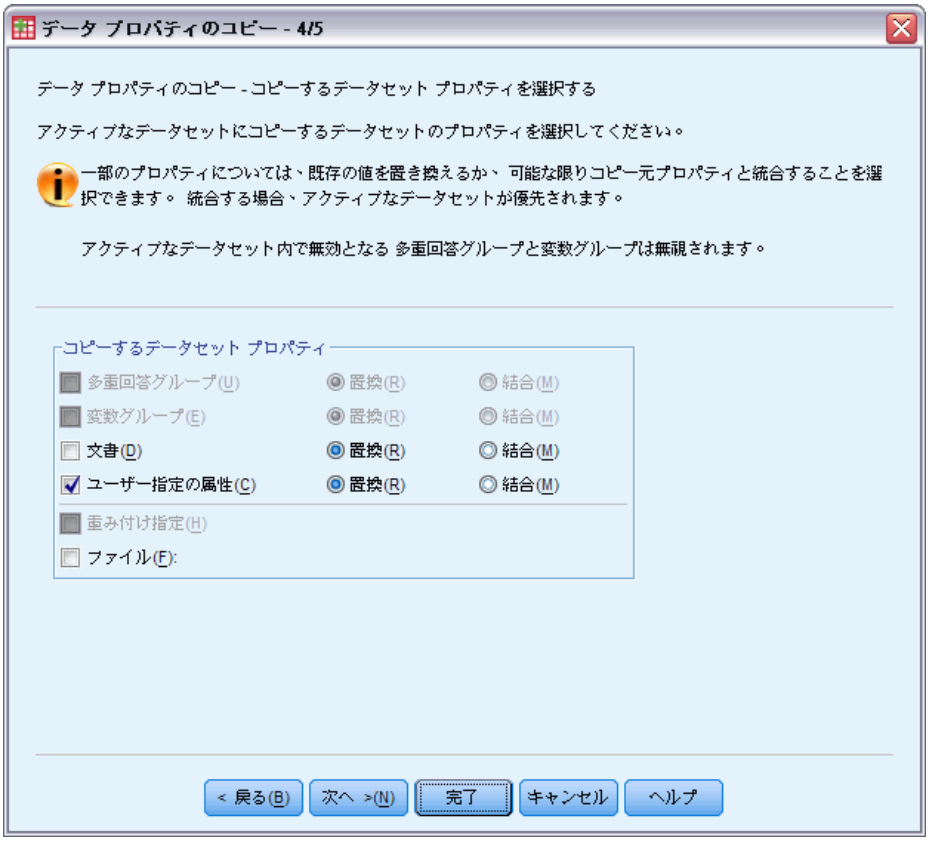
▶ [次へ] をクリックします。

図 7-8
[データ プロパティのコピー] - ステップ 3 (変数プロパティの選択)



- ▶ [**ユーザー指定の属性**] を除くすべてのプロパティの選択を解除します。
- ▶ [**次へ**] をクリックします。

図 7-9
[データ プロパティのコピー] - ステップ 4 (データセット プロパティの選択)

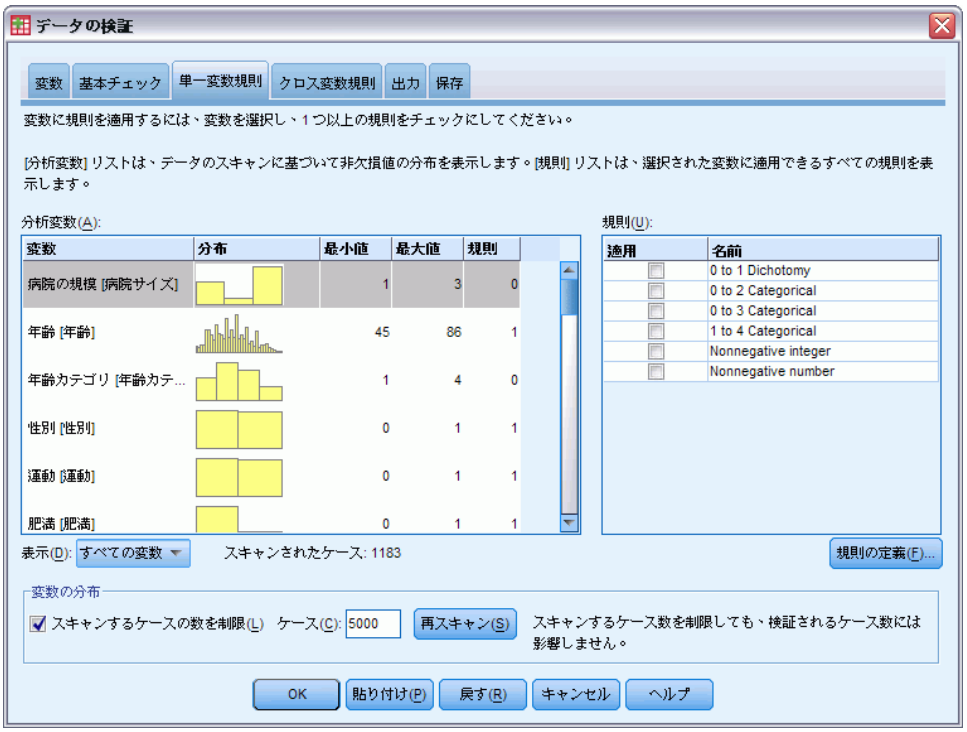


► [ユーザー指定の属性] を選択します。

► [完了] をクリックします。

これで、検証規則を再利用できるようになります。

図 7-10
[データの検証] ダイアログ ボックスの [単一変数規則] タブ



▶ コピーした規則を使用して stroke_invalid.sav のデータを検証するには、[ダイアログ リコール] ツールバー ボタンをクリックし、**[データの検証]** を選択します。

▶ **[単一変数規則]** タブをクリックします。

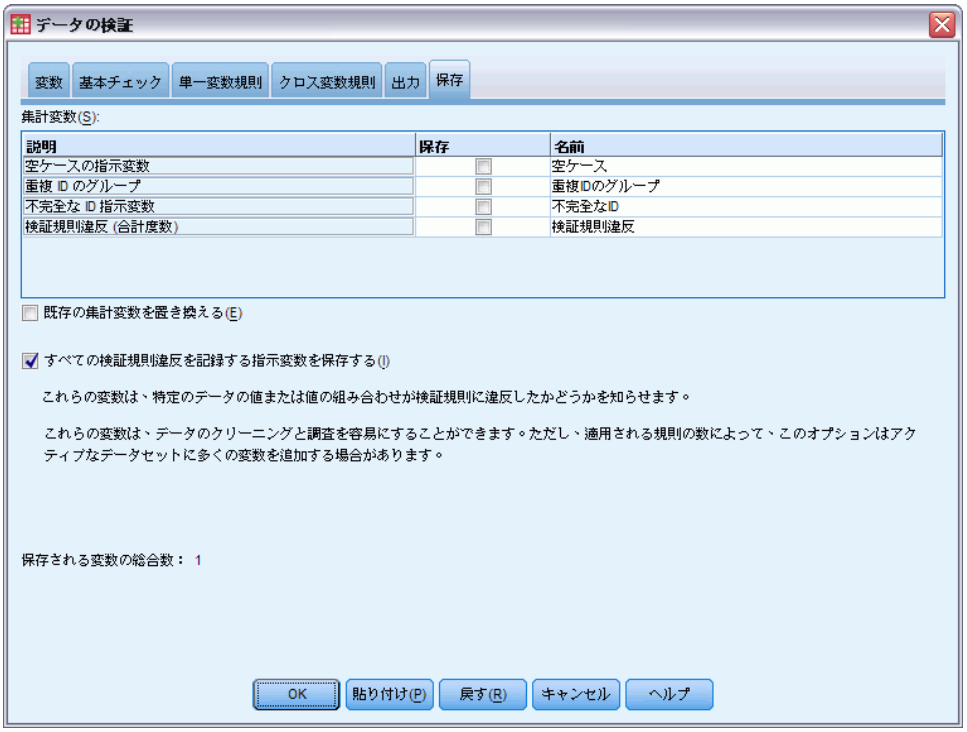
[分析変数] リストには、[変数] タブで選択された変数、それらの分布に関する要約情報、および各変数に適用された規則の数が表示されます。patient_los.sav からプロパティがコピーされた変数には、なんらかの規則が適用されています。

[規則] リストには、データ ファイルで使用できる単一変数検証規則が表示されます。これらの規則はすべて、patient_los.sav からコピーされたものです。これらの規則のいくつかは、一方のデータ ファイルの中に対応する変数が存在しない変数にも適用できます。

図 7-11
[データの検証] ダイアログ ボックスの [単一変数規則] タブ

- 「心房細動」、「一過性脳虚血発作の病歴」、「CAT スキャンの結果」、および「病院での死亡」を選択し、[0 to 1 Dichotomy] 規則を適用します。
- [0 to 3 Categorical] を「リハビリ後」に適用します。
- [0 to 2 Categorical] を「予防的処置手術後」に適用します。
- [Nonnegative integer] を「リハビリでの滞在期間」に適用します。
- [1 to 4 Categorical] を、「1 か月後のレコードバーセルインデックス」から「6 か月後のレコードバーセルインデックス」までの変数に適用します。
- [保存] タブをクリックします。

図 7-12
[データの検証] ダイアログ ボックスの [保存] タブ



- [すべての検証規則違反を記録する指示変数を保存する] を選択します。このオプションにより、単一変数規則に違反するケースと変数を結び付けやすくなります。
- [OK] をクリックします。

規則の説明

図 7-13
規則の説明

規則	説明
Nonnegative integer	型: 数値型 ドメイン型: 範囲型 ユーザー欠損値に星印を付ける: いいえ システム欠損値に星印を付ける: はい 最小値: 0 範囲内でラベルのない値に星印を付ける: いいえ 範囲内で整数でない値に星印を付ける: はい $VD.SRule[5]: 規則
0 to 1 Dichotomy	型: 数値型 ドメイン型: ケースのリスト ユーザー欠損値に星印を付ける: いいえ システム欠損値に星印を付ける: はい ケースのリスト: 0, 1 $VD.SRule[1]: 規則
0 to 2 Categorical	型: 数値型 ドメイン型: ケースのリスト ユーザー欠損値に星印を付ける: いいえ システム欠損値に星印を付ける: はい ケースのリスト: 0, 1, 2 $VD.SRule[2]: 規則

少なくとも 1 回違反した規則が表示されます。

規則の説明表には、違反のあった規則に関する説明が表示されます。この機能は、数多くの検証規則を把握するのに非常に有用です。

変数の要約

図 7-14
変数の要約

変数の要約

	規則	違反の数
年齢	0 to 1Dichotomy	1
	合計	1
性別	0 to 1Dichotomy	1
	合計	1
angina	0 to 1Dichotomy	1
	合計	1
time	0 to 1Dichotomy	2
	合計	2
doa	0 to 1Dichotomy	1
	合計	1

変数の要約表には、1 つ以上の検証規則に違反した変数、違反のあった規則、各規則に対する違反の回数、および各変数の規則ごとの違反回数が一覧として表示されます。

ケースのレポート

図 7-15
ケースのレポート

ケース報告書[b]

ケース	検証規則違反 単一変数[a]	識別子 病院ID	患者ID	担当医ID
175	0 to 1 Dichotomy (1)	OZN	0333204686	883285
274	0 to 1 Dichotomy (1)	OZN	1038840465	103254
310	Nonnegative integer (1)	OZN	2090290204	883285
437	0 to 1 Dichotomy (1)	WPA	2349729006	723384
752	Nonnegative integer (1)	GFG	4993307441	828754
1173	1 to 4 Categorical (1)	ALK	8737661990	185787

a. その規則に違反した変数の数はそれぞれの規則を満たしています。
b. 規則違反のあるケースが 100 個よりも多く存在しました。最初の 100 だけが表示されます。

ケースのレポート表には、1 つ以上の検証規則に違反したケース（ケース番号とケース ID)、違反のあった規則、および各規則に対するそのケースの違反回数が一覧として表示されます。また無効な値は、データ エディタに表示されます。

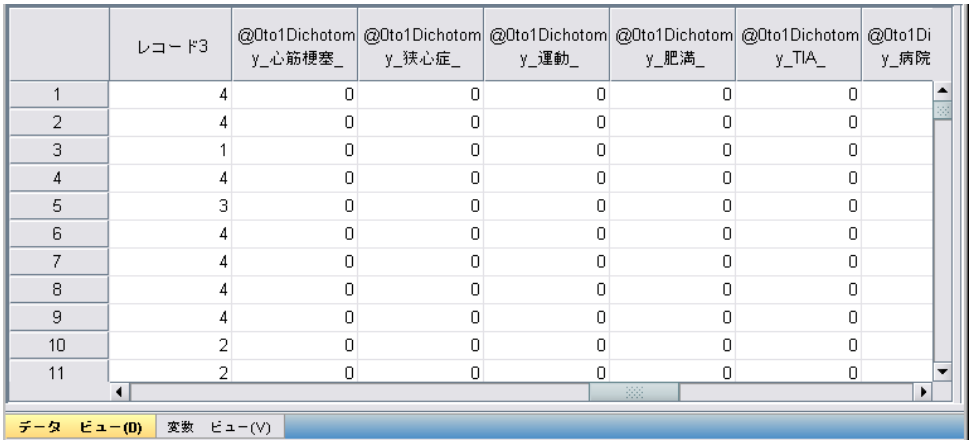
図 7-16
保存された規則違反指標を表示したデータ エディタ

	レコード3	@0to3Categorical_ 血栓溶解_	@0to3Categorical_ リハビリ_	@0to1Dichotomy_ 肥満_	@0to1Dichotomy_ 病院死亡_	@0to1Dichotomy_ TIA_	@0to2C: 手
1	4	0.00	0.00	0.00	0.00	0.00	
2	4	0.00	0.00	0.00	0.00	0.00	
3	1	0.00	0.00	0.00	0.00	0.00	
4	4	0.00	0.00	0.00	0.00	0.00	
5	3	0.00	0.00	0.00	0.00	0.00	
6	4	0.00	0.00	0.00	0.00	0.00	
7	4	0.00	0.00	0.00	0.00	0.00	
8	4	0.00	0.00	0.00	0.00	0.00	
9	4	0.00	0.00	0.00	0.00	0.00	
10	2	0.00	0.00	0.00	0.00	0.00	
11	2	0.00	0.00	0.00	0.00	0.00	

データ ビュー(D) 変数 ビュー(V)

検証規則の適用ごとに、指標変数が個別に生成されます。たとえば、@0to3Categorical_clotsolv_ は、[0 to 3 Categorical] 単一変数検証規則を、「Clot-dissolving drugs (血栓熔解薬)」変数に適用した場合に生成される指標です。与えられたケースに対して、どの変数の値が無効であるかを判別するには、指標の値をスキャンするのが最も簡単な方法です。値 1 は、関連する変数の値が無効であることを示しています。

図 7-17
ケース 175 に関する規則違反指標を表示したデータ エディタ

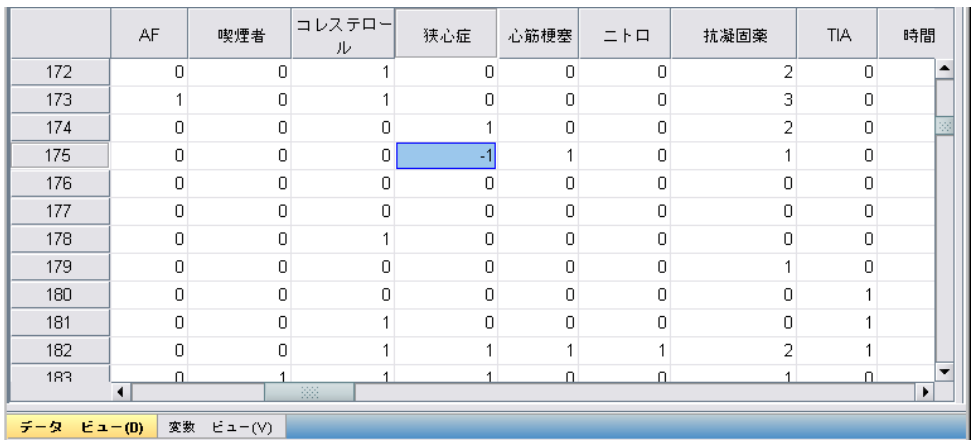


	レコード3	@0to1Dichotomy_心筋梗塞_	@0to1Dichotomy_狭心症_	@0to1Dichotomy_運動_	@0to1Dichotomy_肥満_	@0to1Dichotomy_TIA_	@0to1Diy_病院
1	4	0	0	0	0	0	
2	4	0	0	0	0	0	
3	1	0	0	0	0	0	
4	4	0	0	0	0	0	
5	3	0	0	0	0	0	
6	4	0	0	0	0	0	
7	4	0	0	0	0	0	
8	4	0	0	0	0	0	
9	4	0	0	0	0	0	
10	2	0	0	0	0	0	
11	2	0	0	0	0	0	

データ ビュー(D) 変数 ビュー(V)

規則違反のある最初のケースである、ケース 175 に移動します。変数の要約表で変数に対応する指標を確認すると、検索を効率的に行えます。狭心症の病歴に無効な値があることがすぐに確認できます。

図 7-18
狭心症の病歴について無効な値が表示されているデータ エディタ



	AF	喫煙者	コレステロール	狭心症	心筋梗塞	ニトロ	抗凝固薬	TIA	時間
172	0	0	1	0	0	0	2	0	
173	1	0	1	0	0	0	3	0	
174	0	0	0	1	0	0	2	0	
175	0	0	0	-1	1	0	1	0	
176	0	0	0	0	0	0	0	0	
177	0	0	0	0	0	0	0	0	
178	0	0	1	0	0	0	0	0	
179	0	0	0	0	0	0	1	0	
180	0	0	0	0	0	0	0	1	
181	0	0	1	0	0	0	0	1	
182	0	0	1	1	1	1	2	1	
183	0	1	1	1	0	0	1	0	

データ ビュー(D) 変数 ビュー(V)

