IBM SPSS Data Preparation 26

注記

本書および本書で紹介する製品をご使用になる前に、 7 ページの『特記事項』に記載されている情報をお読みください。

本書は、IBM® SPSS Statistics バージョン 26 リリース 0 モディフィケーション 0 および新しい版で明記されない限り、以降のすべてのリリースおよびモディフィケーションに適用されます。

お客様の環境によっては、資料中の円記号がバックスラッシュと表示されたり、バックスラッシュが円記号と表示されたりする場合があります。

原典： IBM SPSS Data Preparation 26

発行： 日本アイ・ビー・エム株式会社

担当： トランスレーション・サービス・センター

目次

データ準備

以下のデータ準備機能が、SPSS® Statistics Professional Edition または Data Preparation オプションに含まれています。

データ準備の概要

演算システムの処理能力の向上に伴い、それに比例して情報に対する需要も増大するため、より多くのデータ収集され、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予測モデルの予測における問題となるため、データを「クリーン」に保つ必要があります。ただし、増大するウェアハウス格納データの量は、ケースを手動で確認する能力を遥かに超えているため、データ検証の自動プロセスを実装することが不可欠です。

データ準備アドオン 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。

データ準備プロシージャの使用

データ準備プロシージャの使用方法は、目的に応じて異なります。データのロード後の標準的な処理の順序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、実施する価値があります。

データの検証
: 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基づくチェックを実行します。無効なデータが見つかったら、原因を調べ、修正します。これには、メタデータの準備におい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、モデル作成を向上させ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可能性がある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が原因で発生します。これには、メタデータの準備におい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「クリーン」になったら、他のアドオン モジュールからモデルを作成できます。

例外ケースの特定

異常検出プロシージャでは、クラスタ グループの基準値からの偏差に基づいて例外的なケースを検索します。このプロシージャーは、任意の推論的データ分析に先立つ予備的なデータ分析ステップで、データ監査の目的で例外的なケースを素早く検出するために設計されています。このアルゴリズムは一般的な異常値検出のために設計されています。つまり、異常ケースの定義は、異常値の定義が適切にできる、医療保険業界での普通でない支払いパターンの検出や金融業界での不正資金浄化 (マネー・ロンダリング) の検出などのような特定の用途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、予測モデルを作成するデータ・アナリストはデータの品質に注意します。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、データ検証のプロシージャーでは検出できない観測値もあります。「例外ケースの特定」プロシージャーは、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

このプロシージャーは、ピア・グループ、連続型変数とカテゴリー変数のピア・グループ・ノルム、ピア・グループ・ノルムの偏差に基づく異常値指標、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: このプロシージャーは、連続型変数およびカテゴリー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、それぞれの列はピア・グループの基となる異なる変数を表します。データ・ファイルでは出力をマークするためにケース識別変数を使用できますが、分析では使用されません。欠損値は使用できます。重み付け変数が指定されている場合、重み付け変数は無視されます。

検出モデルは、新しい検定データ・ファイルに適用できます。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、モデルの作成に使用される欠損値の処理が、スコアリングの前に検定データ・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ソートされたケースを使用していくつかの異なる解を取得します。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ソートさ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、および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。各連続型変数は正規（ガウス）分布であると仮定し、各カテゴリー変数は多項分布であると仮定します。経験的内部検定は、このプロシージャーが独立仮定および分布仮定の両方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。

例外ケースの特定

1. 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...」
2. 1 つ以上の分析変数を選択します。
3. オプションで、出力のラベル付けに使用するケース識別子変数を選択します。
4. 「適用」をクリックします。

不明な尺度のフィールド

データセット内の 1 つ以上の変数 (フィールド) の測定レベルが不明な場合、測定レベルの警告が表示されます。測定レベルはこの手続きの結果の計算に影響を与えるため、すべての変数について測定レベルを定義する必要があります。

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、デフォルトの測定レベルを、測定レベルが現在不明なすべてのフィールドに割り当てます。データ・セットのサイズが大きい場合、この処理には時間がかかります。

手動で割り当て
: 不明な測定レベルを持つフィールドをすべてリストします。測定レベル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「変数リスト」ペインでも測定レベルを割り当てることができます。

この手続きでは測定レベルが重要であるため、すべてのフィールドに対して測定レベルが定義されるまで、この手続きを実行することはできません。

例外ケースの特定: 出力

「出力」ダイアログは、表形式の出力を生成するためのオプションを提供します。

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
: このオプションを選択すると、3 つの表が作成されます。

- 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値指標の値を表示します。
- 異常ケースのピア ID リストは、例外ケースと、それに対応するピア・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、異常値指標の降順でソートされます。さらに、「変数」ダイアログでケース識別子変数が指定されている場合は、ケース識別子が表示されます。

