バージョン13

予測モデルおよび発展的なモデル
第2版

「真の発見の旅とは、新しい風景を探すことではなく、新たな視点を持つことである。」
マルセル・プルースト

JMP, A Business Unit of SAS
SAS Campus Drive
Cary, NC 27513

13.1

このマニュアルを引用する場合は、次の正式表記を使用してください: SAS Institute Inc. 2017.
『JMP® 13 予測モデルおよび発展的なモデル 第 2 版』 Cary, NC: SAS Institute Inc.

JMP® 13 予測モデルおよび発展的なモデル 第 2 版

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513-2414.

2017 年 2 月

テクノロジーライセンスに関する通知

- Scintilla - Copyright © 1998-2014 by Neil Hodgson <neilh@scintilla.org>.

  All Rights Reserved.

  何らかの目的でこのソフトウェアとそのマニュアルを手数料なしで使用、コピー、変更および配布することは、これをもって許可されます。ただし、すべてのコピーに上記の著作権に関する通知が記載されていること、および補助的なマニュアルに著作権に関する通知とこの許可に関する通知の両方が記載されていることを条件とします。

  NEIL HODGSONは、商業性および適合性の黙示的な保証を含め、このソフトウェアに関するすべての保証を放棄します。NEIL HODGSONは、いかなる場合においても、それが契約、過失、もしくは他の不法行為のどれであれ、このソフトウェアの使用もしくは性能から生じた、もしくはそれに関連して生じた使用、データ、もしくは利益の損失の結果として生じる特別損害、間接損害、もしくは付随的損害を始めとするいかなる損害に対しても責任を負いません。

- Telerik RadControls: Copyright © 2002-2012, Telerik. 含まれているTelerik RadControlsをJMP以外で使用することは許可されていません。
- ZLIB圧縮ライブラリ - Copyright © 1995-2005, Jean-Loup Gailly and Mark Adler.
- Natural Earthを使用して作成。無料のベクトルおよびラスター地図データ @ naturalearthdata.com.
- パッケージ - Copyright © 2009-2010, Stéphane Sudre (s.sudre.free.fr). All rights reserved.

  ソースおよびバイナリの形で、そのまま、もしくは変更を加えて再配布および使用することは、次のような条件を満たす限り、許可されます。

  再配布するソースコードには、上記の著作権に関する通知、この条件リスト、これに続く放棄声明が記載されていなければなりません。

  バイナリ形式で再配布する場合は、共に提供されるマニュアルなどの資料に上記の著作権に関する通知、この条件リスト、これに続く放棄声明が記載されていなければなりません。

  事前に書面による許可を得ることなく、このソフトウェアから派生した製品の推奨または宣伝のためにWhiteBoxの名前やその貢献者の名前を使用することはできません。

  このソフトウェアは、著作権保有者および貢献者によって「現状のままで」提供され、商業性および特定の目的に対する適合性に関する黙示的な保証を含むがそれに限らない、いかなる明示的もしくは黙示的な保証も行われません。いかなる場合においても、著作権保有者または貢献者は、損害の原因が何であれ、そして法的責任の根拠が何であれ、つまり、契約、厳格責任、不法行為（過失その他を含む）のどれであれ、かかる損害の発生する可能性を事前に知らされていたとしても、このソフトウェアをどのように使用して生じた損害であれ、いかなる直接損害、間接損害、付随的損害、特別損害、懲罰的損害、もしくは結果損害（代替品または代替サービスの調達、使用機会、データもしくは利益の損失、業務の中断を含むがそれに限らない）に対しても責任を負いません。

- iODBC ソフトウェア - Copyright © 1995-2006, OpenLink Software Inc and Ke Jin (www.iodbc.org). All rights reserved.