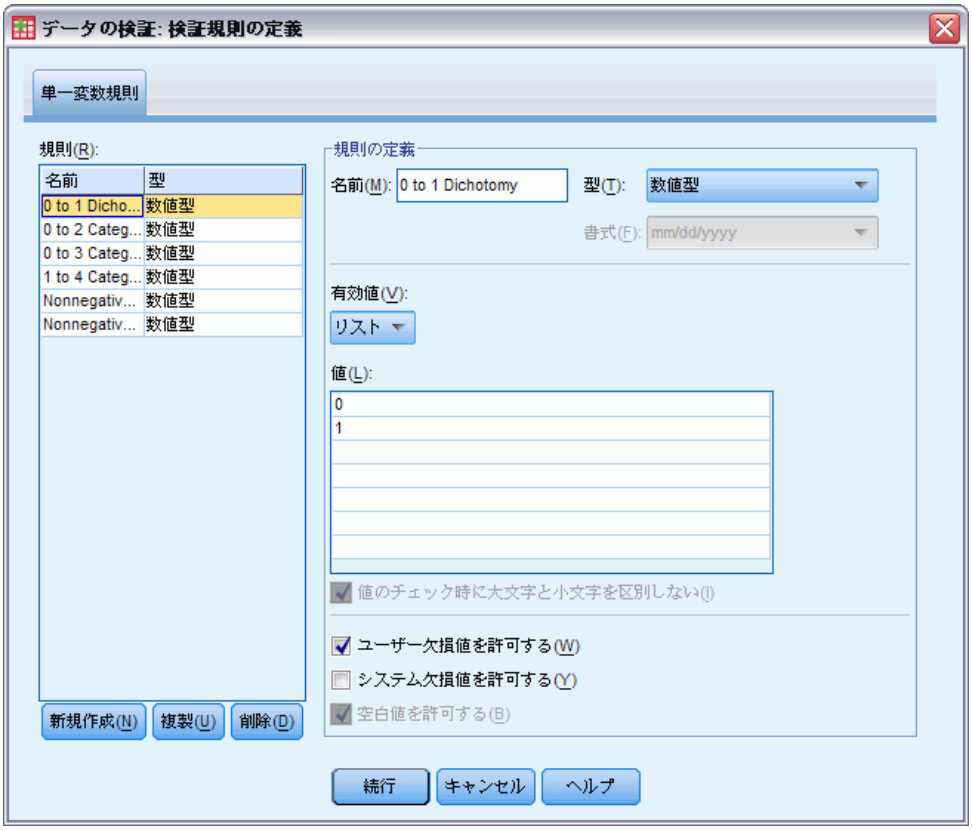
狭心症の病歴は、1 という値になります。 この値は、データ ファイル内の治療変数および結果変数に対しては有効な欠損値ですが、患者の病歴の値に対しては現在ユーザー欠損値が定義されていないため、ここでは無効になります。

独自の規則の定義

ここまでは、patient_los.sav からコピーされた検証規則を使用することが非常に有効でしたが、この作業を完了するには、さらにいくつかの規則を定義する必要があります。また、病院到着時に死亡した患者は、病院内で死亡したと誤って記録されることがあります。単一変数検証規則ではこの誤りを検出できないため、これに対応できるようにクロス変数規則を定義する必要があります。

- ［ダイアログ リコール］ツールバー ボタンをクリックし、[**データの検証**]を選択します。
- [**単一変数規則**]タブをクリックします。(病院の規模、ランキンスコアを測定するための変数、および記録されていないバーセルインデックスに対応する変数についての各規則を定義する必要があります)。
- [**規則の定義**]をクリックします。

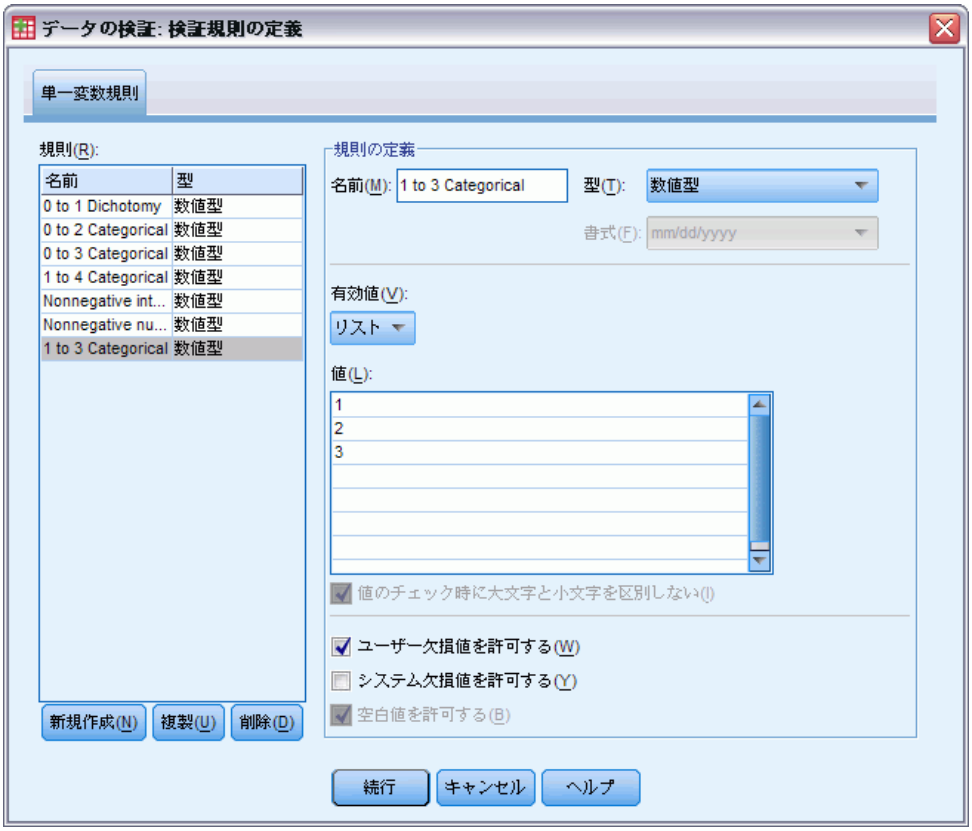
図 7-19
[検証規則の定義] ダイアログ ボックスの [単一変数規則] タブ



現在定義されている規則が表示されます。[規則] リストでは [0 to 1 Dichotomy] が選択され、[規則の定義] グループにその規則のプロパティが表示されています。

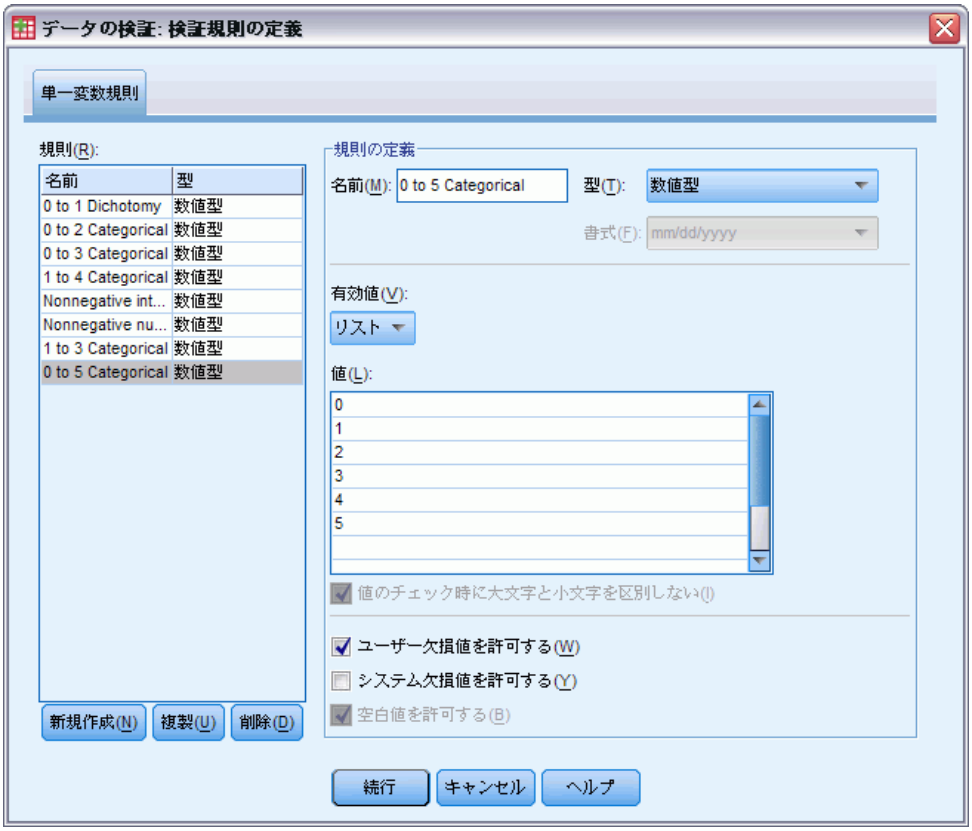
▶ 規則を定義するには、[新規] をクリックします。

図 7-20
[検証規則の定義] ダイアログ ボックスの [単一変数規則] タブ ([1 to 3 Categorical] が定義された場合)



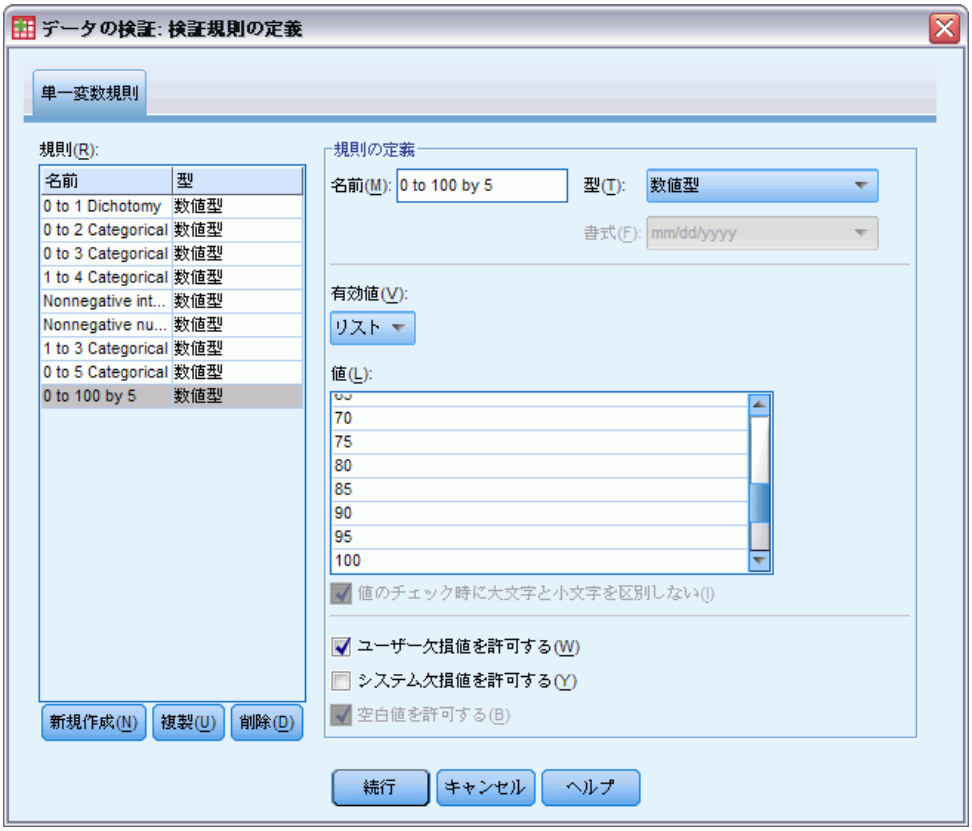
- 規則名に「1 to 3 Categorical」と入力します。
- [有効値] で、[リスト] を選択します。
- 値として、「1」、「2」、および「3」を入力します。
- [システム欠損値を許可する] の選択を解除します。
- ランキンスコアに対する規則を定義するには、まず [新規] をクリックします。

図 7-21
[検証規則の定義] ダイアログ ボックスの [単一変数規則] タブ ([0 to 5 Categorical] が定義された場合)



- 規則名として、「0 to 5 Categorical」と入力します。
- [有効値] で、[リスト] を選択します。
- 値として、「0」、「1」、「2」、「3」、「4」、および「5」を入力します。
- [システム欠損値を許可する] の選択を解除します。
- バーセルインデックスに対する規則を定義するには、まず [新規] をクリックします。

図 7-22
[検証規則の定義] ダイアログ ボックスの [単一変数規則] タブ ([0 to 100 by 5 defined] が定義された場合)



- 規則名として、「0 to 100 by 5」と入力します。
- [有効値] で、[リスト] を選択します。
- 値として、「0」、「5」、. . . 、「100」 を入力します。
- [システム欠損値を許可する] の選択を解除します。
- [続行] をクリックします。

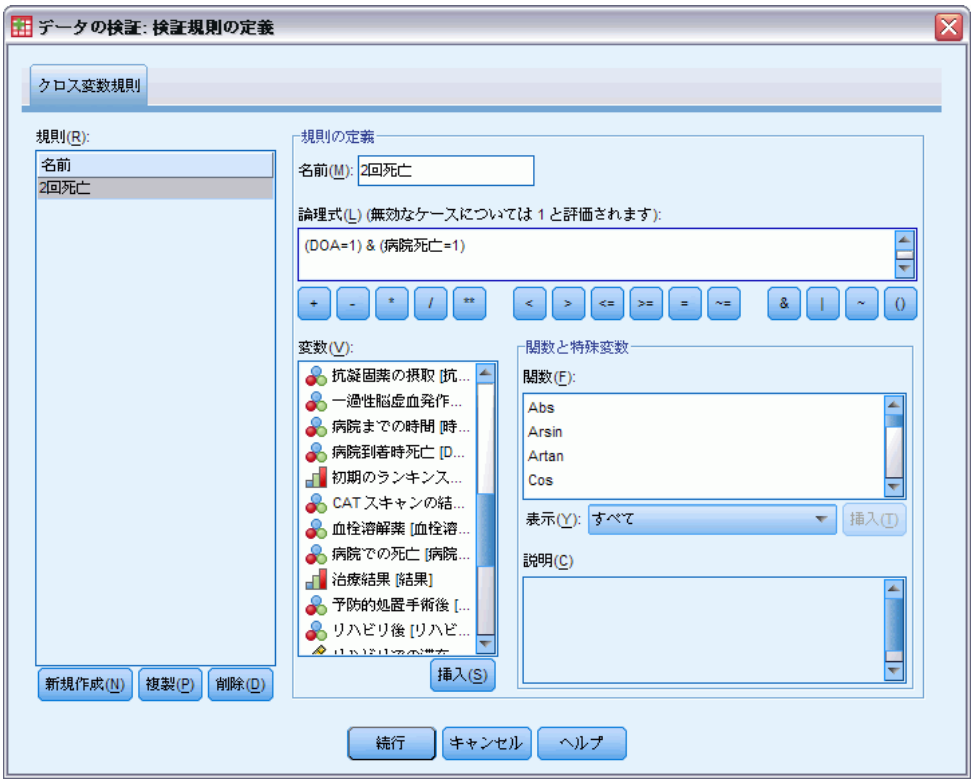
図 7-23
[データの検証] ダイアログ ボックスの [単一変数規則] タブ ([0 to 100 by 5 defined] が定義された場合)

次に、定義した規則を分析変数に適用する必要があります。

- [1 to 3 Categorical] を「病院の規模」に適用します。
- [0 to 5 Categorical] を、「初期のランキンスコア」、および「1 か月後のランキンスコア」から「6 か月後のランキンスコア」までの変数に適用します。
- [0 to 100 by 5] を「1 か月後のバーセルインデックス」から「6 か月後のバーセルインデックス」までの変数に適用します。
- **[クロス変数規則]** タブをクリックします。

  現在定義されている規則はありません。
- **[規則の定義]** をクリックします。

図 7-24
[検証規則の定義] ダイアログ ボックスの [クロス変数規則] タブ



規則がない場合は、新しいプレースホルダ規則が自動的に作成されます。

- 規則名として、「**2度死亡**」と入力します。
- 論理式として、「(DOA=1) & (dhosp=1)」と入力します。これにより、1 人の患者について「病院到着時死亡」と「病院での死亡」という 2 つの記録がなされている場合は戻り値が 1 となります。
- [続行] をクリックします。

  [クロス変数規則] タブでは、新規に定義された規則が自動的に選択されます。
- [OK] をクリックします。

クロス変数規則

図 7-25
クロス変数規則

変数間規則

規則	違反の数	規則式
2度死亡	45	(DOA = 1) & (病院ID=1)

クロス変数規則についての集計画面には、1 つ以上違反のあったクロス変数規則、違反のあった回数、および違反のあった規則の説明が一覧として表示されます。

ケースのレポート

図 7-26
ケースのレポート

ケース報告書[b]

ケース	検証規則違反		識別子		
	単一変数[a]	変数間	病院ID	患者ID	担当医ID
29	0 to 100 by 5 (2)		NHV	4592215163	942982
30	0 to 100 by 5 (2)	2度死亡	NHV	7628592330	371884
31	0 to 100 by 5 (2)	2度死亡	NHV	7628592330	371884
32	0 to 100 by 5 (2)		NHV	3842780503	190855
33	0 to 100 by 5 (2)		NHV	8401049557	371884
34	0 to 100 by 5 (2)		NHV	0980186054	942982
35	0 to 100 by 5 (2)	2度死亡	NHV	8509099172	237418
36	0 to 100 by 5 (2)		NHV	0548592276	371884
37	0 to 100 by 5 (2)		NHV	8090627400	942982
100	0 to 100 by 5 (2)		QWS	5856145337	817329

a. その規則に違反した変数の数はそれぞれの規則を満たしています。
b. 規則違反のあるケースが 100 個よりも多く存在しました。最初の 100 だけが表示されます。

ケースのレポートには、クロス変数規則に違反したケースのほか、以前単一変数規則に対する違反を検出されたケースが表示されます。これらのケースはすべて、データ入力チームに報告して修正する必要があります。

要約表

以上で、分析担当者は、データ入力責任者向けの予備レポートに必要な情報を準備することができました。

関連手続き

データの検証手続きは、データの品質を管理する上で有用な手段です。

- 例外ケースの特定手続きでは、データ内のパターンを分析し、類型からの逸脱が顕著な値が含まれるケースを特定できます。

章

8

自動データ準備

分析に向けてデータを準備することは、プロジェクトにおいて最も重要な手順の 1 つですが、従来は最も時間を消費する手順の 1 つでもありました。自動データ準備 (ADP) は、データ分析および修正の特定、問題となる、または有用でないと考えられるフィールドの除外、必要に応じた新しい属性の取得、高度なスクリーニング手法を用いたパフォーマンスの改善を行い、タスクを処理します。完全に**自動化**した方法でアルゴリズムを使用して、修正を選択または適用したり、**インタラクティブ**な方法を使用して、必要に応じて変更を実行、承認または拒否する前に変更をプレビューすることができます。

ADP を使用すると、実行する統計の概念の事前情報を必要とせず、モデルを迅速かつ用意に作成できるよう、データを準備することができます。モデルはより迅速に構築およびスコアリングするようになります。また、ADP を使用すると、、自動モデル作成プロセスの強固さをより向上させます。