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、「連続型変数ノルム」表 (分析で連続型変数が使用されている場合) または「カテゴリー変数ノルム」表 (分析でカテゴリー変数が使用されている場合) が表示されます。「連続型変数ノルム」表には、ピア・グループごとに、各連続型変数の平均および標準偏差が表示されます。また「カテゴリー変数ノルム」表には、ピア・グループごとに、各カテゴリー変数の最頻値 (度数が最も大きいカテゴリー)、度数、および度数パーセントが表示されます。連続型変数の平均とカテゴリー変数の最頻値は、分析のノルム値として使用されます。

異常指数
: 異常値指標の要約には、異常度が最も高いと特定されたケースの異常値指標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。

ケースが処理されました
: ケース処理要約には、アクティブなデータ・セットにおけるすべてのケースの度数とその度数のパーセント、分析に組み込まれたケースと除外されたケース、および各ピア・グループのケースが表示されます。

例外ケースの特定: 保存

「保存」ダイアログは、変数およびモデルの保存オプションを提供します。

変数の保存
: このグループのコントロールにより、モデル変数をアクティブ・データ・セットに保存できます。また、保存する変数と名前が競合する既存の変数を置き換えることもできます。

  異常指数
  : 各ケースの異常値指標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、ピア・グループの ID、ケース度数、および割合 (パーセント) で表されたサイズを、指定されたルート名の変数に保存します。例えば、ルート名として *Peer* が指定されると、変数 *Peerid*、*PeerSize*、および*PeerPctSize* が生成されます。*Peerid* はケースのピア・グループ ID、*PeerSize* はそのグループのサイズ、および *PeerPctSize* はグループのサイズ (パーセント) です。

  理由
  : 理由変数のセットを指定されたルート名で保存します。理由変数のセットは、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。セット数は、「オプション」タブで要求された理由の数に応じて変わります。例えばルート名 *Reason* が指定されている場合に生成される変数は *ReasonVar_k*、*ReasonMeasure_k*、*ReasonValue_k*、および *ReasonNorm_k* です (*k* は 理由の順序 (*k* 番目) です)。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。

  名前またはルート名が同一の既存の変数を置換する
  : 選択すると、保存する変数と名前が競合する既存の変数が置き換えられます。

モデル ファイルをエクスポート
: モデルを外部 XML ファイルに保存できます。

例外ケースの特定: 欠損値

「欠損値」ダイアログでは、ユーザ欠損値とシステム欠損値の処理方法を制御します。

分析から欠損値を除外する
: 欠損値を持つケースが分析から除外されます。

分析に欠損値を含める
: 連続型変数の欠損値は対応する全平均に置換され、カテゴリー変数の欠損カテゴリーはグループ化され、有効なカテゴリーとして扱われます。その後、処理された変数が分析で使用されます。必要であれば、ケースごとの欠損変数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

「オプション」ダイアログには、例外ケースの基準および同位グループの数の範囲を定義するための設定があります。

例外ケースを特定する基準
: 以下の設定により、異常値リストに含めるケースの数が決まります。

  異常指数の最も高い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

分析に使用されるアクティブなデータ・セット内のケースの総数以下の正整数を指定します。

異常指数値が最小値以上のケースのみを特定する

負ではない数値を指定します。ケースの異常値指標の値が指定されたカットオフ点以上の場合、そのケースは異常と見なされます。このオプションを使用する場合は、「ケースのパーセント」と「ケースの固定数」オプションを指定してください。例えば、ケースの固定数として 50 を指定し、カットオフ値として 2 を指定した場合、異常値リストには最大で 50 個のケースが含まれ、各ケースは 2 以上の異常値指標値を持ちます。

同位グループの数

このプロシージャは、指定された最小値から最大値までの範囲内で最適な数の同位グループを検索します。これらの値は正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場合、プロシージャーは固定数のピア・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、プロシージャーが作成するピア・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、この理由の変数名、変数の値、および対応するピア・グループの値で構成されます。負ではない整数を指定してください。この値が、分析で使用される処理済み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

このコマンド・シンタックス言語により、以下の操作が可能です。

- すべての分析変数を明示的に指定せずに、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)。
- 連続型変数とカテゴリー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)。

シンタックスの詳細については、「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。

特記事項

本書は米国 IBM が提供する製品およびサービスについて作成したものです。この資料は、IBM から他の言語でも提供されている可能性があります。ただし、これを入手するには、本製品または当該言語版製品を所有している必要がある場合があります。

本書に記載の製品、サービス、または機能が日本においては提供されていない場合があります。日本で利用可能な製品、サービス、および機能については、日本 IBM の営業担当員にお尋ねください。本書で IBM 製品、プログラム、またはサービスに言及していても、その IBM 製品、プログラム、またはサービスのみが使用可能であることを意味するものではありません。これらに代えて、IBM の知的所有権を侵害することのない、機能的に同等の製品、プログラム、またはサービスを使用することができます。ただし、IBM 以外の製品とプログラムの操作またはサービスの評価および検証は、お客様の責任で行っていただきます。