  ソースおよびバイナリの形で、そのまま、もしくは変更を加えて再配布および使用することは、次のような条件を満たす限り、許可されます。

  – 再配布するソースコードには、上記の著作権に関する通知、この条件リスト、これに続く放棄声明が記載されていなければなりません。

  – バイナリ形式で再配布する場合は、共に提供されるマニュアルなどの資料に上記の著作権に関する通知、この条件リスト、これに続く放棄声明が記載されていなければなりません。

  – 事前に書面による許可を得ることなく、このソフトウェアから派生した製品の推奨または宣伝のために OpenLink Software Inc. の名前やその貢献者の名前を使用することはできません。

  このソフトウェアは、著作権保有者および貢献者によって「現状のままで」提供され、商業性および特定の目的に対する適合性に関する黙示的な保証を含むがそれに限らない、いかなる明示的もしくは黙示的な保証も行われません。いかなる場合においても、OPENLINK または貢献者は、損害の原因が何であれ、そして法的責任の根拠が何であれ、つまり、契約、厳格責任、不法行為（過失その他を含む）のどれであれ、かかる損害の発生する可能性を事前に知らされていたとしても、このソフトウェアをどのように使用して生じた損害であれ、いかなる直接損害、間接損害、付随的損害、特別損害、懲罰的損害、もしくは結果損害（代替品または代替サービスの調達、使用機会、データもしくは利益の損失、業務の中断を含むがそれに限らない）に対しても責任を負いません。

- bzip2、関連ライブラリの「libbzip2」、およびすべてのマニュアル: Copyright © 1996-2010, Julian R Seward. All rights reserved.

  ソースおよびバイナリの形で、そのまま、もしくは変更を加えて再配布および使用することは、次のような条件を満たす限り、許可されます。

  再配布するソースコードには、上記の著作権に関する通知、この条件リスト、これに続く放棄声明が記載されていなければなりません。

  このソフトウェアの供給源は正しく表記しなければならず、使用者が元のソフトウェアを記述したと主張することはできません。ある製品の中でこのソフトウェアを使用する場合は、その製品のマニュアルに謝辞を記載してもらえるとありがたいですが、必須ではありません。

  ソースに変更を加えたバージョンには、その旨を明記しなければならず、元のソフトウェアとは違うものであることを明確にしてください。

  事前に書面による許可を得ることなく、このソフトウェアから派生した製品の推奨または宣伝のために作成者の名前を使用することはできません。

  このソフトウェアは、作成者によって「現状のままで」提供され、商業性および特定の目的に対する適合性に関する黙示的な保証を含むがそれに限らない、いかなる明示的もしくは黙示的な保証も行われません。いかなる場合においても、作成者は、損害の原因が何であれ、そして法的責任の根拠が何であれ、つまり、契約、厳格責任、不法行為（過失その他を含む）のどれであれ、かかる損害の発生する可

能性を事前に知らされていたとしても、このソフトウェアをどのように使用して生じた損害であれ、いかなる直接損害、間接損害、付随的損害、特別損害、懲罰的損害、もしくは結果損害（代替品または代替サービスの調達、使用機会、データもしくは利益の損失、業務の中断を含むがそれに限らない）に対しても責任を負いません。

- R ソフトウェア : Copyright © 1999-2012, R Foundation for Statistical Computing.
- MATLAB ソフトウェア : Copyright © 1984-2012, The MathWorks, Inc. 米国特許法および国際特許法によって保護されています。www.mathworks.com/patents を参照してください。MATLAB および Simulink は、The MathWorks, Inc. の登録商標です。他の商標は、www.mathworks.com/trademarks に一覧されています。他の製品名やブランド名は、それぞれの所有者の商標または登録商標である可能性があります。
- libopc: Copyright © 2011, Florian Reuter. All rights reserved.