自動データ準備をインタラクティブに使用

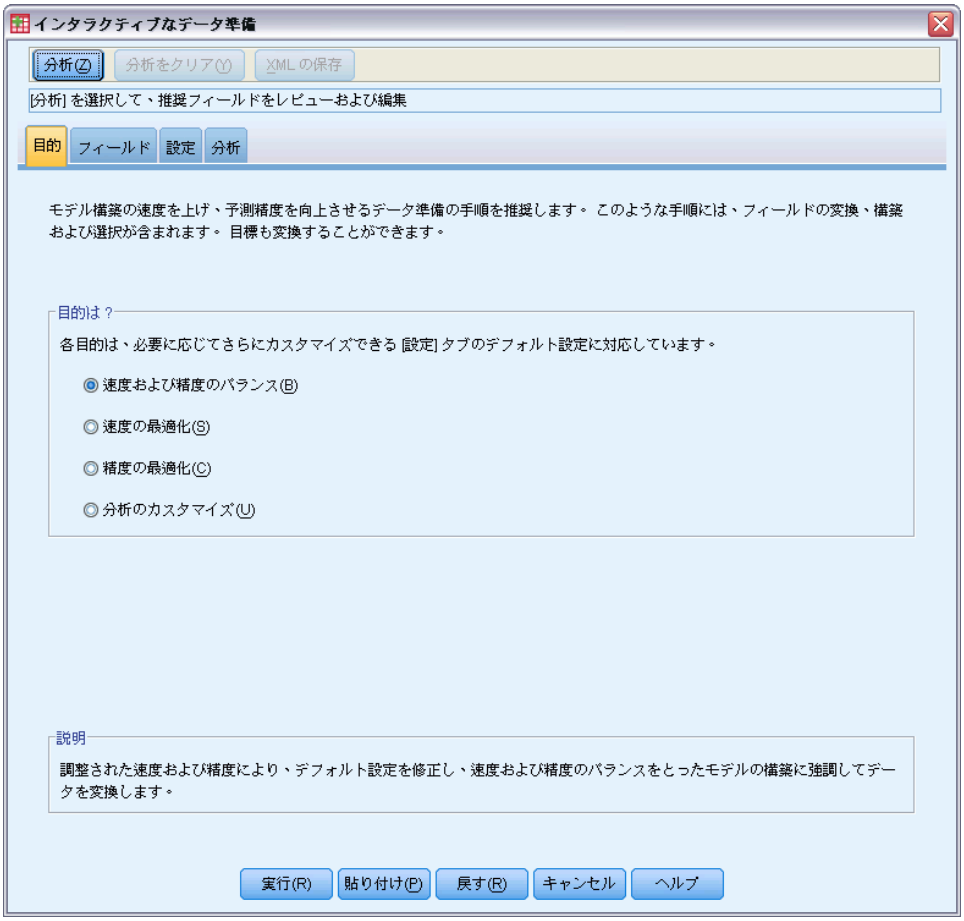
世帯主の保険請求を調査するためのリソースが制限されている保険会社が、不正請求の恐れのある疑いを区別するためのモデルを作成したいと考えています。その会社には、insurance_claims.sav で収集された以前の請求についての情報のサンプルがあります。詳細は、 A 付録 p.149 サンプル ファイル を参照してください。 モデルを作成する前に、自動データ準備を使用して、モデル作成のためのデータを準備します。変換が適用される前に提案される変換を確認できる必要があるため、自動データ準備をインタラクティブ モードで使用します。

目的の選択

- 自動データ準備をインタラクティブに実行するには、メニューから次の項目を選択します。

  変換(T) > モデル作成のデータ準備 > インタラクティブ...

図 8-1
[目的] タブ



最初のタブでは、デフォルト設定を制御する目的を要求しますが、目的の実際の違いはどのようになっているのでしょう?各目的を使用して手順を実行して、結果の違いを確認することができます。

► [**速度および精度のバランス**] が選択されていることを確認し、[**分析**] を選択します。

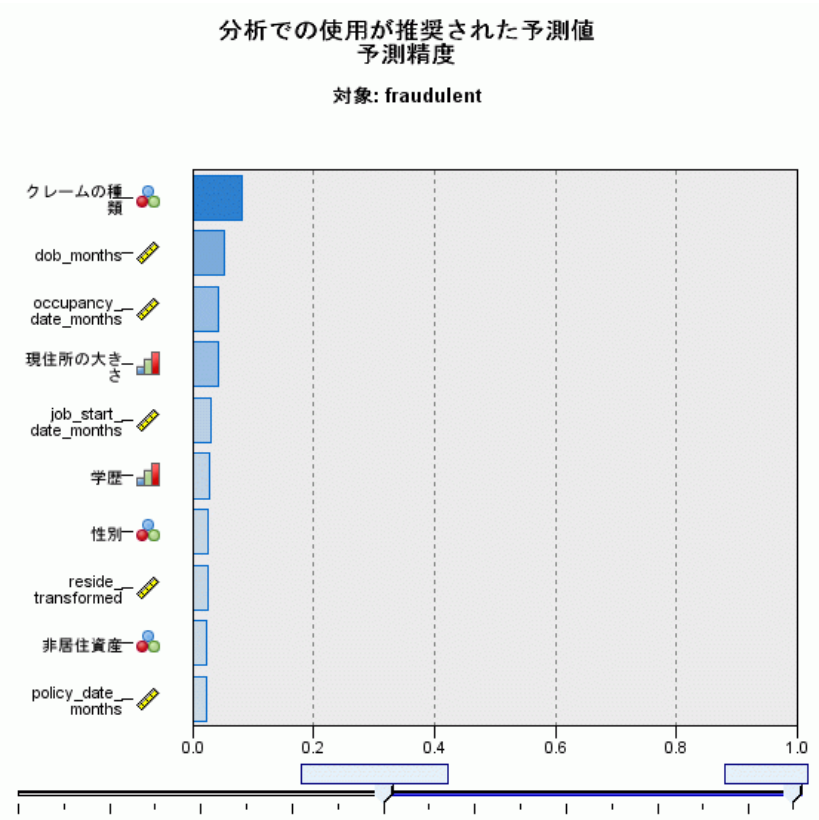
図 8-2
[分析] タブ、調整された目的のフィールド処理の要約

処理したフィールドの要約

フィールド		N
対象		1
予測値		18
分析での使用に推奨される予測値	合計	18
	元のフィールド (未変換)	8
	元のフィールドの変換	5
	日付と時間から派生	5
	構造化	0
未使用の予測値		0

手続きでデータが処理されているとき、フォーカスは自動的に［分析］タブに切り替わります。デフォルトのメイン ビューは、［フィールド処理の要約］で、自動データ準備でフィールドがどのように処理されるかについての概要が表示されます。モデル作成に目標が 1 つ、18 の入力および 18 のフィールドが推奨されています。モデル作成に推奨されているフィールドのうち、9 つが元の入力フィールド、4 つが元の入力フィールドの変換、5 つが日付および時刻フィールドから派生したものです。

図 8-3
[分析] タブ、調整された目的の予測精度



デフォルトの補助ビューは［予測精度］で、推奨フィールドのうちどれがモデル作成に最も役立つかについて、すばやく表示します。18 の予測フィールドが分析に推奨されますが、デフォルトでは、最初の 10 個のフィールドのみが予測精度グラフに表示されます。フィールドをより多くまたはより少なく表示するには、グラフの下のスライド コントロールを使用します。

［速度および精度のバランス］を目的に指定し、請求の種類を「最適な」予測値として特定し、その後「家族の人数」および請求者の年齢（誕生日から現在の日付までの期間・月数）が続きます。

▶ **［分析をクリア］**をクリックして、［目的］タブをクリックします。

▶ **［速度の最適化］**を選択して、**［分析］**をクリックします。

図 8-4
[分析] タブ、速度に最適化された場合のフィールド処理の要約

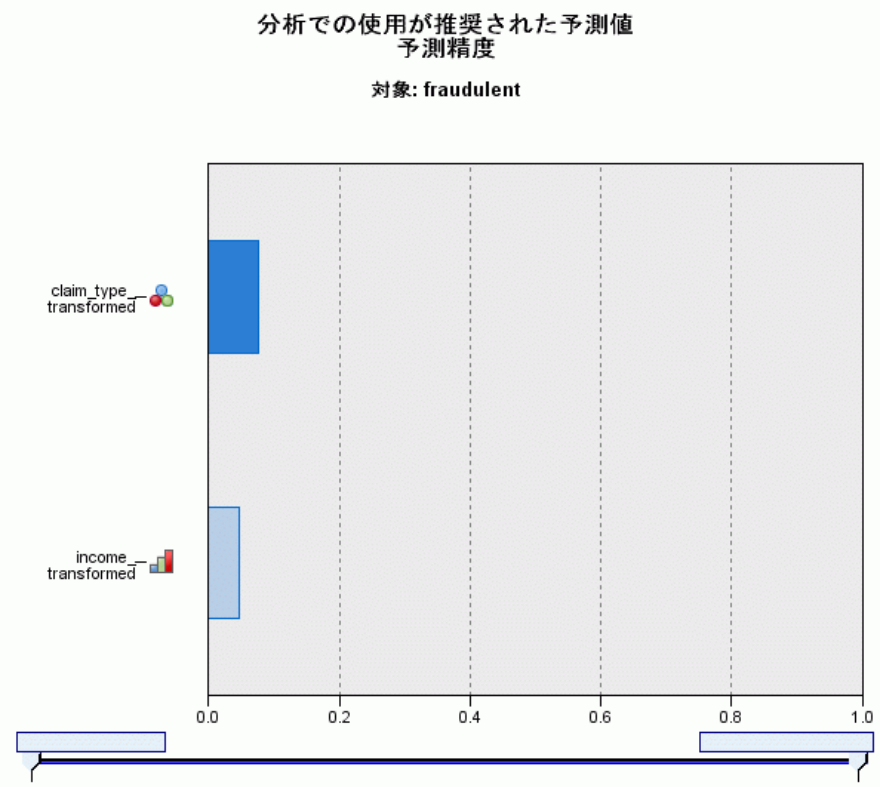
処理したフィールドの要約

フィールド		N
対象		1
予測値		18
分析での使用に推奨される予測値	合計	2
	元のフィールド (オ変換)	0
	元のフィールドの変換	2
	日付と時間から派生	0
	構造化	0
未使用の予測値		16

- 役立つ予測値を構築できませんでした。 一般的な理由は、対象との関連が強い連続型予測フィールドが少なすぎること、またはすべての連続型の予測フィールドが独立していたことです。

手続きでデータが処理されているとき、フォーカスは再度自動的に［分析］タブに切り替わります。この場合、モデル作成には 2 つのフィールドのみが推奨され、いずれのフィールドも元のフィールドから変換されたものとなります。

図 8-5
[分析] タブ、速度に最適化された場合の予測精度



目的に [**速度の最適化**] を指定した場合、claim_type_transformed が「最適な」予測フィールドとして指定され、その次に income_transformed が指定されます。

▶ [**分析をクリア**] をクリックして、[目的] タブをクリックします。

▶ [**精度の最適化**] を選択して、[**分析**] をクリックします。

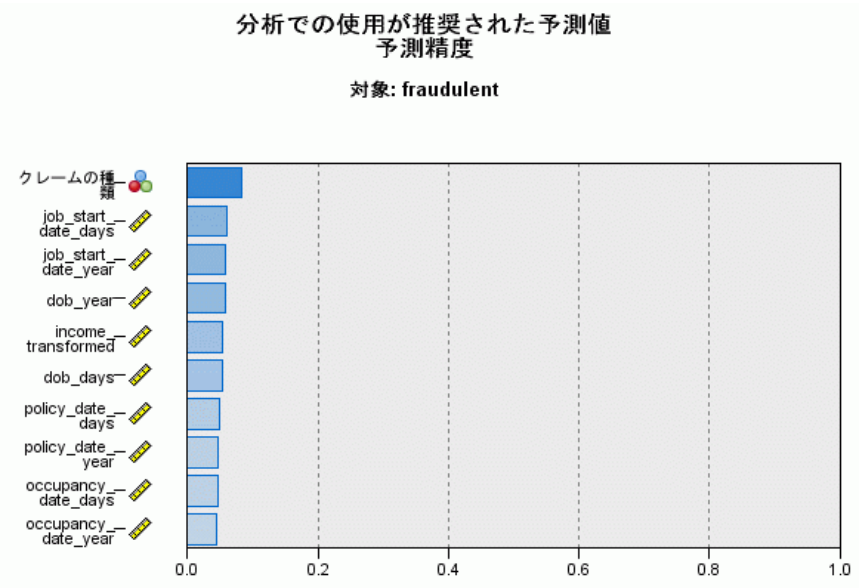
図 8-6
[分析] タブ、精度に最適化された場合の予測精度

処理したフィールドの要約

フィールド		N
対象		1
予測値		18
分析での使用に推奨される予測値	合計	32
	元のフィールド (未変換)	8
	元のフィールドの変換	5
	日付と時間から派生	19
	構造化	0
未使用の予測値		0

目的に [**精度の最適化**] を指定した場合、日付から日、月、年、そして時刻から時、分、秒を取得して、モデル フィールドを取得するため、32 個のフィールドがモデル作成に推奨されます。

図 8-7
[分析] タブ、精度に最適化された場合の予測精度



[請求の種類] が「最適な」予測フィールドとして指定され、その次に請求者が勤務を開始してからの日数（金部開始日から現在の日付まで算出された期間）、および現在の勤務を開始した年（勤務開始日から算出）が指定されます。

要約

- **[速度および精度のバランス]** は、日付からのモデル作成に役立つフィールドを作成し、より正規分布になるよう「同居人数」のような連続からフィールドを変換します。
- **[精度の最適化]** は日付から追加フィールドをいくつか作成します（また、外れ値をチェックし、目標が連続型である場合は、より正規分布になるよう変換します）。
- **[速度の最適化]** は、日付フィールドを準備せず、連続型フィールドを尺度化しませんが、目標がカテゴリ型である場合、カテゴリ型予測フィールドのカテゴリを結合し、連続型予測フィールドを分割します（また、目標が連続型の場合、フィールド選択およびフィールド構築を実行します）。

保険会社は、**[精度の最適化]** の結果をより詳細に調査します。

▶ メイン ビューのドロップダウンから、**[フィールド]** を選択します。

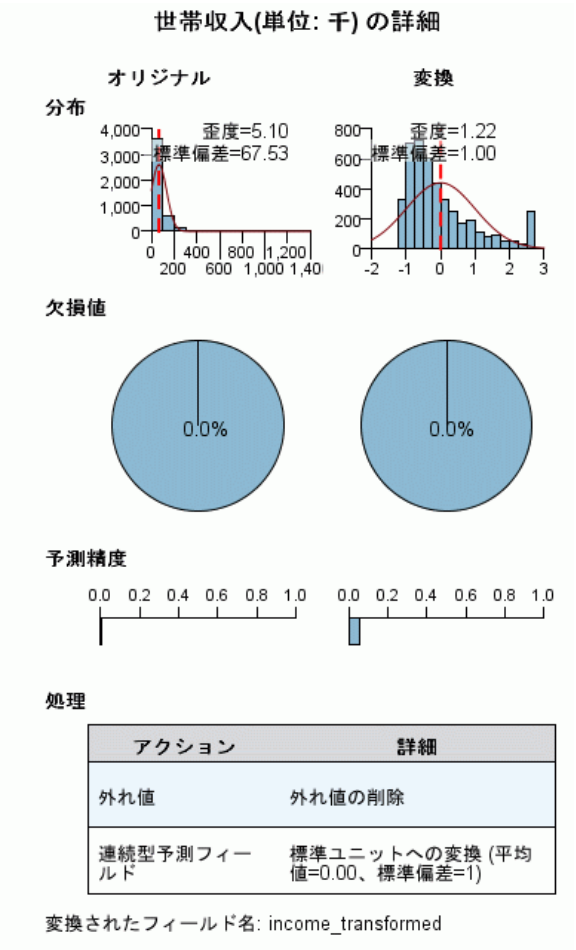
フィールドおよびフィールドの詳細

図 8-8
フィールド

［フィールド］ビューには、処理済みフィールドと、ADP がモデル作成にそれらのフィールドの使用を推奨するかどうかを表示します。フィールド名をクリックすると、フィールドに関する詳細情報がリンク ビューに表示されます。

▶ **［収入］**をクリックします。

図 8-9
世帯の収入（千単位）に関するフィールドの詳細

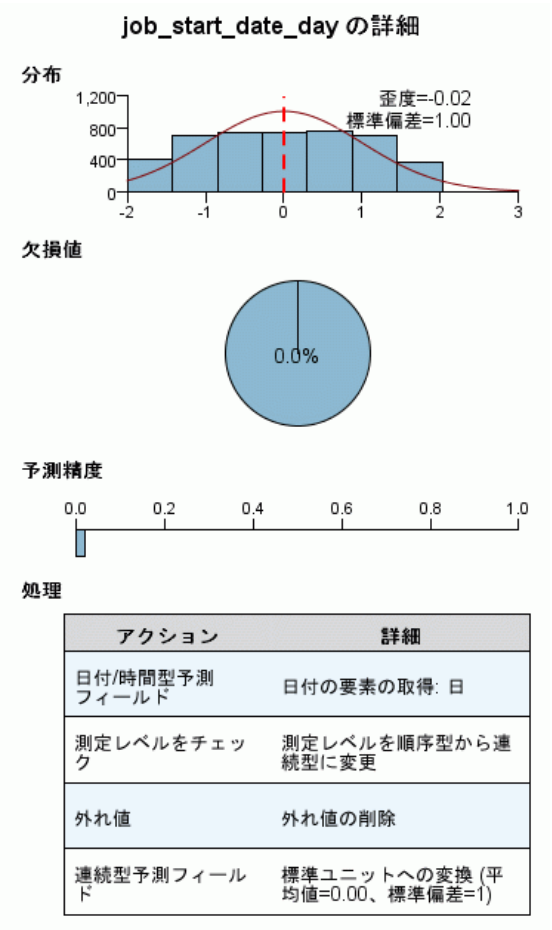


［フィールドの詳細］ビューには、［世帯の収入（千単位)］の元のフィールドと変換されたフィールドの分布を表示します。処理表に従い、値を外れ値を決定する分割点に設定して外れ値として特定されたレコードを選定し、フィールドを標準化して平均値 0 および標準偏差 1 となります。変換されたフィールドのヒストグラムの右側にある「バンプ」は、おそらく 200 を超えるレコード数が外れ値として特定されます。収入は非常に歪んだ分布であるため、これはデフォルトの分割点を使用して外れ値を決定するには、あまりに強引なケースとなります。