IBM は、本書に記載されている内容に関して特許権 (特許出願中のものを含む) を保有している場合があります。本書の提供は、お客様にこれらの特許権について実施権を許諾することを意味するものではありません。実施権についてのお問い合わせは、書面にて下記宛先にお送りください。

〒*103-8510*
東京都中央区日本橋箱崎町*19*番*21*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

IBM およびその直接または間接の子会社は、本書を特定物として現存するままの状態で提供し、商品性の保証、特定目的適合性の保証および法律上の瑕疵担保責任を含むすべての明示もしくは黙示の保証責任を負わないものとします。国または地域によっては、法律の強行規定により、保証責任の制限が禁じられる場合、強行規定の制限を受けるものとします。

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。本書は定期的に見直され、必要な変更は本書の次版に組み込まれます。IBM は予告なしに、随時、この文書に記載されている製品またはプログラムに対して、改良または変更を行うことがあります。

本書において IBM 以外の Web サイトに言及している場合がありますが、便宜のため記載しただけであり、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。それらの Web サイトにある資料は、この IBM 製品の資料の一部ではありません。それらの Web サイトは、お客様の責任でご使用ください。

IBM は、お客様が提供するいかなる情報も、お客様に対してなんら義務も負うことのない、自ら適切と信ずる方法で、使用もしく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保持者で、(i) 独自に作成したプログラムとその他のプログラム (本プログラムを含む) との間での情報交換、および (ii) 交換された情報の相互利用を可能にすることを目的として、本プログラムに関する情報を必要とする方は、下記に連絡してください。

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

本プログラムに関する上記の情報は、適切な使用条件の下で使用することができますが、有償の場合もあります。

本書で説明されているライセンス・プログラムまたはその他のライセンス資料は、IBM 所定のプログラム契約の契約条項、IBM プログラムのご使用条件、またはそれと同等の条項に基づいて、IBM より提供されます。

記載されている性能データとお客様事例は、例として示す目的でのみ提供されています。実際の結果は特定の構成や稼働条件によって異なります。

IBM 以外の製品に関する情報は、その製品の供給者、出版物、もしくはその他の公に利用可能なソースから入手したものです。IBM は、それらの製品のテストは行っておりません。したがって、他社製品に関する実行性、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者にお願いします。

IBM の将来の方向または意向に関する記述については、予告なしに変更または撤回される場合があり、単に目標を示しているものです。

本書には、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。より具体性を与えるために、それらの例には、個人、企業、ブランド、あるいは製品などの名前が含まれている場合があります。これらの名前はすべて架空のものであり、名前や住所が類似する個人や企業が実在しているとしても、それは偶然にすぎません。

著作権使用許諾:

本書には、様々なオペレーティング・プラットフォームでのプログラミング手法を例示するサンプル・アプリケーション・プログラムがソース言語で掲載されています。お客様は、サンプル・プログラムが書かれているオペレーティング・プラットフォームのアプリケーション・プログラミング・インターフェースに準拠したアプリケーション・プログラムの開発、使用、販売、配布を目的として、いかなる形式においても、IBM に対価を支払うことなくこれを複製し、改変し、配布することができます。このサンプル・プログラムは、あらゆる条件下における完全なテストを経ていません。従って IBM は、これらのサンプル・プログラムについて信頼性、利便性もしくは機能性があることをほのめかしたり、保証することはできません。これらのサンプル・プログラムは特定物として現存するままの状態で提供されるものであり、いかなる保証も提供されません。 IBM は、お客様の当該サンプル・プログラムの使用から生ずるいかなる損害に対しても一切の責任を負いません。

それぞれの複製物、サンプル・プログラムのいかなる部分、またはすべての派生的創作物にも、次のように、著作権表示を入れていただく必要があります。

商標

IBM、IBM ロゴおよび ibm.com は、世界の多くの国で登録された International Business Machines Corporation の商標です。他の製品名およびサービス名等は、それぞれ IBM または各社の商標である場合があります。現時点での IBM の商標リストについては、http://www.ibm.com/legal/copytrade.shtml をご覧ください。

Adobe、Adobe ロゴ、PostScript、PostScript ロゴは、Adobe Systems Incorporated の米国およびその他の国における登録商標または商標です。

インテル、Intel、Intel ロゴ、Intel Inside、Intel Inside ロゴ、Centrino、Intel Centrino ロゴ、Celeron、Xeon、Intel SpeedStep、Itanium、および Pentium は、Intel Corporation または子会社の米国およびその他の国における商標または登録商標です。

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

Microsoft、Windows、Windows NT および Windows ロゴは、Microsoft Corporation の米国およびその他の国における商標です。

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

Java およびすべての Java 関連の商標およびロゴは Oracle やその関連会社の米国およびその他の国における商標または登録商標です。

索引

日本語, 数字, 英字, 特殊文字の順に配列されています。なお, 濁音と半濁音は清音と同等に扱われています。

［ア行］

［カ行］

［ハ行］

［ラ行］

Printed in Japan

日本アイ・ビー・エム株式会社
〒103-8510 東京都中央区日本橋箱崎町19-21