  ソースおよびバイナリの形で、そのまま、もしくは変更を加えて再配布および使用することは、次のような条件を満たす限り、許可されます。

  – 再配布するソースコードには、上記の著作権に関する通知、この条件リスト、これに続く放棄声明が記載されていなければなりません。

  – バイナリ形式で再配布する場合は、共に提供されるマニュアルなどの資料に上記の著作権に関する通知、この条件リスト、これに続く放棄声明が記載されていなければなりません。

  – 事前に書面による許可を得ることなく、このソフトウェアから派生した製品の推奨または宣伝のために Florian Reuter の名前やその貢献者の名前を使用することはできません。

  このソフトウェアは、著作権保有者および貢献者によって「現状のままで」提供され、商業性および特定の目的に対する適合性に関する黙示的な保証を含むがそれに限らない、いかなる明示的もしくは黙示的な保証も行われません。いかなる場合においても、著作権保有者または貢献者は、損害の原因が何であれ、そして法的責任の根拠が何であれ、つまり、契約、厳格責任、不法行為（過失その他を含む）のどれであれ、かかる損害の発生する可能性を事前に知らされていたとしても、このソフトウェアをどのように使用して生じた損害であれ、いかなる直接損害、間接損害、付随的損害、特別損害、懲罰的損害、もしくは結果損害（代替品または代替サービスの調達、使用機会、データもしくは利益の損失、業務の中断を含むがそれに限らない）に対しても責任を負いません。

- libxml2 - ソースコードに特に記載がある場合を除く（たとえば、使用しているライセンスは類似しているが、著作権の通知が異なる hash.c、list.c ファイルや trio ファイル）、すべてのファイル :

  Copyright © 1998 - 2003 Daniel Veillard. All Rights Reserved.

  これをもって、このソフトウェアのコピーと関連する文書ファイル（「本ソフトウェア」）を入手した人すべてに対し、無料で本ソフトウェアを使用、コピー、変更、マージ、パブリッシュ、配布、サブライセンスする、もしくはコピーを販売する権利を含むがそれに限定せず、本ソフトウェアを制限なく取り扱う権利、および本ソフトウェアの供給相手に対してそうすることを許可する権利が付与されます。ただし、以下の条件を満たさなければなりません。

上記の著作権に関する通知とこの許可に関する通知が、本ソフトウェアのコピーのすべてまたは大部分に記載されていること。

このソフトウェアは、 現状のままで」提供され、商業性および特定の目的に対する適合性、および非侵害の保証を含むがそれに限らない、いかなる明示的もしくは黙示的な保証も行われません。DANIEL VEILLARDは、いかなる場合においても、それが契約、過失、もしくは他の不法行為のどれであれ、本ソフトウェアから、もしくは本ソフトウェアに関連して、または本ソフトウェアの使用もしくは他の取り扱いに関連して生じた申し立て、損害賠償もしくは他の義務に対し、責任を負いません。

この通知に含まれているものを除き、Daniel Veillardから事前に書面による許可を得ることなく、本ソフトウェアの広告、またはその他の手段による本ソフトウェアの販売、使用もしくは他の取り扱いの宣伝にDaniel Veillardの名前を使用することはできません。

- UNIXファイルに使用された解凍アルゴリズムについて:

  Copyright © 1985, 1986, 1992, 1993

  カリフォルニア大学評議員。All rights reserved.

  このソフトウェアは、評議員および貢献者によって「現状のままで」提供され、商業性および特定の目的に対する適合性に関する黙示的な保証を含むがそれに限らない、いかなる明示的もしくは黙示的な保証も行われません。いかなる場合においても、評議員または貢献者は、損害の原因が何であれ、そして法的責任の根拠が何であれ、つまり、契約、厳格責任、不法行為（過失その他を含む）のどれであれ、かかる損害の発生する可能性を事前に知らされていたとしても、このソフトウェアをどのように使用して生じた損害であれ、いかなる直接損害、間接損害、付随的損害、特別損害、懲罰的損害、もしくは結果損害（代替品または代替サービスの調達、使用機会、データもしくは利益の損失、業務の中断を含むがそれに限らない）に対しても責任を負いません。

  1. 再配布するソースコードには、上記の著作権に関する通知、この条件リスト、これに続く放棄声明が記載されていなければなりません。

  2. バイナリ形式で再配布する場合は、共に提供されるマニュアルなどの資料に上記の著作権に関する通知、この条件リスト、これに続く放棄声明が記載されていなければなりません。

  3. 事前に書面による許可を得ることなく、このソフトウェアから派生した製品の推奨または宣伝のために大学の名前や貢献者の名前を使用することはできません。

- Snowball - Copyright © 2001, Dr Martin Porter, Copyright © 2002, Richard Boulton.