また、元のフィールドより変換されたフィールドが予測精度が高くなります。これは有用な変換であることがわかります。

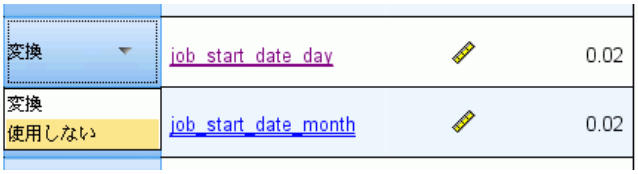
▶ ［フィールド］ビューで、[job_start_date_day] をクリックします（[job_start_date_days］とは異なりますので注意してください)。

図 8-10
job_start_date_day のフィールドの詳細



フィールド［job_start_date_day］は、［雇用開始日［job_start_date］］から抽出された日です。このフィールドが請求が不正であるかについての意味があるか、可能性は非常に低いため、保険会社はモデル作成の候補から削除したいと考えます。

図 8-11
世帯の収入（千単位）に関するフィールドの詳細



- ［フィールド］ビューで、[job_start_date_day] 行の［使用バージョン］ドロップダウンから［**使用しない**］を選択します。_day や _month の接尾辞を持つすべてのフィールドに同じ操作を実行します。
- 変換を適用するには、[**実行**] をクリックします。

  すべての推奨予測フィールド（新旧ともに）の役割を［入力］に、推奨されていない予測フィールドの役割を［なし］に設定して、データセットのモデル作成の準備が整いました。推奨予測フィールドのみを持つデータセットを作成するには、ダイアログで［変換の適用］設定を使用します。

自動データ準備を自動で使用

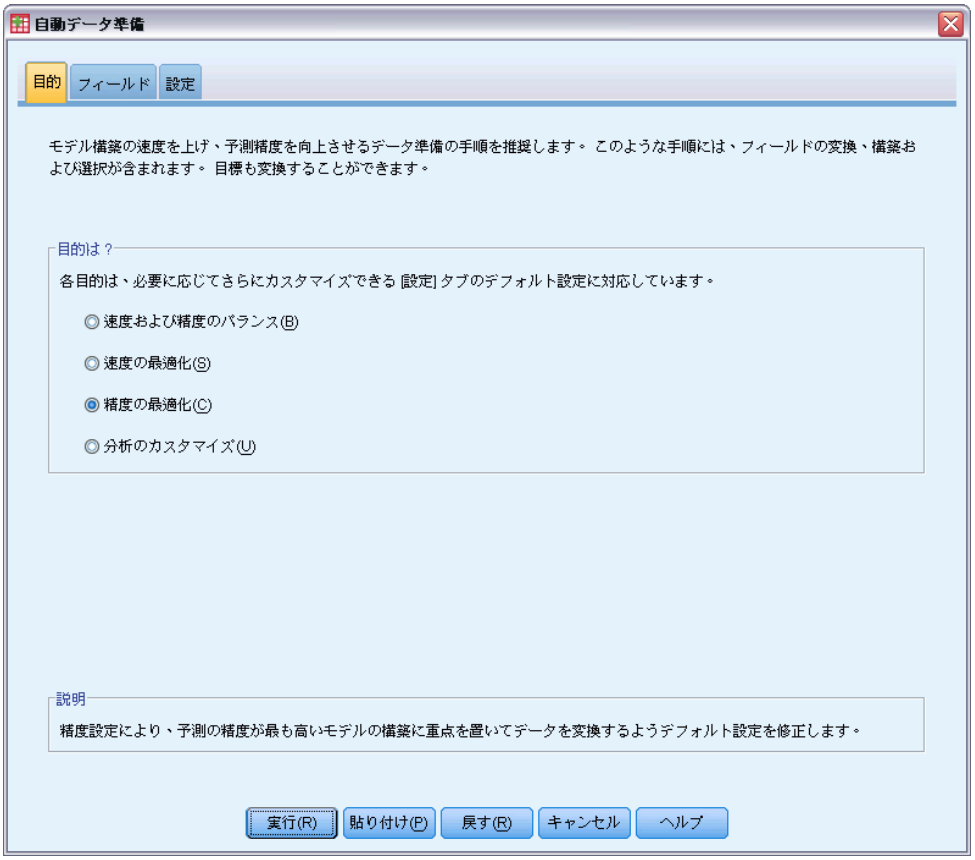
自動車産業グループは、さまざまな個人用自動車の売り上げを記録します。採算ベースを上回るモデルおよび下回るモデルを特定できるように、自動車の売り上げと自動車の特性との関係を確立したいと考えます。この情報は、car_sales_unprepared.sav に収集されています。詳細は、 A 付録 p. 149 サンプル ファイル を参照してください。 自動データを使用して、分析するデータを準備します。また、準備「前」および準備「後」のデータを使用してモデルを作成し、結果を比較できるようにします。

データの準備

- 自動データ準備を自動モードで実行するには、メニューから次の項目を選択します。

  変換(T) > モデル作成のデータ準備 > 自動…

図 8-12
[目的] タブ

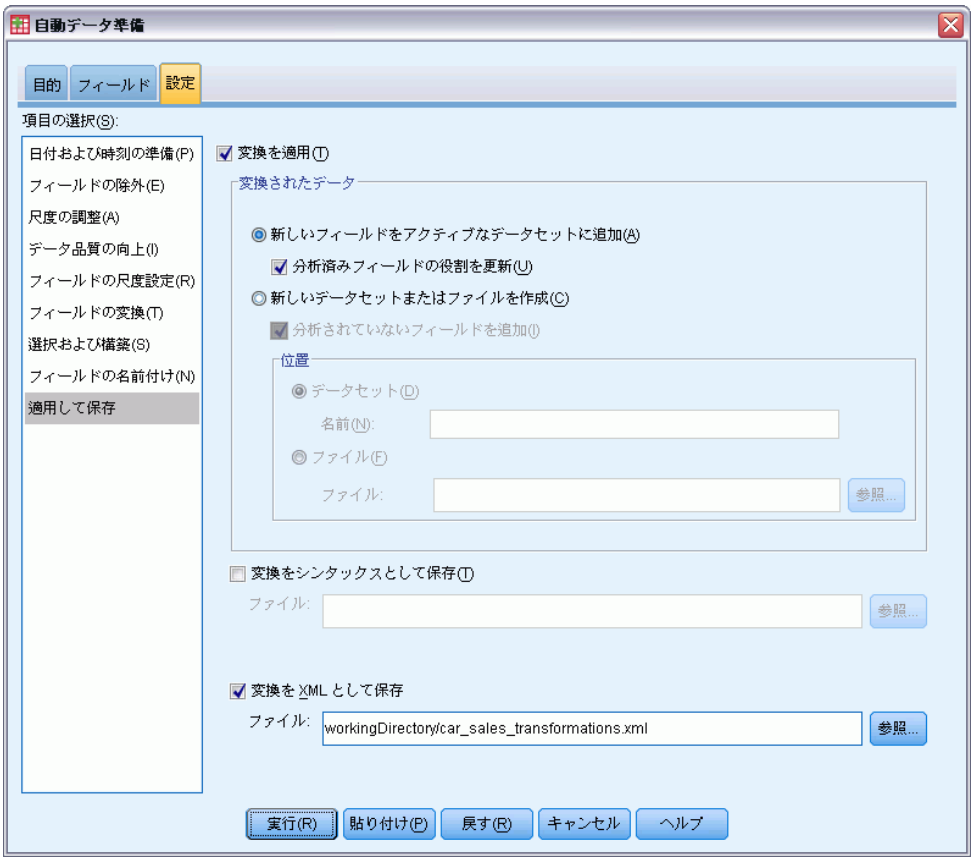


► **[精度の最適化]** を選択します。

目標フィールド、[売上額（単位:千)] が連続型フィールドで、自動データ準備中に変換できた場合、[スコアの後方変換] ダイアログを使用して、変換された目標フィールドの予測値を元のスケールに戻すことができるよう、変換を XML ファイルに保存する必要があります。

► **[設定]** タブをクリックし、**[適用して保存]** 設定をクリックします。

図 8-13
[適用して保存] 設定



- [変換の保存] を XML で選択し、[参照] をクリックして workingDirectory/car_sales_transformations.xml に移動、ファイルを保存するパスに workingDirectory を指定します。
- [実行] をクリックします。

以上の選択により、次のコマンド シンタックスが生成されます。

```
*Automatic Data Preparation.
ADP
  /FIELDS TARGET=sales INPUT=resale type price engine_s horsepow wheelbas wid
    curb_wgt fuel_cap mpg
  /PREPDATETIME DATEDURATION=YES(REFERENCE=YMD('2009-06-04') UNIT=AUTO)
    TIMEDURATION=YES(REFERENCE=HMS('08:43:35') UNIT=AUTO) EXTRACTYEAR=YES(SUF
    EXTRACTMONTH=YES(SUFFIX='_month') EXTRACTDAY=YES(SUFFIX='_day')
    EXTRACTHOUR=YES(SUFFIX='_hour') EXTRACTMINUTE=YES(SUFFIX='_minute')
    EXTRACTSECOND=YES(SUFFIX='_second')
   /SCREENING PCTMISSING=YES(MAXPCT=50) UNIQUECAT=YES(MAXCAT=100) SINGLECAT=N
  /ADJUSTLEVEL INPUT=YES TARGET=YES MAXVALORDINAL=10 MINVALCONTINUOUS=5
  /OUTLIERHANDLING INPUT=YES TARGET=NO CUTOFF=SD(3) REPLACEWITH=CUTOFFVALUE
  /REPLACEMISSING INPUT=YES TARGET=NO
```

```
  /REORDERNOMINAL INPUT=YES TARGET=NO
  /RESCALE INPUT=ZSCORE(MEAN=0 SD=1) TARGET=BOXCOX(MEAN=0 SD=1)
  /TRANSFORM MERGESUPERVISED=NO MERGEUNSUPERVISED=NO BINNING=NONE SELECTION=N
    CONSTRUCTION=NO
  /CRITERIA SUFFIX(TARGET='_transformed' INPUT='_transformed')
  /OUTFILE PREPXML='/workingDirectory/car_sales_transformations.xml'.
TMS IMPORT
  /INFILE TRANSFORMATIONS='/workingDirectory/car_sales_transformations.xml'
    MODE=FORWARD (ROLES=UPDATE)
  /SAVE TRANSFORMED=YES.
EXECUTE.
```

- ADP コマンドは、目標フィールド［売上額］と入力フィールド［再販］から［マイル/ガロン］を準備します。
- PREPDATETIME サブコマンドが指定されますが、日付/時刻フィールドがないため、使用されません。
- ADJUSTLEVEL サブコマンドは、値が 10 を超える順序型フィールドを連続型フィールドに、値が 5 より小さい連続型フィールドを順序型フィールドに変更します。
- OUTLIERHANDLING サブコマンドは、平均値からの標準偏差が 3 を超える連続型入力フィールド（目標フィールドではない）の値を、平均値からの標準偏差が 3 である値に置き換えます。
- REPLACEMISSING サブコマンドは、欠損値である入力フィールド（目標ではない）の値を置き換えます。
- REORDERNOMINAL サブコマンドは、最も頻繁に発生しない名義型入力フィールドの値を最も頻繁に発生する入力フィールドの値に再コード化します。
- RESCALE サブコマンドは z スコア変換を使用して連続型入力フィールドを標準化し、平均値が 0、標準偏差が 1 になるように、また Box-Cox 変換を使用して連続型目標フィールドを標準化して平均値が 0、標準偏差が 1 になるようにします。
- TRANSFORM サブコマンドは、このサブコマンドで指定されたすべてのデフォルト操作をオフにします。
- CRITERIA サブコマンドは、目標フィールドおよび入力フィールドの変換にデフォルトの接尾辞を指定します。
- OUTFILE サブコマンドは、変換を /workingDirectory/car_sales_transformations.xml に保存するよう指定します。/workingDirectory は、car_sales_transformations.xml を保存するパスです。
- TMS IMPORT コマンドは car_sales_transformations.xml の変換を読み込み、その変換をアクティブ データセットに適用して、変換された既存フィールドの役割を更新します。
- EXECUTE コマンドにより、変換を処理します。これをシンタックスの長いストリームの一部として使用する場合、EXECUTE コマンドを削除して、所持時間を短くできる場合があります。

準備されていないデータのモデル作成

▶ 準備されていないデータでモデルを作成するには、メニューから次の項目を選択します。
分析(A) > 一般線型モデル > 1 変量...

図 8-14
[GLM – 1 変量] ダイアログ

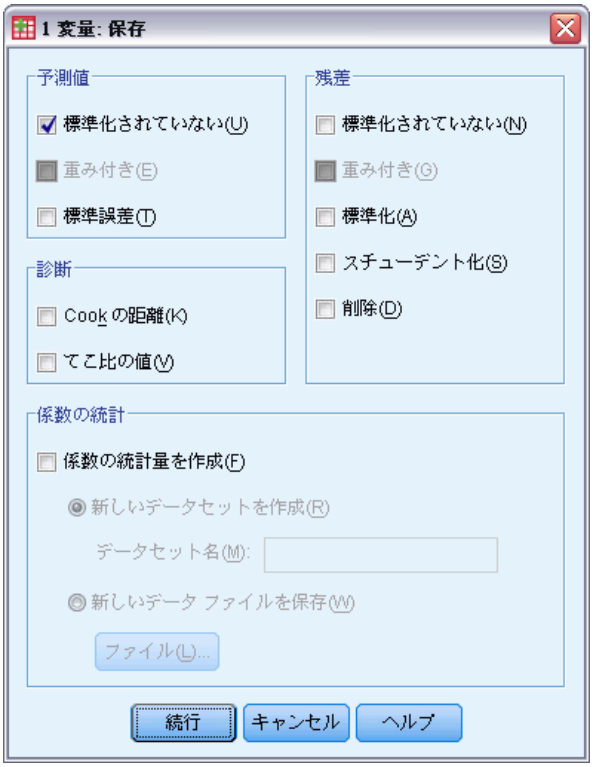
▶ 従属変数として「売上額（単位:千）[売上]」を選択します。

▶ 固定因子として、[車両タイプ [タイプ]] を選択します。

▶ [4 年再販価格 [再販]] から [燃料効率(マイル/ガロン) [mpg]] を共変量として選択します。

▶ **[保存]** をクリックします。

図 8-15
[保存] ダイアログ



- ［予測値］グループの［**標準化されていない**］を選択します。
- ［**続行**］をクリックします。
- ［GLM 1 変量］ダイアログ ボックスで［OK］をクリックします。

以上の選択により、次のコマンド シンタックスが生成されます。

```
UNIANOVA sales BY type WITH resale price engine_s horsepow wheelbas width len
    curb_wgt fuel_cap mpg
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /SAVE=PRED
  /CRITERIA=ALPHA(0.05)
  /DESIGN=resale price engine_s horsepow wheelbas width length curb_wgt fuel_
    mpg type.
```

図 8-16
準備されていないデータに基づくモデルに対する被験者間の効果

従属変数:Sales in thousands

ソース	タイプ III 平方和	自由度	平均平方	F 値	有意確率
修正モデル	226123.658[a]	11	20556.696	5.050	.000
切片	12227.688	1	12227.688	3.004	.086
resale	50.702	1	50.702	.012	.911
price	471.630	1	471.630	.116	.734
engine_s	19872.712	1	19872.712	4.882	.029
horsepow	9644.486	1	9644.486	2.369	.127
wheelbas	29824.272	1	29824.272	7.327	.008
width	263.465	1	263.465	.065	.800
length	1374.525	1	1374.525	.338	.562
curb_wgt	32762.692	1	32762.692	8.049	.005
fuel_cap	1124.237	1	1124.237	.276	.600
mpg	337.585	1	337.585	.083	.774
type	17668.779	1	17668.779	4.341	.040
誤差	427402.183	105	4070.497		
総和	1062354.955	117			
修正総和	653525.841	116			

a. R2 乗 = .346 (調整済み R2 乗 = .277)

デフォルトの GLM 1 変量出力には、分散分析表である被験者間の効果が含まれています。モデル内の各項目および全体としてモデルが、従属変数の変動を説明する機能についてテストされます。この表には、変数ラベルは表示されません。

予測フィールドは、さまざまな有意水準を示します。有意値が 0.05 より小さい予測フィールドは通常、モデルに役立つとみなされます。

準備されたデータのモデル作成

図 8-17
[GLM - 1 変量] ダイアログ

- 準備されたデータのモデルを作成するには、[GLM - 1 変量] ダイアログを再度呼び出します。
- 「売上額（単位:千）[売上]」の選択を解除し、従属変数として「sales_transformed」を選択します。
- [4 年再販価格 [再販]] から [燃料効率(マイル/ガロン) [mpg]] の選択を解除し、「resale_transformed 」から「mpg_transformed 」を共変量として選択します。
- [OK] をクリックします。

以上の選択により、次のコマンド シンタックスが生成されます。

```
UNIANOVA sales_transformed BY type WITH resale_transformed price_transformed
    engine_s_transformed horsepow_transformed wheelbas_transformed width_tran
    length_transformed curb_wgt_transformed fuel_cap_transformed mpg_transfor
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /SAVE=PRED
  /CRITERIA=ALPHA(0.05)
  /DESIGN=resale_transformed price_transformed engine_s_transformed horsepow_
    wheelbas_transformed width_transformed length_transformed curb_wgt_transf
    fuel_cap_transformed mpg_transformed type.
```

図 8-18
準備されたデータに基づくモデルに対する被験者間の効果

従属変数:sales_transformed

ソース	タイプ III 平方和	自由度	平均平方	F 値	有意確率
修正モデル	79.327[a]	11	7.212	13.638	.000
切片	2.436	1	2.436	4.606	.034
resale_transformed	.954	1	.954	1.804	.181
price_transformed	9.271	1	9.271	17.533	.000
engine_s_transformed	2.885	1	2.885	5.456	.021
horsepow_transformed	.034	1	.034	.064	.801
wheelbas_transformed	1.213	1	1.213	2.293	.132
width_transformed	.037	1	.037	.071	.791
length_transformed	.265	1	.265	.501	.480
curb_wgt_transformed	.103	1	.103	.194	.660
fuel_cap_transformed	.132	1	.132	.249	.618
mpg_transformed	3.390	1	3.390	6.411	.012
type	4.007	1	4.007	7.579	.007
誤差	76.673	145	.529		
総和	156.000	157			
修正総和	156.000	156			

a. R2 乗 = .509 (調整済み R2 乗 = .471)

準備されていないデータに作成されたモデルと準備されていないデータに作成されたモデルの被験者間の効果については、重要な違いがいくつかあります。まず、全体の自由度が増加します。これは自動データ準備中に欠損値が代入値に置き換えられるためで、最初のモデルからリストごとに削除されたレコードは 2 番目のモデルに使用できます。とりわけ、特定の予測フィールドの有意度が変わります。2 つのモデルはエンジン サイズ [engine_s] と車両タイプ [type] がモデルに有用で、ホイールベース [wheelbas] および車両総重量 [curb_wgt] があまり重要ではなく、車の価格 [price_transformed] および燃費 [mpg_transformed] が現在重要です。

このような変化はなぜ起こるのでしょう?売上額は歪んだ分布であるため、ホイールベースと車両総重量には影響を与えるレコードがいくつかありますが、売上額が変換されるとその影響はなくなります。別の可能性として、欠損値の置換によって使用できる追加のケースによって、これらの変数の統計的な重要度が変化したということが考えられます。いずれの場合でも、購入しないことについてのより詳細な調査が必要です。

準備データに作成されたモデルの R 2 乗が大きくなりますが、売上額が変換されているため、各モデルのパフォーマンスを比較するのに最適な測定ではありません。代わりに、観測値および予測値の 2 つのセット間のノンパラメトリック相関を計算できます。

予測値の比較

▶ 2 つのモデルから予測された値の相関を取得するには、メニューから次の項目を選択します。
分析(A) > 相関 > 2 変量...

図 8-19
[2 変量の相関分析] ダイアログ

▶ [売上額（単位:千）[売上]]、[売上予測値 [PRE_1]]、[売上予測値（変換）[PRE_2]] を分析変数として選択します。

▶ [相関係数] グループの [Pearson] の選択を解除し、[Kendall のタウ b] および [Spearman] を選択します。

[売上予測値（変換）[PRE_2]] を使用して、ノンパラメトリック相関を計算できます。後方変換しても予測値の順位は変わらないため、元のスケールに後方変換する必要はありません。

▶ [OK] をクリックします。

以上の選択により、次のコマンド シンタックスが生成されます。

```
NONPAR CORR
  /VARIABLES=sales PRE_1 PRE_2
  /PRINT=BOTH TWOTAIL NOSIG
  /MISSING=PAIRWISE.
```

図 8-20
ノンパラメトリック相関

			Sales in thousands	Sales に対する予測値	Sales_transformed に対する予測値
Kendallのタウb	Sales in thousands	相関係数	1.000	.376**	.484**
		有意確率 (両側)	.	.000	.000
		N	157	117	157
	Sales に対する予測値	相関係数	.376**	1.000	.655**
		有意確率 (両側)	.000	.	.000
		N	117	117	117
	Sales_transformed に対する予測値	相関係数	.484**	.655**	1.000
		有意確率 (両側)	.000	.000	.
		N	157	117	157
Spearmanのロー	Sales in thousands	相関係数	1.000	.530**	.666**
		有意確率 (両側)	.	.000	.000
		N	157	117	157
	Sales に対する予測値	相関係数	.530**	1.000	.831**
		有意確率 (両側)	.000	.	.000
		N	117	117	117
	Sales_transformed に対する予測値	相関係数	.666**	.831**	1.000
		有意確率 (両側)	.000	.000	.
		N	157	117	157

**. 相関は、1 % 水準で有意となります (両側)。

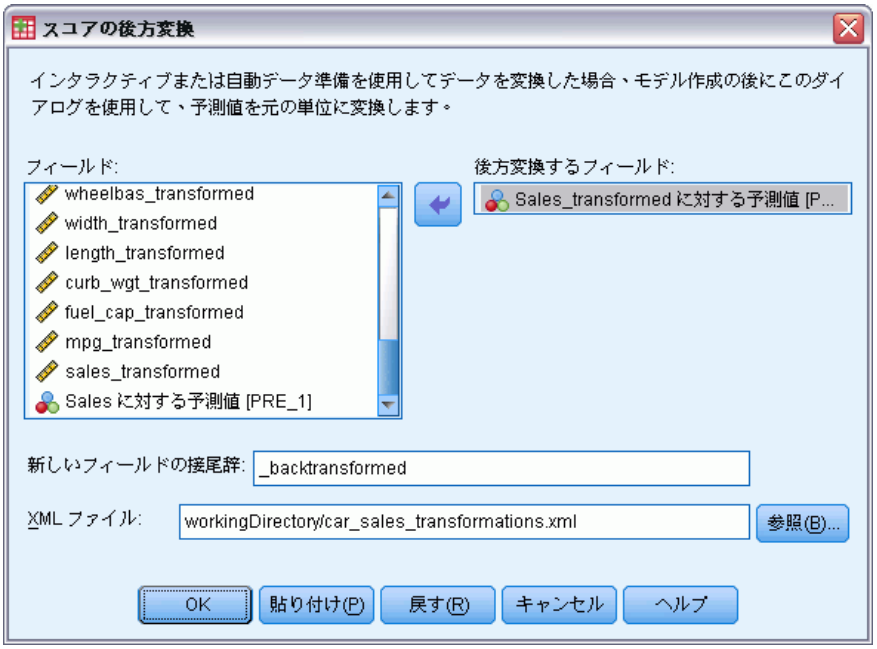
最初の列には、準備されたデータを使用して作成されたモデルの予測値が、Kendall のタウ b と Spearman のロー測定によって観測された値とより強く相関していることを示します。つまり、自動データ準備の実行によってモデルが改善されることを示します。

予測値の後方変換

▶ 準備されたデータには売上額の変換が含まれるため、このモデルからの予測値はスコアとして直接使用できません。予測値を元のスケールに変換するには、メニューから次の項目を選択します。

変換(T) > モデル作成のデータ準備 > スコアの後方変換...

図 8-21
[スコアの後方変換] ダイアログ



- ▶ [売上予測値（変換）[PRE_2]] を後方変換するフィールドとして選択します。
- ▶ 新しいフィールドの接尾辞として「_backtransformed」と入力します。
- ▶ 「workingDirectory¥car_sales_transformations.xml」と入力し、変換を含む XML ファイルの場所として、ファイルのパスを workingDirectory に代入します。
- ▶ [OK] をクリックします。

以上の選択により、次のコマンド シンタックスが生成されます。

```
TMS IMPORT
  /INFILE TRANSFORMATIONS='workingDirectory/car_sales_transformations.xml'
    MODE=BACK (PREDICTED=PRE_2 SUFFIX='_backtransformed').
EXECUTE.
```

- TMS IMPORT コマンドは car_sales_transformations.xml の変換を読み込み、後方変換を PRE_2 に適用します。
- 後方変換を含む新しいフィールドの名前は PRE_2_backtransformed となります。
- EXECUTE コマンドにより、変換を処理します。これをシンタックスの長いストリームの一部として使用する場合、EXECUTE コマンドを削除して、所持時間を短くできる場合があります。

要約

自動データ準備を使用して、モデルを改善できるデータの変換を迅速に取得できます。目標フィールドが変換された場合、変換を XML ファイルに保存して、[スコアの後方変換] ダイアログを使用して、変換された目標の予測値を元のスケールに変換できます。

章
9

例外ケースの特定

異常検知手続きは、クラスタ グループのノルムからの偏差に基づいて異常ケースを検索します。この手続きは、推論的データ分析の前に、探索的データ分析手順において、データ監査の目的で異常ケースをすばやく検索するように設計されています。このアルゴリズムは、汎用的な異常検知用に設計されています。つまり、この異常ケースの定義は、医療産業における異常な支払いパターンの検知や、金融業会におけるマネー ロンダリングの検知など、異常の定義を正確に定義できる特定の応用例に固有のものではありません。

例外ケースの特定アルゴリズム

このアルゴリズムは、次の 3 つの段階に区分されます。

モデリング。この手続きは、データセット内の自然なグループ（またはクラスタ）を明確化する上で不可欠なクラスタ モデルを作成するためのものです。クラスタ化は、一組の入力変数に基づいて行われます。作成されたクラスタ モデルおよびクラスタ グループのノルムを計算するための十分統計量は、以後の処理で使用できるよう保存されます。

得点化。モデルが各ケースに適用され、そのクラスタ グループが特定されます。また、そのクラスタ グループに関してケースの異常度を測定するための指数が、ケースごとにいくつか作成されます。この異常指数の値を基準にして、すべてのケースが並べ替えられます。このリストの上位に位置するケースは、異常度が高いと判断されます。

理由の提示。異常と判断されたケースごとに、対応する変数の偏差指標を基準にして変数が並べ替えられます。その場合、ケースが異常であるという判断の根拠となった変数とその値、およびそれに対応するノルム値が、リストの上位に表示されます。

医療データベースにおける例外ケースの特定

脳卒中の治療結果に関する予測モデルは、異常な観測値の影響を受けやすいため、モデルを作成するデータ分析の担当者はデータの品質に気を使います。こうした異常な観測値の中には、非常に特異なケースを表しているため予測に使用するのは適当でないものがあります。また、技術

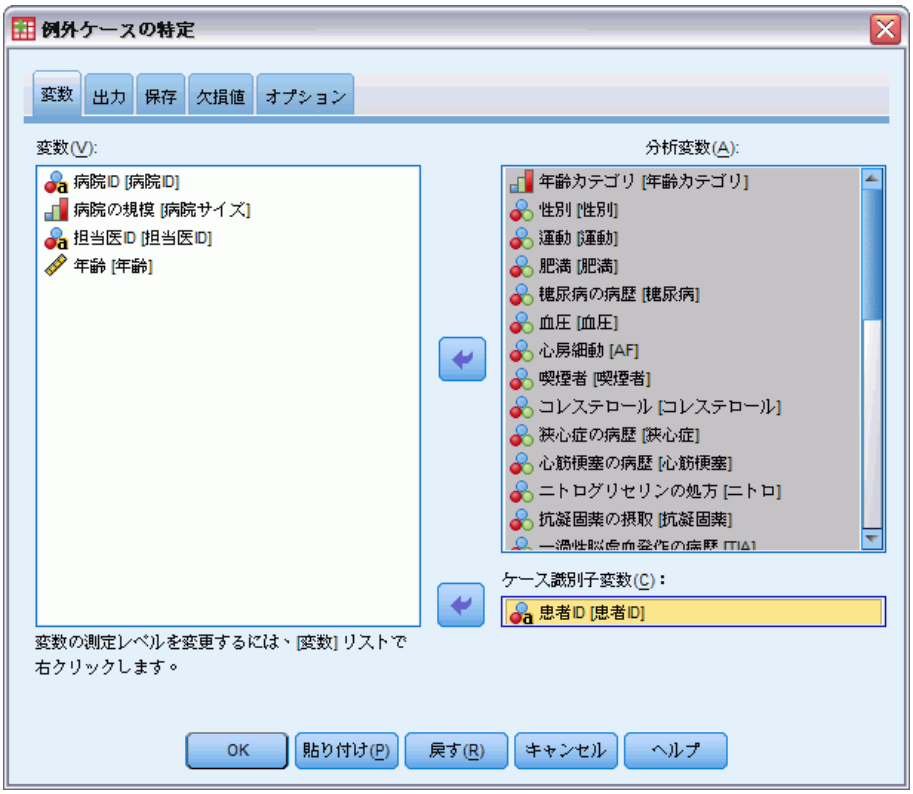
的には「正しい」値であっても、誤って入力されたために、データ検証の手続きでは検出できない観測値もあります。

この情報は、stroke_valid.sav に収集されています。詳細は、 A 付録 p. 149 サンプル ファイル を参照してください。［例外ケースの特定］手続きを使用すると、データ ファイル内のデータを整理できます。これらの分析結果を再生成するためのシンタックスは、detectanomaly_stroke.sps にあります。

分析の実行

▶ 例外ケースを特定するには、メニューから次の項目を選択します。
データ > 例外ケースの特定(I)...