  All rights reserved.

  ソースおよびバイナリの形で、そのまま、もしくは変更を加えて再配布および使用することは、次のような条件を満たす限り、許可されます。

  1. 再配布するソースコードには、上記の著作権に関する通知、この条件リスト、これに続く放棄声明が記載されていなければなりません。

  2. バイナリ形式で再配布する場合は、共に提供されるマニュアルなどの資料に上記の著作権に関する通知、この条件リスト、これに続く放棄声明が記載されていなければなりません。

3. 事前に書面による許可を得ることなく、このソフトウェアから派生した製品の推奨または宣伝のために著作権保有者の名前や貢献者の名前を使用することはできません。

このソフトウェアは、著作権保有者および貢献者によって「現状のままで」提供され、商業性および特定の目的に対する適合性に関する黙示的な保証を含むがそれに限らない、いかなる明示的もしくは黙示的な保証も行われません。いかなる場合においても、著作権保有者または貢献者は、損害の原因が何であれ、そして法的責任の根拠が何であれ、つまり、契約、厳格責任、不法行為（過失その他を含む）のどれであれ、かかる損害の発生する可能性を事前に知らされていたとしても、このソフトウェアをどのように使用して生じた損害であれ、いかなる直接損害、間接損害、付随的損害、特別損害、懲罰的損害、もしくは結果損害（代替品または代替サービスの調達、使用機会、データもしくは利益の損失、業務の中断を含むがそれに限らない）に対しても責任を負いません。

目次
予測モデルおよび発展的なモデル

15 時系列分析

16 対応のあるペア分析

第 1 章

JMP の概要
マニュアルとその他のリソース

この章には以下の情報が記載されています。

- 本書の表記法
- JMP のマニュアル
- JMP ヘルプ
- その他のリソース
  - その他の JMP のドキュメンテーション
  - チュートリアル
  - 索引
  - Web リソース
  - テクニカルサポートのオプション

表記規則

マニュアルの内容と画面に表示される情報を対応付けるために、次のような表記規則を使っています。

- サンプルデータ名、列名、パス名、ファイル名、ファイル拡張子、およびフォルダ名は「」で囲んで表記しています。
- スクリプトのコードは`Lucida Sans Typewriter`フォントで表記しています。
- スクリプトコードの結果（ログに表示されるもの）は*`Lucida Sans Typewriter`*（斜体）フォントで表記し、先に示すコードよりインデントされています。
- クリックまたは選択する項目は [] で囲んで太字で表記しています。これには以下の項目があります。
  - ボタン
  - チェックボックス
  - コマンド
  - 選択可能なリスト項目
  - メニュー
  - オプション
  - タブ名
  - テキストボックス
- 次の項目の表記規則は下記のとおりです。
  - 重要な単語や句、JMPに固有の定義を持つ単語や句は太字または「」で囲んで表記
  - マニュアルのタイトルは『』で囲んで表記
  - 変数名は斜体で表記
  - スクリプトの出力は斜体で表記
- JMP Proのみの機能にはJMP Proアイコン JMP PRO がついています。JMP Proの機能の概要については https://www.jmp.com/ja_jp/software/predictive-analytics-software.html をご覧ください。