図 9-1
[例外ケースの特定] ダイアログ ボックスの [変数] タブ

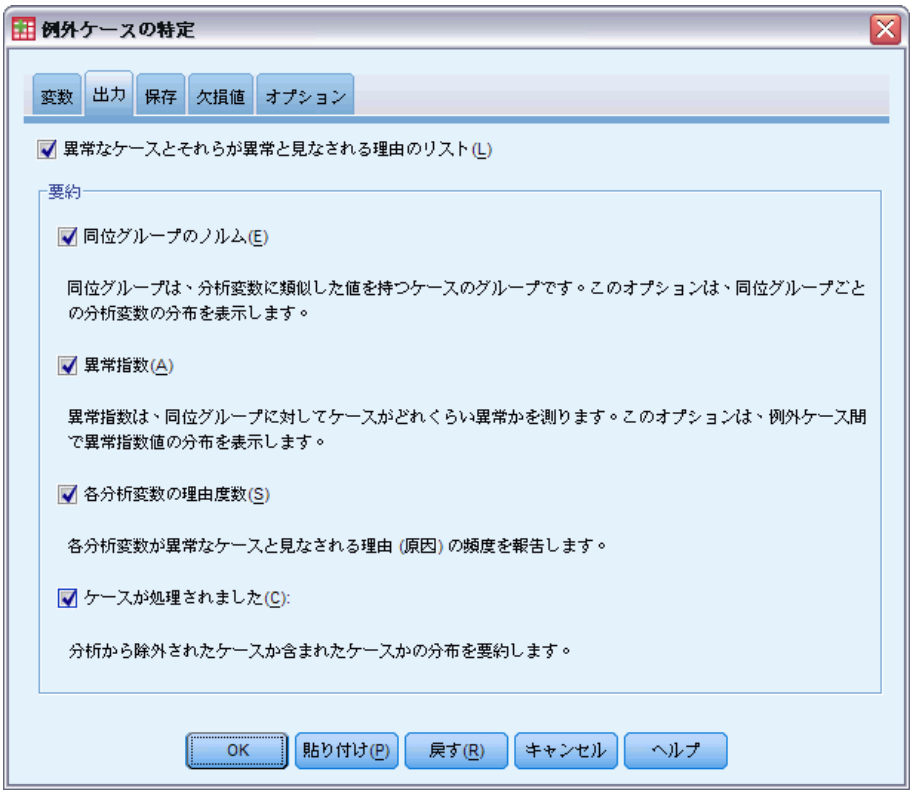


▶ 分析変数として、「年齢カテゴリ」から「3 か月から 6 か月以内の発作」までを選択します。

▶ またケース識別変数として、「患者 ID」を選択します。

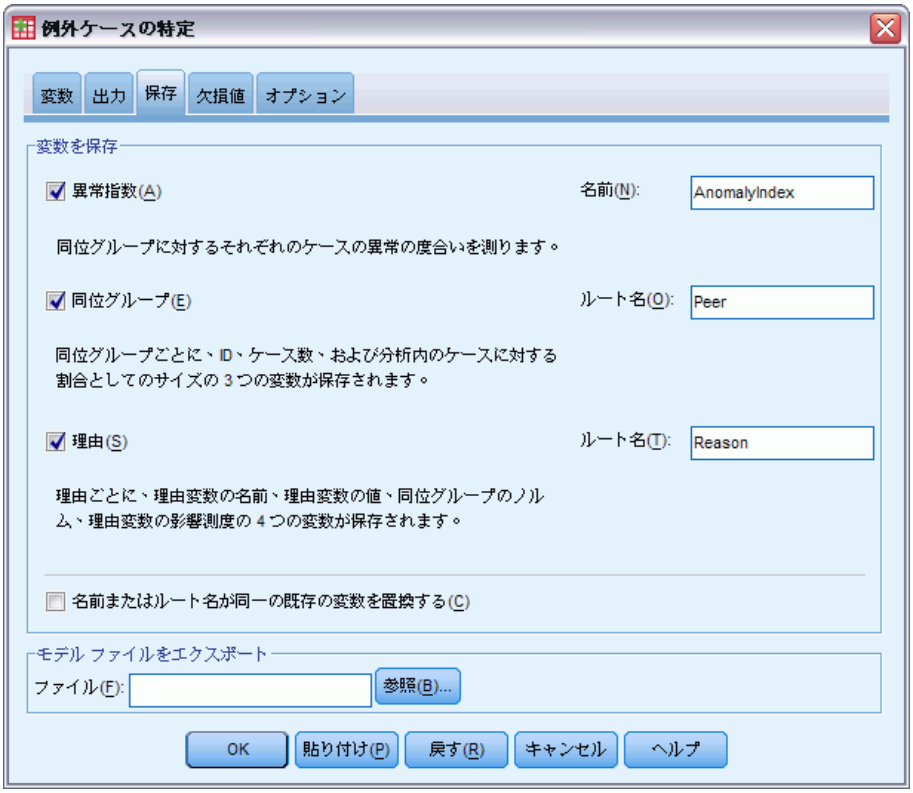
▶ [出力] タブをクリックします。

図 9-2
[例外ケースの特定] ダイアログ ボックスの [出力] タブ



- ▶ [同位グループのノルム]、[異常指数]、[各分析変数の理由度数]、および [処理されたケース] を選択します。
- ▶ [保存] タブをクリックします。

図 9-3
[例外ケースの特定] ダイアログ ボックスの [保存] タブ

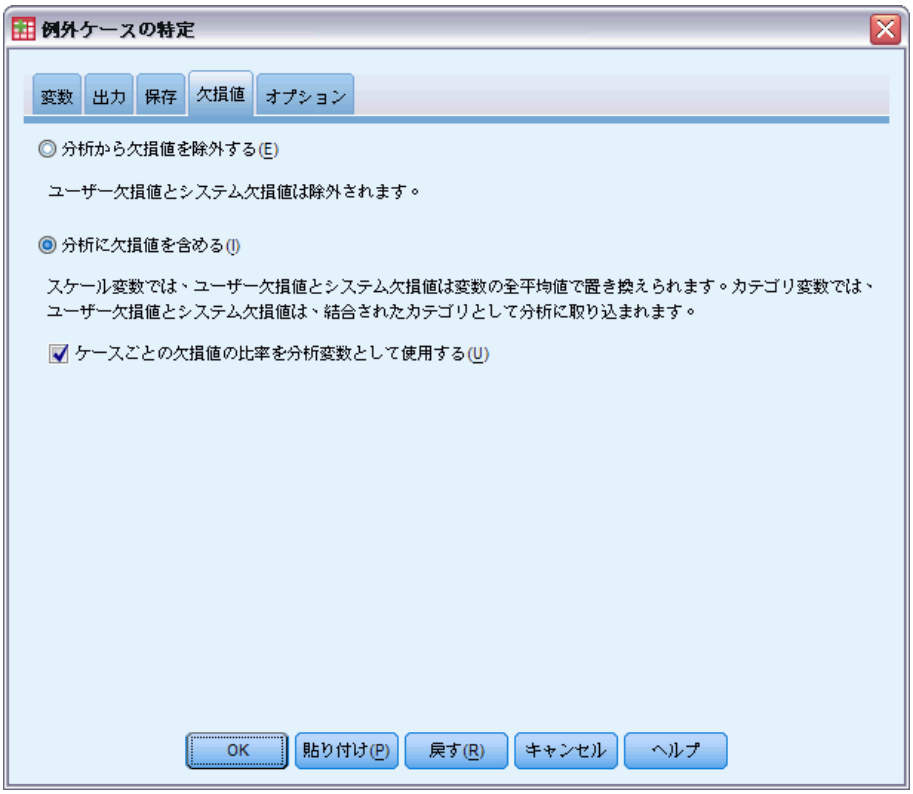


▶ [異常指数]、[同位グループ]、および [理由] を選択します。

これらの結果を保存すると、それを集計した便利な散布図を作成できます。

▶ [欠損値] タブをクリックします。

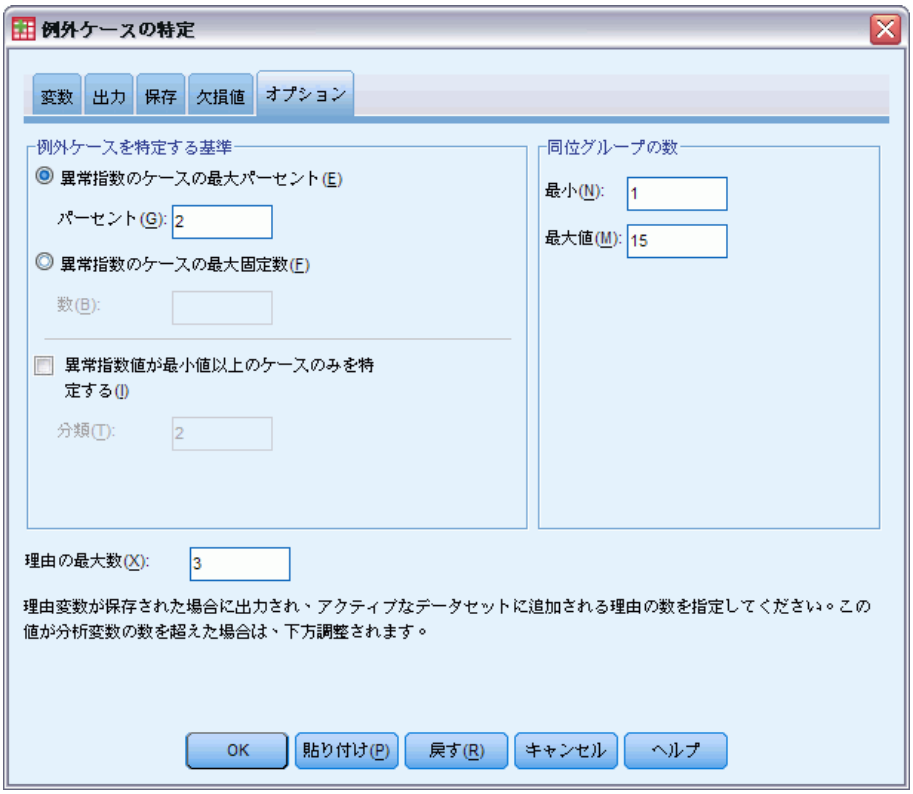
図 9-4
[例外ケースの特定] ダイアログ ボックスの [欠損値] タブ



► [**分析に欠損値を含める**] を選択します。このオプションが必要となるのは、治療前または治療中に死亡した患者を扱うためのユーザー欠損値が数多く存在するからです。分析には、ケースごとの欠損値の比率を測定する新たな変数がスケール変数として追加されます。

► [**オプション**] タブをクリックします。

図 9-5
[例外ケースの特定] ダイアログ ボックスの [オプション] タブ



- 異常指数のケースの最大パーセントとして「2」を入力します。
- [異常指数値が最小値以上のケースのみを特定する] の選択を解除します。
- 理由の最大数として「3」を入力します。
- [OK] をクリックします。

ケース処理の要約(O)

図 9-6
ケース処理要約(S)

処理したケースの要約

		N	結合の割合 (%)	合計数に対する割合 (%)
同位 ID	1	1183	100.0%	100.0%
結合		1183	100.0%	100.0%
合計		1183		100.0%

各ケースは、類似したケースで構成される同位グループに分類されます。ケース処理の要約には、作成された同位グループの数のほか、各同位グループに含まれるケースの数と割合（パーセント）が表示されます。

異常ケースの指数リスト

図 9-7
異常ケースの指数リスト

異常ケースの指数リスト

ケース	患者ID	異常指数
49	8717862852	1.599
162	5546809538	1.533
1010	5284009939	1.529
804	1336411777	1.506
974	4626554391	1.502
1124	0428895740	1.497
458	9159094175	1.489
258	9146506080	1.484
639	3962622031	1.484
132	7513809991	1.325
354	7216951816	1.325
1015	5857375438	1.324

異常指数は、ピアグループに関してケースの異常（例外）を反映した測度です。異常指数の値が高い上位 2% のケースが、ケース番号およびケース ID とともに表示されます。リストには 21 個のケースが表示されており、値は最小で 1.736、最大で 2.837 となっています。リスト内の 1 番目のケースと 2 番目のケースとでは、異常指数の値に比較的大きな差があります。これは、ケース 843 が異常なケースである可能性が高いことを示唆しています。その他のケースについては、状況に応じて判断する必要があります。

異常ケースの同位 ID リスト

図 9-8
異常ケースの同位 ID リスト

異常ケースの同位 ID リスト

ケース	患者ID	同位 ID	同位サイズ	同位サイズ パーセント
49	8717862852	1	1183	100.0%
162	5546809538	1	1183	100.0%
1010	5284009939	1	1183	100.0%
804	1336411777	1	1183	100.0%
974	4626554391	1	1183	100.0%
1124	0428895740	1	1183	100.0%
458	9159094175	1	1183	100.0%
258	9146506080	1	1183	100.0%
639	3962622031	1	1183	100.0%
132	7513809991	1	1183	100.0%
354	7216951816	1	1183	100.0%
1015	5857375438	1	1183	100.0%

異常である可能性を持つケースが、所属する同位グループについての情報とともに表示されます。先頭の 10 ケースを含め全部で 15 個のケースが、同位グループ 3 に属しており、その他は同位グループ 1 に属しています。

異常ケースの理由リスト

図 9-9
異常ケースの理由リスト

異常ケースの理由リスト

理由: 1

ケース	患者ID	理由変数	変数の影響	変数値	変数ノルム
49	8717862852	バーセル2	.071	0	100
162	5546809538	バーセル2	.074	0	100
1010	5284009939	バーセル2	.074	0	100
804	1336411777	治療費	.238	198.25	44.96
974	4626554391	バーセル2	.068	5	100
1124	0428895740	バーセル2	.071	10	100
458	9159094175	バーセル3	.076	10	100
258	9146506080	バーセル2	.074	25	100
639	3962622031	バーセル3	.074	40	100
447	7163481282	バーセル3	.071	0	100
132	7513809991	バーセル2	.083	25	100
354	7216951816	バーセル2	.080	10	100
1015	5857375438	バーセル3	.092	15	100

理由変数は、あるケースが異常ケースとして分類されるのに最も寄与する変数です。各異常ケースに関する最も重要な理由変数が、その影響度、ケースに対する値、および同位グループのノルムとともに表示されます。カテゴリ変数に関する同位グループのノルム（欠損値）は、その同位グループに属する複数のケースが、変数に欠損値を持つことを示しています。

変数の影響を表す統計量は、同位グループでのケースの偏差に対する、理由変数の寄与率を表しています。分析に使用されている変数は、欠損した比率変数を含め 38 個あるため、各変数の影響度の期待値は 1/38 = 0.026 となります。それに対してケース 843 での変数 治療費の影響度は 0.411 と、比較的大きくなっています。また同位グループ 3 での平均値は 19.83 であるのに対して、ケース 843 における 治療費の値は 200.51 となっています。

ここではダイアログ ボックスの設定により、上位 3 つの理由に関する結果が表示されています。

- その他の理由に関する結果を表示する場合は、ダブルクリック操作により表をアクティブにしてください。
- 理由を層次元から行次元に移動します。

図 9-10
異常ケースの理由リスト（先頭の 8 ケース）

異常ケースの理由リスト

理由: 3

ケース	患者ID	理由変数	変数の影響	変数値	変数ノルム
49	8717862852	バーセル3	.066	0	100
162	5546809538	バーセル1	.065	5	80
1010	5284009939	バーセル3	.069	0	100
804	1336411777	バーセル2	.062	45	100
974	4626554391	バーセル1	.067	5	80
1124	0428895740	治療費	.067	121.37	44.96
458	9159094175	バーセル1	.067	5	80
258	9146506080	バーセル1	.067	15	80
678	1299881501	バーセル1	.077	10	80
132	7513809991	バーセル3	.075	35	100
354	7216951816	バーセル3	.075	35	100
1015	5857375438	バーセル1	.076	5	80

この設定により、各ケースに対する上位 3 つの理由の寄与率を容易に比較することができます。ケース 843 は、治療費の値が異常に大きいため、推測したとおり、異常ケースと判断されます。これに対して、ケース 501 の異常度に対する寄与率は、どの理由においても 0.10 以下です。

スケール変数のノルム

図 9-11
スケール変数のノルム

スケール変数ノルム

		同位 ID	
		1	結合
リハビリでの滞在期間	平均値	17.65	17.65
	標準偏差	12.638	12.638
治療とリハビリ合計の治療費(千単位)	平均値	44.9607	44.9607
	標準偏差	26.99713	26.99713
欠損比率	平均値	.035	.035
	標準偏差	.072	.072

スケール変数のノルムには、各変数について、それぞれの同位グループおよび全体における平均値と標準偏差が表示されます。この値を比較することが、同位グループの構成にどの変数が寄与しているかについての目安となります。

たとえば、リハビリでの滞在期間の平均値は、3 つの同位グループすべてでほぼ一定しており、この変数は同位グループの構成に寄与していないことがわかります。それに対し、治療とリハビリ合計の治療費（千単位）と 欠損比率はそれぞれ、同位グループの構成に関する判断材料となります。同位グループ 1 は、治療費の平均値が最も高く、欠損値が最も少なくなっています。また同位グループ 2 は、治療費が全体として非常に低く、欠損値は多くなっています。同位グループ 3 は、治療費も欠損値も中間の値です。

このことから、同位グループ 2 は、病院到着時に死亡していた患者で構成されており、したがって治療費も非常に低く、治療変数およびリハビリ変数がすべて欠損していると推測されます。また同位グループ 3 は、治療中に死亡した患者が多く含まれており、したがって治療費が発生しているもののリハビリの費用はなく、リハビリ変数が欠損していると推測されます。さらに同位グループ 1 は、治療およびリハビリを通して生存していた患者で大部分が構成されており、したがって最も高い治療費が発生したと推測されます。

カテゴリ変数のノルム

図 9-12
カテゴリ変数のノルム（先頭の 10 変数）

カテゴリ変数ノルム

		同位 ID 1	結合
年齢カテゴリ	最も人気のあるカテゴリ	2	2
	度数	424	424
	パーセント	35.8%	35.8%
性別	最も人気のあるカテゴリ	0	0
	度数	592	592
	パーセント	50.0%	50.0%
運動	最も人気のあるカテゴリ	0	0
	度数	596	596
	パーセント	50.4%	50.4%
肥満	最も人気のあるカテゴリ	0	0
	度数	893	893
	パーセント	75.5%	75.5%
糖尿病の病歴	最も人気のあるカテゴリ	0	0
	度数	1061	1061
	パーセント	89.7%	89.7%
血圧	最も人気のあるカテゴリ	1	1
	度数	712	712
	パーセント	60.2%	60.2%
心房細動	最も人気のあるカテゴリ	0	0
	度数	1059	1059
	パーセント	89.5%	89.5%
喫煙者	最も人気のあるカテゴリ	0	0
	度数	911	911
	パーセント	77.0%	77.0%
コレステロール	最も人気のあるカテゴリ	0	0
	度数	669	669
	パーセント	56.6%	56.6%
狭心症の病歴	最も人気のあるカテゴリ	0	0
	度数	794	794
	パーセント	67.1%	67.1%

カテゴリ変数のノルムは、スケール変数のノルムとほとんど同じ役割を果たしますが、このノルムでは、最頻の（度数が最も大きい）カテゴリや、そのカテゴリに属する同位グループ内のケースの数および割合（パーセント）が表示されます。 値の比較は処理が難しい場合があります。たとえば、[喫煙者] の最頻カテゴリが 3 つの同意グループで同じであり、[性別] が同意グループ 3 で異なるため、[性別]が [喫煙者] に比べクラスタ情報に寄与していると考えられる場合があります。 differs on peer group 3.ただし、[性別] の値は 2 つだけであるため、同位グループ 3 のケースの 49.2% の値が 0 となり、他の同位グループのパーセンテージと近くなります。一方、喫煙者の割合の範囲は、72.2 ～ 81.4% となっています。

図 9-13
カテゴリ変数のノルム（選択した変数）

カテゴリ変数ノルム

		同位 ID	
		1	結合
病院での死亡	最も人気のあるカテゴリ	2	2
	度数	424	424
	パーセント	35.8%	35.8%
治療結果	最も人気のあるカテゴリ	0	0
	度数	592	592
	パーセント	50.0%	50.0%
予防的処置手術後	最も人気のあるカテゴリ	0	0
	度数	596	596
	パーセント	50.4%	50.4%
リハビリ後	最も人気のあるカテゴリ	0	0
	度数	893	893
	パーセント	75.5%	75.5%
1ヵ月後のランキンスコア	最も人気のあるカテゴリ	0	0
	度数	1061	1061
	パーセント	89.7%	89.7%
3ヵ月後のランキンスコア	最も人気のあるカテゴリ	1	1
	度数	712	712
	パーセント	60.2%	60.2%
6ヵ月後のランキンスコア	最も人気のあるカテゴリ	0	0
	度数	1059	1059
	パーセント	89.5%	89.5%
1ヵ月後のバーセルインデックス	最も人気のあるカテゴリ	0	0
	度数	911	911
	パーセント	77.0%	77.0%
3ヵ月後のバーセルインデックス	最も人気のあるカテゴリ	0	0
	度数	669	669
	パーセント	56.6%	56.6%
6ヵ月後のバーセルインデックス	最も人気のあるカテゴリ	0	0
	度数	794	794
	パーセント	67.1%	67.1%

スケール変数ノルムから推測された事実について、[カテゴリ変数ノルム]表でさらに詳しく確認します。同位グループ 2 は、すべて病院到着時に死亡していた患者で構成されているため、治療変数とリハビリ変数が欠損しています。同位グループ 3 に属する患者は、その多く（69.0%）が治療中に死亡した患者であるため、リハビリ変数に対する最頻カテゴリは（欠損値）となります。

異常指数の要約

図 9-14
異常指数の要約

異常指数の要約

	異常リスト内の項目数	最小値	最大値	平均値	標準偏差
異常指数	59	1.324	1.599	1.401	.064

異常リスト内の項目数は、指定によって決まります。異常のパーセントは 5% です。

この表には、異常リストに含まれるケースの異常指数値に対する要約統計量が表示されます。

理由の要約

図 9-15
理由の要約 (治療変数とリハビリ変数)

Reason 1

	理由としての出現		変数の影響の統計量			
	度数	パーセント	最小値	最大値	平均値	標準偏差
年齢カテゴリ	0	.0%	.	.	.	.
性別	0	.0%	.	.	.	.
運動	0	.0%	.	.	.	.
肥満	0	.0%	.	.	.	.
糖尿病の病歴	0	.0%	.	.	.	.
血圧	0	.0%	.	.	.	.
心房細動	0	.0%	.	.	.	.
喫煙者	0	.0%	.	.	.	.
コレステロール	0	.0%	.	.	.	.
狭心症の病歴	0	.0%	.	.	.	.
心筋梗塞の病歴	0	.0%	.	.	.	.
ニトログリセリンの処方	0	.0%	.	.	.	.
抗凝固薬の摂取	0	.0%	.	.	.	.
一過性脳虚血発作の病歴	0	.0%	.	.	.	.
病院までの時間	0	.0%	.	.	.	.
病院到着時死亡	0	.0%	.	.	.	.
初期のランキンスコア	0	.0%	.	.	.	.
CAT スキャンの結果	0	.0%	.	.	.	.
血栓溶解薬	0	.0%	.	.	.	.
病院での死亡	0	.0%	.	.	.	.
治療結果	0	.0%	.	.	.	.
予防的処置手術後	0	.0%	.	.	.	.
リハビリ後	0	.0%	.	.	.	.
1ヵ月後のランキンスコア	0	.0%	.	.	.	.
3ヵ月後のランキンスコア	0	.0%	.	.	.	.
6ヵ月後のランキンスコア	0	.0%	.	.	.	.
1ヵ月後のバーセルインデックス	6	10.2%	.063	.077	.069	.007
3ヵ月後のバーセルインデックス	27	45.8%	.068	.085	.077	.005
6ヵ月後のバーセルインデックス	21	35.6%	.070	.092	.078	.005
1ヵ月後のレコードバーセルインデックス	0	.0%	.	.	.	.
3ヵ月後のレコードバーセルインデックス	0	.0%	.	.	.	.
6ヵ月後のレコードバーセルインデックス	0	.0%	.	.	.	.
0 to 100 by 5年齢	0	.0%	.	.	.	.
2度死亡: (DOA = 1) & (コレステロール=1)	0	.0%	.	.	.	.
リハビリでの滞在期間	0	.0%	.	.	.	.
治療とリハビリ合計の治療費(千単位)	4	6.8%	.069	.273	.164	.106
欠損比率	1	1.7%	.078	.078	.078	.
全体	59	100.0%	.063	.273	.083	.033

この表には、分析に使用される変数ごとに、その主要な理由としての役割がまとめて表示されます。病院到着時死亡から リハビリ後までの変数を含め、大部分の変数は、いずれかのケースが異常リストに含まれる主要な理由とはなりません。理由になっているものとしては 1 か月目のバーセ

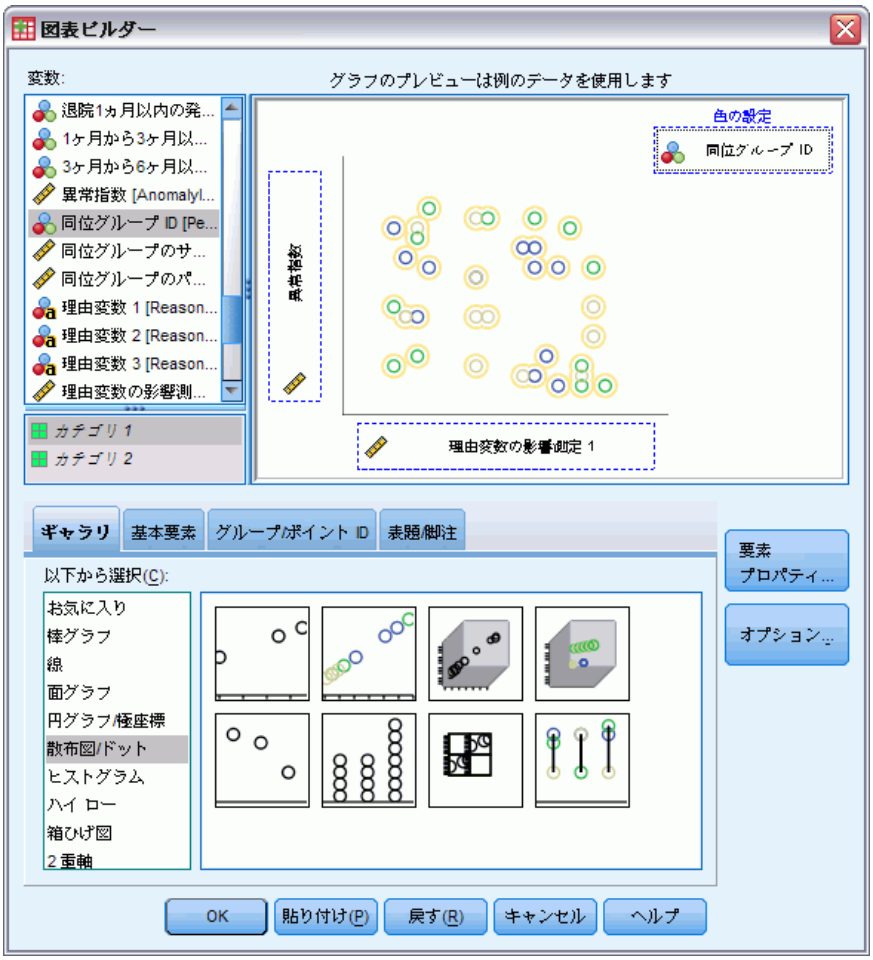
ルインデックスが最も多く、次いで多いのが 治療とリハビリ合計の治療費（千単位）です。変数の影響度を表す統計量としては、各変数に関する影響度の最大値、最小値、平均値のほか、複数のケースにおいて理由となった変数についての標準偏差が表示されます。

変数の影響度による異常指数の散布図

表には有用な情報が数多く含まれていますが、各情報間の関係を把握するのが困難な場合もあります。保存されている変数を使用してグラフを作成することにより、各情報間の関係を理解しやすくなります。

▶ この散布図を作成するには、メニューから次の項目を選択します。
グラフ(G) > 図表ビルダー(C)...

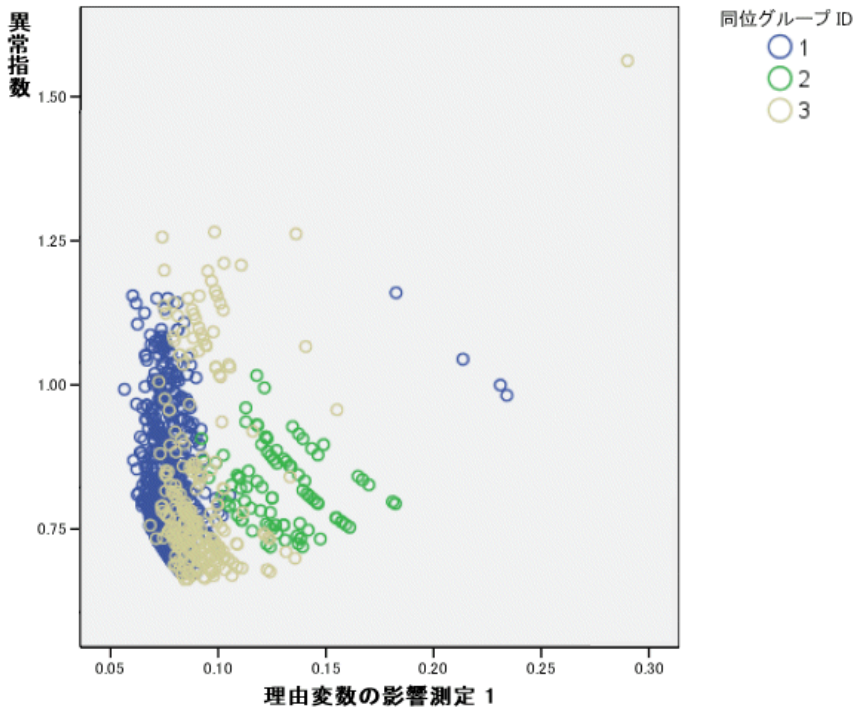
図 9-16
[図表ビルダー] ダイアログ ボックス



- ▶ [散布図/ドット] ギャラリを選択し、[グループ化散布図] アイコンをキャンバス上にドラッグします。
- ▶ y 変数として「異常指数」を、x 変数として「理由変数の影響測定 1」をそれぞれ選択します。
- ▶ 色を設定するための変数として、「同位グループ ID」を選択します。
- ▶ [OK] をクリックします。

この選択により、散布図が作成されます。

図 9-17
最初の理由変数の影響度測定による異常指数の散布図



グラフを調べることで、いくつかの事実を観測できます。

- 右上隅に存在するケースは同位グループ 3 に属し、最も異常度の高いケースであると同時に、単一変数の寄与が最も大きいケースでもあります。
- y 軸に沿って下方に移動すると、同位グループ 3 に属するケースが 3 つあり、いずれも異常指数値が 2.00 をわずかに上回っています。これらのケースは、異常ケースとしてさらに詳しく調べる必要があります。
- x 軸に沿って移動すると、同位グループ 1 に属するケースが 4 つあり、いずれも変数影響度の測定値がほぼ 0.23 から 0.33 の間に存在します。これらのケースは、こうした値により散布図の大部分の点から孤立しているため、さらに詳しく調べる必要があります。
- 同位グループ 2 は、異常指数および変数影響度の値の中に中心傾向から大きく外れたものがなく、その意味でかなりの等質性を持つと思われます。

要約

［例外ケースの特定］手続きを使用することにより、さらに検証が必要なケースをいくつか特定しました。異常ケースかどうかは、（変数の値そのものだけでなく）変数間の関係に基づいて判断されるため、ここで特定されたケースは、その他の検証手続きでは特定できないケースです。

同位グループがほとんど 2 つの変数病院到着時死亡と病院での死亡に基づいて構築されている場合があります。さらに詳しい分析としては、たとえば、作成する同位グループの数を増加させることによって現れる影響を調べる、または治療によって命を取りとめた患者だけを対象とした分析を行う、などができます。

関連手続き

［例外ケースの特定］手続きは、データ ファイル内の異常ケースを検出するための有用な手段です。

- ［データの検証］手続きは、アクティブなデータセット内で、無効の疑いがあるかまたは実際に無効なケース、変数、およびデータ値を特定するためのものです。

章
10

最適カテゴリ化

［最適カテゴリ化］手続きは、各スケール変数の値をビンに分配して、1 つ以上のスケール変数（以下 **ビン（分割）入力変数**と呼びます）を離散化するためのものです。ビンの構成は、ビン分割プロセスを「監視」するカテゴリ ガイド変数に基づいて最適化されます。元のデータ値の代わりにビンを使用することにより、カテゴリ変数を使用することが必須または適切な手続きを使ってさらに詳しい分析を行えます。

最適カテゴリ化のアルゴリズム

最適カテゴリ化のアルゴリズムの基本的な手順は次のとおりです。

前処理（省略可）。 ビン（分割）入力変数は n 個のビンに分割されます（n は任意に指定する数値）。それぞれのビンには、同数または可能な範囲で同数に近いケースが含まれます。

分割点の候補の特定。 ビン（分割）入力変数の値のうち、その次に大きな値と同じガイド変数のカテゴリには属さないものが、分割点の候補となります。

分割点の選択。 分割点の候補のうち情報利得が最大になるものに対して、MDLP 判定基準による評価が行われます。判定基準を満たす分割点の候補がなくなるまで、繰り返し評価が行われます。判定基準を満たした分割点が、ビンの終点となります。

最適カテゴリ化による融資申請者データの離散化

銀行の融資担当者は、債務不履行率を低減させる取り組みの一環として、債務不履行の確率を予測するモデルを作成するため、過去および現在の顧客に関する財務情報と人口統計情報を収集しました。予測変数の候補としてスケール変数が使用できますが、融資担当者は、カテゴリ予測変数を使って適切な処理のできるモデルにしたいと考えています。

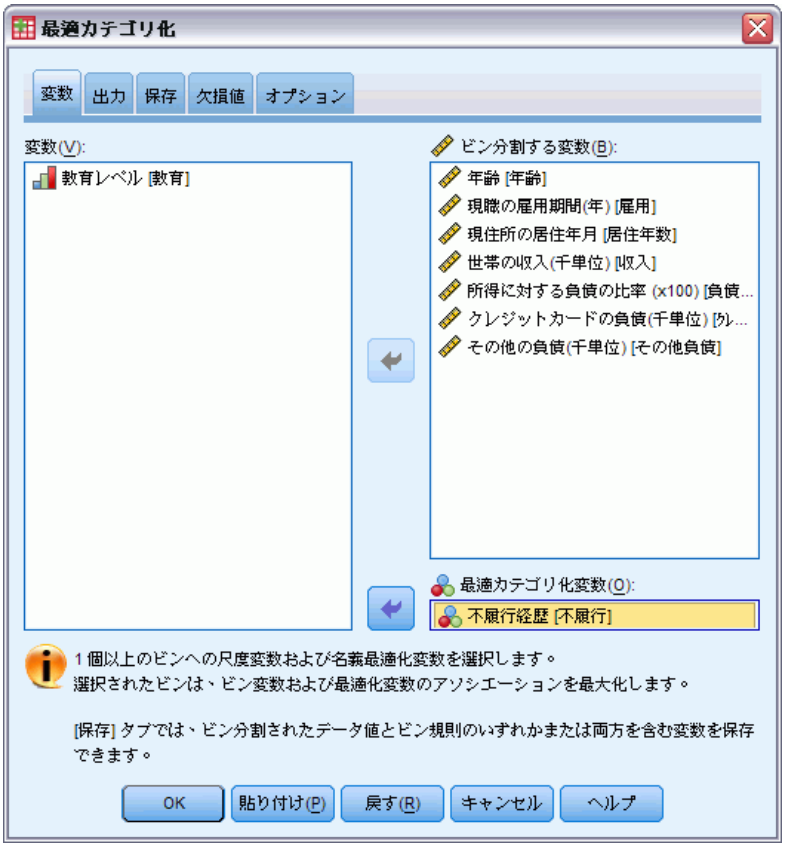
過去の顧客 5000 人分の情報はすべて、bankloan_binning.sav に収集されています。詳細は、 A 付録 p.149 サンプル ファイル を参照してください。［最適カテゴリ化］手続きを使用してスケール予測変数のビン規則を生成し、その規則に基づいて bankloan.sav の処理を行います。さらに処理されたデータセットを使用することで、予測モデルを作成できます。

分析の実行

▶ ［最適カテゴリ化］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。

変換(T) > 最適カテゴリ化...

図 10-1
[最適カテゴリ化] ダイアログ ボックスの [変数] タブ



▶ ビン分割する変数として、「年齢」、および「現職の雇用期間（年)」から「その他の負債（千単位)」までの変数を選択します。

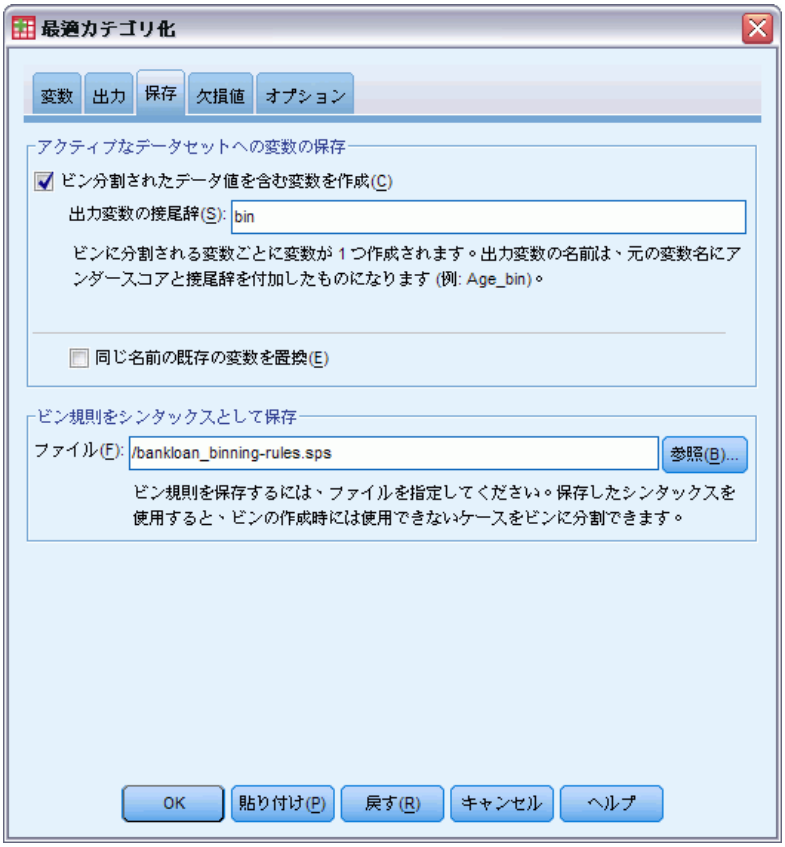
▶ またガイド変数として、「不履行履歴」を選択します。

▶ ［出力］タブをクリックします。

図 10-2
[最適カテゴリ化] ダイアログ ボックスの [出力] タブ

- ビン分割される変数に対して、[**記述統計量**] および [**モデル エントロピー**] を選択します。
- [**保存**] タブをクリックします。

図 10-3
[最適カテゴリ化] ダイアログ ボックスの [保存] タブ



- [ビン分割されたデータ値を含む変数を作成] を選択します。
- 生成されたビン規則を保存するシンタックス ファイルのパスおよびファイル名を入力します。この例では、/bankloan_binning-rules.sps を使用しました。
- [OK] をクリックします。

以上の選択により、次のコマンド シンタックスが生成されます。

```
* Optimal Binning.
OPTIMAL BINNING
  /VARIABLES GUIDE=default BIN=age employ address income debtinc creddebt
  othdebt SAVE=YES (INTO=age_bin employ_bin address_bin income_bin debtinc_bi
  creddebt_bin othdebt_bin)
  /CRITERIA METHOD=MDLP
   PREPROCESS=EQUALFREQ (BINS=1000)
   FORCEMERGE=0
   LOWERLIMIT=INCLUSIVE
   LOWEREND=UNBOUNDED
   UPPEREND=UNBOUNDED
  /MISSING SCOPE=PAIRWISE
```

```
/OUTFILE RULES='/bankloan_binning-rules.sps'
/PRINT ENDPOINTS DESCRIPTIVES ENTROPY.
```

- この手続きでは、ガイド変数不履行を基に MDLP ビン分割を使用して、年齢、雇用、居住年数、収入、負債比、クレジット負債、その他負債の各ビン（分割）入力変数を離散化します。
- これらの変数に関して離散化された値はそれぞれ、新しい変数「年齢_bin」、「employ_bin（雇用_bin)」、「居住年数_bin」、「収入_bin」、「負債比_bin」、「クレジット負債_bin」、および「その他負債_bin」に格納されます。
- ビン（分割）入力変数の値の個数が 1000 を超える場合は、等度数法によって値の個数を 1000 にした上で、MDLP ビン分割が実行されます。
- ビン規則を表すコマンド シンタックスは、/bankloan_binning-rules.sps というファイルに保存されています。
- ビン（分割）入力変数に対しては、ビンの終点、記述統計量、およびモデル エントロピーの値が必要です。
- その他のビン分割条件には、それぞれのデフォルト値が設定されます。

記述統計

図 10-4
記述統計(S)

	N	最小値	最大値	異なる数値の数	ビン (分割) の数
年齢	5000	20	58	39	2
現職の雇用期間(年)	5000	0	38	39	4
現住所の居住年月	5000	0	37	38	3
世帯の収入(千単位)	5000	12.10	2461.70	1100	2
所得に対する負債の比率 (x100)	5000	.08	44.62	2060	5
クレジットカードの負債(千単位)	5000	.01	139.58	5000	4
その他の負債(千単位)	5000	.01	416.52	4999	2

記述統計量表には、ビン（分割）入力変数に関する要約情報が表示されます。先頭の 4 つの列は、ビン分割前の値に関するものです。

- [N] は、分析に使用されるケースの数を表します。欠損値をリストごとに削除する場合、この値はすべての変数に対して一定になります。欠損値をペアごとに削除する場合は、必ずしもこの値は一定になりません。このデータセットには欠損値が含まれていないため、この値はケースの数そのものに一致します。

- [**最小値**] 列および [**最大値**] 列には、各ビン（分割）入力変数に対する、データセット内の（ビン分割前の）最小値および最大値が表示されます。これらの値は、各変数に対する観測値の範囲を確定するだけでなく、期待範囲に含まれない値を特定する場合にも利用されます。
- [**異なる数値の数**] では、等度数アルゴリズムを使用して前処理されたのはどの変数かを知ることができます。デフォルトでは、変数（世帯の収入（千単位）から その他の負債（千単位）まで）のうち、異なる値の数が 1000 個を超えるものは、事前に 1000 個のビンに分割されます。前処理されたこれらのビンは、ガイド変数に基づき MDLP 法に従ってビンに分割されます。前処理機能については、[オプション] タブで設定できます。
- [**ビン（分割）の数**] は、手続きを通して最終的に生成されたビンの数であり、異なる値の数よりもはるかに少なくなります。

モデル エントロピー

図 10-5
モデル エントロピー

	モデル エントロピー
年齢	.788
現職の雇用期間(年)	.754
現住所の居住年月	.781
世帯の収入(千単位)	.803
所得に対する負債の比率 (x100)	.711
クレジットカードの負債(千単位)	.776
その他の負債(千単位)	.801