メモ: 特別な情報および制限事項には、この文のように「メモ」という見出しがついています。

ヒント: 役に立つ情報には「ヒント」という見出しがついています。

JMPのマニュアル

JMP では、PDF 形式のマニュアルが用意されています。

- PDF版は［**ヘルプ**］>［**ドキュメンテーション**］メニューまたはJMPオンラインヘルプのフッタから開くことができます。
- 検索しやすいようにすべてのドキュメンテーションが1つのPDFファイルにまとめられた『JMP ドキュメンテーションライブラリ』と呼ばれるファイルがあります。『JMP ドキュメンテーションライブラリ』のPDFファイルは［**ヘルプ**］>［**ドキュメンテーション**］メニューから開くことができます。

JMP ドキュメンテーションライブラリ

以下の表は、JMP ライブラリに含まれている各ドキュメンテーションの目的および内容をまとめたものです。

マニュアル	目的	内容
『はじめてのJMP』	JMPをあまりご存知ない方を対象とした入門ガイド	JMPの紹介と、データを作成および分析し始めるための情報
『JMPの使用法』	JMPのデータテーブルと、基本操作を理解する	一般的なJMPの概念と、データの読み込み、列プロパティの変更、データの並べ替え、SASへの接続など、JMP全体にわたる機能の説明
『基本的な統計分析』	このマニュアルを見ながら、基本的な分析を行う	［分析］メニューからアクセスできる以下のプラットフォームの説明： • 一変量の分布 • 二変量の関係 • 表の作成 • テキストエクスプローラ ［分析］>［二変量の関係］で二変量、一元配置分散分析、分割表に対する分析を実行する方法の説明。ブートストラップを使用した標本分布の近似方法やシミュレーションの機能を使用したパラメトリックな標本再抽出の実行方法の説明も含まれています。

マニュアル	目的	内容
『グラフ機能』	データに合った理想的なグラフを見つける	［グラフ］メニューからアクセスできる以下のプラットフォームの説明： • グラフビルダー • 重ね合わせプロット • 三次元散布図 • 等高線図 • バブルプロット • パラレルプロット • セルプロット • ツリーマップ • 散布図行列 • 三角図 • チャート このマニュアルには背景マップやカスタムマップの作成方法も記載されています。
『プロファイル機能』	対話式のプロファイルツールの使い方を学ぶ。任意の応答曲面の断面を表示できるようになります。	［グラフ］メニューに表示されるすべてのプロファイルについて。誤差因子の分析が、ランダム入力を使用したシミュレーションの実行とともに含まれています。
『実験計画(DOE)』	実験の計画方法と適切な標本サイズの決定方法を学ぶ	［実験計画（DOE）］メニューと［分析］>［発展的なモデル］メニューの「発展的な実験計画モデル」に関するすべてのトピックについて。

マニュアル	目的	内容
『基本的な回帰モデル』	「モデルのあてはめ」プラットフォームとその多くの手法について学ぶ	［分析］メニューの「モデルのあてはめ」プラットフォームで使用できる、以下の手法の説明： • 標準最小2乗 • ステップワイズ • 一般化回帰 • 混合モデル • MANOVA • 対数線形-分散 • 名義ロジスティック • 順序ロジスティック • 一般化線形モデル

マニュアル	目的	内容
『予測モデルおよび発展的なモデル』	さらなるモデリング手法について学ぶ	［分析］>［予測モデル］メニューで使用できる以下のプラットフォームの説明： • モデル化ユーティリティ • ニューラル • パーティション • ブートストラップ森 • ブースティングツリー • K近傍法 • 単純Bayes • モデルの比較 • 計算式デポ ［分析］>［発展的なモデル］メニューで使用できる以下のプラットフォームの説明： • 曲線のあてはめ • 非線形回帰 • Gauss過程 • 時系列分析 • 対応のあるペア ［分析］>［スクリーニング］メニューで使用できる以下のプラットフォームの説明： • 応答のスクリーニング • 工程のスクリーニング • 説明変数のスクリーニング • アソシエーション分析 ［分析］>［発展的なモデル］>［発展的な実験計画モデル］で使用できるプラットフォームについては、『実験計画(DOE)』に説明があります。