モデル エントロピーが小さいほど、ガイド変数 不履行経歴に基づくビン化変数の予測精度が高いことを表します。

モデル エントロピーは、債務不履行の確率に関する予測モデルにおいて、各変数がどの程度の有用性を持つかの目安になります。

- 予測変数としては、生成されたビンごとにガイド変数と同じ値を持つケースを含み、それによってガイド変数が完全に予測できるようなものが最も理想的です。ただし、このような予測変数のモデル エントロピーは定義されません。通常、このような状況が現実に起こることはなく、起こったとすれば個別データの質に問題があると考えられます。
- 一方、最も不適当な予測変数は、値を予測する根拠がほとんど見当たらないようなものです。この場合のモデル エントロピーの値は、データによって異なります。このデータセットでは、全部で 5000 人の顧客のうち、1256 人 (0.2512) が債務不履行となっており、3744 人 (0.7488) が債務不履行とはなっていません。したがって、予測変数が最も不適当なものであるとすれば、そのモデル エントロピーは、$-0.2512 \times \log_2(0.2512) - 0.7488 \times \log_2(0.7488) = 0.8132$ となります。

よいモデル エントロピーを生み出す要素はアプリケーションやデータによって異なるため、「モデル エントロピーの値が低い変数は予測変数に適している」という事実をより具体的な形で表現することは困難です。ここでは、異なるカテゴリの数に比べ生成されたビンの数が多い変数ほど、モデル エントロピーの値が小さくなっていると考えられます。これらのビン（分割）入力変数に対しては、より高度な手段で変数を選択するための予測モデル手続きを使用して、予測変数としてのさらに詳しい評価が行われます。

ビンの要約

ビンの要約では、ガイド変数の値に基づいて、生成されたビンの上限と下限、および各ビンの度数が表示されます。ビンの要約表は、各ビン（分割）入力変数に対して個別に作成されます。

図 10-6
年齢に関するビンの要約

ビン (分割)	終点		不履行経歴 の水準によるケースの数		
	下限	上限	なし	あり	合計
1	[a]	32	1129	639	1768
2	32	[a]	2615	617	3232
合計			3744	1256	5000

各ビンは、下限 <= 年齢 < 上限として計算されます。
a. 限界値なし

[年齢] の要約は、32 歳以下の 1768 名の顧客はビン 1 に分類され、32 歳以上の 3232 名の顧客はビン 2 二分プイされます。 以前不履行の履歴がある顧客の割合は、ビン 2（617/3232=0.191）よりビン 1（639/1768=0.361）の方が非常に大きくなります。

図 10-7
世帯の収入（千単位）に関するビンの要約

ビン (分割)	終点		不履行経歴 の水準によるケースの数		
	下限	上限	なし	あり	合計
1	[a]	26.70	1054	513	1567
2	26.70	[a]	2690	743	3433
合計			3744	1256	5000

各ビンは、下限 <= 世帯の収入(千単位) < 上限として計算されます。
a. 限界値なし

世帯の収入（千単位）に関する要約でも、26.70 を唯一の分割点として、すでに債務不履行となっている顧客の比率がビン 2（743/3433=0.216）よりもビン 1（513/1567=0.327）の方で高くなっており、前記と同じようなパターンが見られます。ただし、モデル エントロピー統計量から予想されるように、これらの比率の違いは 年齢ほど大きくはありません。

図 10-8
その他の負債（千単位）に関するビンの要約

ビン (分割)	終点		不履行経歴 の水準によるケースの数		
	下限	上限	なし	あり	合計
1	[a]	2.19	2161	539	2700
2	2.19	[a]	1583	717	2300
合計			3744	1256	5000

各ビンは、下限 <= その他の負債(千単位) < 上限として計算されます。
a. 限界値なし

その他の負債（千単位）に関する要約では、2.19 を唯一の分割点として、すでに債務不履行となっている顧客の比率がビン 2（717/2300=0.312）よりもビン 1（539/2700=0.200）の方で低くなっており、前記とは逆のパターンを示しています。ここでも、モデル エントロピー統計量から予想されるように、これらの比率の違いは 年齢ほど大きくはありません。

図 10-9
現職の雇用期間（年）に関するビンの要約

ビン (分割)	終点		不履行経歴 の水準によるケースの数		
	下限	上限	なし	あり	合計
1	[a]	3	629	478	1107
2	3	8	1066	461	1527
3	8	18	1471	268	1739
4	18	[a]	578	49	627
合計			3744	1256	5000

各ビンは、下限 <= 現職の雇用期間(年) < 上限として計算されます。
a. 限界値なし

現職の雇用期間（年）に関するビンの要約を見ると、ビン番号が増加するにつれて債務不履行者の比率が減少するというパターンがあることがわかります。

ビン	債務不履行者の比率
1	0.432
2	0.302
3	0.154
4	0.078

図 10-10
現住所の居住年月に関するビンの要約

ビン(分割)	終点		不履行経歴の水準によるケースの数		
	下限	上限	なし	あり	合計
1	1[a]	7	1652	829	2481
2	7	14	1184	313	1497
3	14	[a]	908	114	1022
合計			3744	1256	5000

各ビンは、下限 <= 現住所の居住年月 < 上限として計算されます。
a. 限界値なし

現住所の居住年月に関する要約でも、同様のパターンが見られます。モデル エントロピー統計量から予想されるように、債務不履行者の比率におけるビン間の差は、現住所の居住年月よりも 現職の雇用期間（年）の方が著しくなっています。

ビン	債務不履行者の比率
1	0.334
2	0.209
3	0.112

図 10-11
クレジット カードの負債（千単位）に関するビンの要約

ビン (分割)	終点		不履行経歴の水準によるケースの数		
	下限	上限	なし	あり	合計
1	[a]	.97	2169	466	2635
2	.97	1.91	848	307	1155
3	1.91	6.05	643	352	995
4	6.05	[a]	84	131	215
合計			3744	1256	5000

各ビンは、下限 <= クレジットカードの負債(千単位) < 上限として計算されます。
a. 限界値なし

クレジット カードの負債（千単位）に関するビンの要約では逆に、ビン番号が増加するにつれて債務不履行者の比率が増加するというパターンが見られます。現職の雇用期間（年）と 現住所の居住年月は、債務不履行者にならない確率が高い顧客の特定に適しているのに対し、クレジット カードの負債（千単位）は、債務不履行者になる確率が高い顧客の特定に適していると考えられます。

ビン	債務不履行者の比率
1	0.177
2	0.266
3	0.354
4	0.609

図 10-12
所得に対する負債の比率 (x100) に関するビンの要約

ビン (分割)	終点		不履行経歴 の水準によるケースの数		
	下限	上限	なし	あり	合計
1	[a]	4.39	912	88	1000
2	4.39	12.09	2006	437	2443
3	12.09	18.71	625	386	1011
4	18.71	31.00	198	303	501
5	31.00	[a]	3	42	45
合計			3744	1256	5000

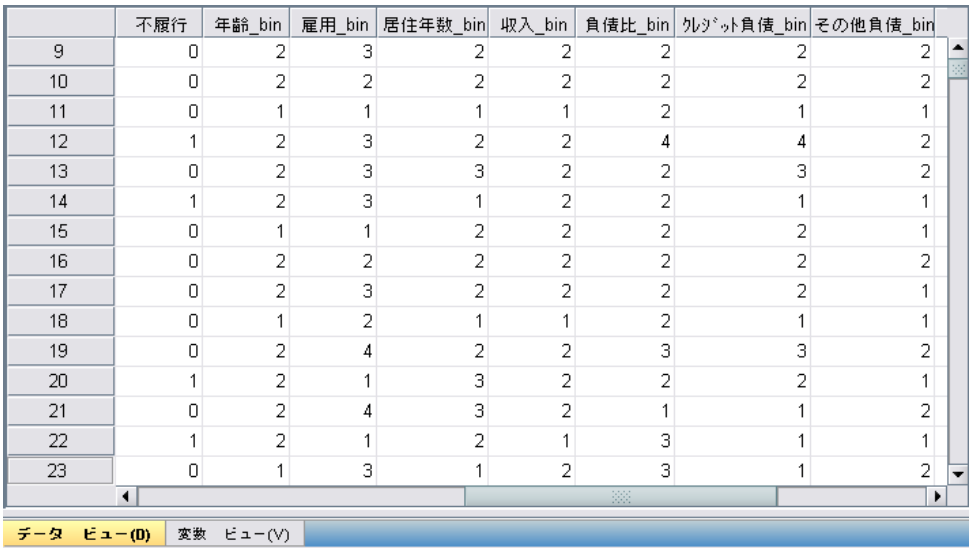
各ビンは、下限 <= 所得に対する負債の比率 (x100) < 上限として計算されます。
a. 限界値なし

所得に対する負債の比率（x100）に関するビンの要約では、クレジットカードの負債（千単位）と同様のパターンが見られます。この変数は、モデル エントロピーの値が最も低く、したがって債務不履行の確率に関する予測変数としては最適です。この変数は、債務不履行者になる確率が高い顧客を分類する上では クレジット カードの負債（千単位）よりも優れており、債務不履行者になる確率が低い顧客を分類する上では 現職の雇用期間（年）と同等です。

ビン	債務不履行者の比率
1	0.088
2	0.179
3	0.382
4	0.605
5	0.933

ビン分割

図 10-13
データ エディタでの bankloan_binning.sav に対するビン分割



	不履行	年齢_bin	雇用_bin	居住年数_bin	収入_bin	負債比_bin	ｸﾚｼﾞｯﾄ負債_bin	その他負債_bin
9	0	2	3	2	2	2	2	2
10	0	2	2	2	2	2	2	2
11	0	1	1	1	1	2	1	1
12	1	2	3	2	2	4	4	2
13	0	2	3	3	2	2	3	2
14	1	2	3	1	2	2	1	1
15	0	1	1	2	2	2	2	1
16	0	2	2	2	2	2	2	2
17	0	2	3	2	2	2	2	1
18	0	1	2	1	1	2	1	1
19	0	2	4	2	2	3	3	2
20	1	2	1	3	2	2	2	1
21	0	2	4	3	2	1	1	2
22	1	2	1	2	1	3	1	1
23	0	1	3	1	2	3	1	2

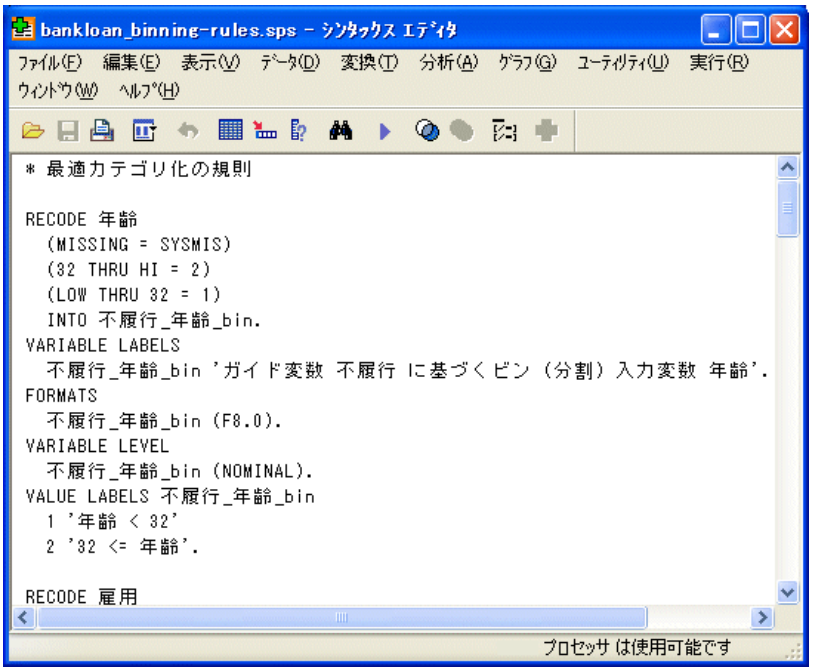
このデータセットにおけるビン分割プロセスの結果が、データ エディタに表示されています。これらのビン分割は、記述統計手続きやレポート手続きを使用し、ビン分割の結果についてカスタマイズした要約情報を作成する場合には有用ですが、これらのケースを使用してビン規則が生成されているため、このデータセットを使用して予測モデルを構成することは勧められません。代替案の 1 つとして、他の顧客に関する情報が保存されている別のデータセットにビン規則を適用することをお勧めします。

シンタックス形式のビン規則の適用

［最適カテゴリ化］手続きを実行中、手続きにより生成されたビン規則を、コマンド シンタックスとして保存するよう指定しました。

- bankloan_binning-rules.sps を開きます。

図 10-14
シンタックス形式の規則ファイル



ビン（分割）入力変数ごとに、ビン分割を実行するコマンド シンタックスのブロックがあります。このブロックでは、変数のラベル、書式およびレベルの設定や、ビンに対する値ラベルの設定が行われます。これらのコマンドは、bankloan_binning.sav と同じ変数を持つ任意のデータセットに適用できます。

- bankloan.sav を開きます。詳細は、 A 付録 p.149 サンプル ファイル を参照してください。
- bankloan_binning-rules.sps の［シンタックス エディタ］ビューに戻ります。

▶ ビン規則を適用するには、シンタックス エディタのメニューから次の項目を選択します。
実行(R) > すべて...

図 10-15
データ エディタでの bankloan.sav に対するビン分割

	不履行予測3	年齢_bin	雇用_bin	居住年数_bin	収入_bin	負債比_bin	ｸﾚｼﾞｯﾄ負債_bin	その他負債_bin
1	0.21304	2	3	2	2	2	4	2
2	0.43690	1	3	1	2	3	2	2
3	0.14102	2	3	3	2	2	1	1
4	0.10442	2	3	3	2	1	3	1
5	0.43690	1	1	1	2	3	2	2
6	0.23358	2	2	1	1	2	1	1
7	0.81709	2	4	2	2	4	3	2
8	0.11336	2	3	2	2	1	1	1
9	0.66390	1	2	1	1	4	2	2
10	0.51553	2	1	2	1	4	3	1
11	0.09055	1	1	1	1	1	1	1
12	0.13631	1	2	1	1	2	1	1
13	0.22890	2	4	3	2	2	3	2
14	0.40484	2	2	2	2	3	2	2
15	0.20866	2	4	3	2	2	3	2

データ ビュー(D) 変数 ビュー(V)

bankloan_binning.sav で［最適カテゴリ化］手続きを実行して生成された規則に基づいて、bankloan.sav に含まれる変数がビン分割されました。これによりこのデータセットは、カテゴリ変数を使用することが適切または必須の予測モデルを構成する際に使用できます。

要約

［最適カテゴリ化］手続きを使用して、債務不履行の確率に関する予測変数の候補となるスケール変数のビン規則を生成し、それを別のデータセットに適用しました。

ビン分割のプロセスを通じて指摘したことは、ビン分割された 現職の雇用期間（年）および 現住所の居住年月は、債務不履行者にならない確率が高い顧客を特定するのに適しており、クレジット カードの負債（千単位）は、債務不履行者になる確率が高い顧客を特定するのに適しているということでした。この興味深い事実は、債務不履行者の確率に関する予測モデルを構成する際に、なんらかの新たな手掛かりを与えてくれるでしょう。ただし資金回収不能の回避を優先する場合は、現職の雇用期間（年）や 現住所の居住年月よりも、クレジット カードの負債（千単位）の方が重要な変数となります。また、顧客基盤の拡大を優先する場合は、現職の雇用期間（年）や 現住所の居住年月が重要な変数となります。

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 (Van der Ham, Meulman, Van Strien, および Van Engeland, 1997) 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 (Price および Bouffard, 1974)。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 (Green および Rao, 1972)。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 (McCullagh および Nelder, 1989) 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（Green および Wind, 1973）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 (Kennedy, Riquier, および Sharp, 1996) 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” (Rickman, Mitchell, Dingman, および Dalen, 1974) 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 (Blake および Merz, 1998) 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 (Bell, 1961) は、予想される社会グループを示す表を作成しました。Guttman (Guttman, 1968)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 (Rosenberg および Kim, 1975) 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 (Breiman および Friedman(F), 1985)、(Hastie および Tibshirani, 1990)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）(Hartigan, 1975) 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（McCullagh など，1989）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 (Collett, 2003) 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 (Collett など, 2003) 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（Verdegaal, 1985）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（Ware, Dockery, Spiro III, Speizer, および Ferris Jr., 1984）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。

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.

参考文献

Bell, E. H. 1961. Social foundations of human behavior:Introduction to the study of sociology. New York: Harper & Row.

Blake, C. L., および C. J. Merz. 1998. "UCI Repository of machine learning databases." Available at http://www.ics.uci.edu/~mlearn/MLRepository.html.

Breiman, L., および J. H. Friedman(F). 1985. Estimating optimal transformations for multiple regression and correlation. Journal of the American Statistical Association, 80, .

Collett, D. 2003. Modelling survival data in medical research, 2 ed. Boca Raton: Chapman & Hall/CRC.

Green, P. E., および V. Rao. 1972. Applied multidimensional scaling. Hinsdale, Ill.: Dryden Press.

Green, P. E., および Y. Wind. 1973. Multiattribute decisions in marketing:A measurement approach. Hinsdale, Ill.: Dryden Press.

Guttman, L. 1968. A general nonmetric technique for finding the smallest coordinate space for configurations of points. Psychometrika, 33, .

Hartigan, J. A. 1975. Clustering algorithms. New York: John Wiley and Sons.

Hastie, T., および R. Tibshirani. 1990. Generalized additive models. London: Chapman and Hall.

Kennedy, R., C. Riquier, および B. Sharp. 1996. Practical applications of correspondence analysis to categorical data in market research. Journal of Targeting, Measurement, and Analysis for Marketing, 5, .

McCullagh, P., および J. A. Nelder. 1989. Generalized Linear Models, 2nd ed. London: Chapman & Hall.

Price, R. H., および D. L. Bouffard. 1974. Behavioral appropriateness and situational constraints as dimensions of social behavior. Journal of Personality and Social Psychology, 30, .

Rickman, R., N. Mitchell, J. Dingman, および J. E. Dalen. 1974. Changes in serum cholesterol during the Stillman Diet. Journal of the American Medical Association, 228, .

Rosenberg, S., および M. P. Kim. 1975. The method of sorting as a data-gathering procedure in multivariate research. Multivariate Behavioral Research, 10, .

Van der Ham, T., J. J. Meulman, D. C. Van Strien, および H. Van Engeland. 1997. Empirically based subgrouping of eating disorders in adolescents:A longitudinal perspective. British Journal of Psychiatry, 170, .

Verdegaal, R. 1985. Meer sets analyse voor kwalitatieve gegevens (in Dutch). Leiden: Department of Data Theory, University of Leiden.

Ware, J. H., D. W. Dockery, A. Spiro III, F. E. Speizer, および B. G. Ferris Jr.. 1984. Passive smoking, gas cooking, and respiratory health of children living in six cities. American Review of Respiratory Diseases, 129, .

索引