マニュアル	目的	内容
『多変量分析』	複数の変数を同時に分析するための手法について理解を深める	［分析］>［多変量］メニューで使用できる以下のプラットフォームの説明： • 多変量の相関 • 主成分分析 • 判別分析 • PLS ［分析］>［クラスター分析］メニューで使用できる以下のプラットフォームの説明： • 階層型クラスター分析 • K Means クラスター分析 • 正規混合 • 潜在クラス分析 • 変数のクラスタリング
『品質と工程』	工程を評価し、向上させるためのツールについて理解を深める	［分析］>［品質と工程］メニューで使用できる以下のプラットフォームの説明： • 管理図ビルダーと個々の管理図 • 測定システム分析 • 計量値/計数値ゲージチャート • 工程能力 • パレート図 • 特性要因図

マニュアル	目的	内容
『信頼性/生存時間分析』	製品やシステムにおける信頼性を評価し、向上させる方法、および人や製品の生存時間データを分析する方法について学ぶ	［分析］>［信頼性/生存時間分析］メニューで使用できる以下のプラットフォームの説明： • 寿命の一変量 • 寿命の二変量 • 累積損傷 • 再生モデルによる分析 • 劣化分析と破壊劣化 • 信頼性予測 • 信頼性成長 • 信頼性ブロック図 • 修理可能システムのシミュレーション • 生存時間分析 • 生存時間(パラメトリック)のあてはめ • 比例ハザードのあてはめ
『消費者調査』	消費者選好を調査し、その洞察を使用してより良い製品やサービスを作成するための方法を学ぶ	［分析］>［消費者調査］メニューで使用できる以下のプラットフォームの説明： • カテゴリカル • 多重対応分析 • 多次元尺度構成 • 因子分析 • 選択モデル • MaxDiff • アップリフト • 項目分析
『スクリプトガイド』	パワフルなJMPスクリプト言語（JSL）の活用方法について学ぶ	スクリプトの作成やデバッグ、データテーブルの操作、ディスプレイボックスの構築、JMPアプリケーションの作成など。
『スクリプト構文リファレンス』	JSL関数、その引数、およびオブジェクトやディスプレイボックスに送信するメッセージについて理解を深める	JSLコマンドの構文、例、および注意書き。

メモ：［**ドキュメンテーション**］メニューでは、印刷可能な2つのリファレンスカードも用意されています。『**メニューカード**』はJMPのメニューをまとめた表で、『**クイックリファレンス**』はJMPのショートカットキーをまとめた表です。

JMPヘルプ

JMPヘルプは、一連のマニュアルの簡易版です。JMPのヘルプは、次のいくつかの方法で開くことができます。

- Windowsでは、F1キーを押すとヘルプシステムウィンドウが開きます。
- データテーブルまたはレポートウィンドウの特定の部分のヘルプを表示します。［**ツール**］メニューからヘルプツールを選択した後、データテーブルやレポートウィンドウの任意の位置でクリックすると、その部分に関するヘルプが表示されます。
- JMPウィンドウ内で［**ヘルプ**］ボタンをクリックします。
- Windowsの場合、［ヘルプ］メニューの［**ヘルプの目次**］、［**ヘルプの検索**］、［**ヘルプの索引**］の各オプションを使用して、JMPヘルプ内を検索し、目的の内容を表示します。Macの場合、［**ヘルプ**］>［**JMPヘルプ**］を選択します。

JMPを習得するためのその他のリソース

JMPのマニュアルとJMPヘルプの他、次のリソースもJMPの学習に役立ちます。

- チュートリアル（「チュートリアル」（27ページ）を参照）
- サンプルデータ（「サンプルデータテーブル」（28ページ）を参照）
- 索引（「統計用語とJSL用語の習得」（28ページ）を参照）
- 使い方ヒント（「JMPを使用するためのヒント」（28ページ）を参照）
- Webリソース（「JMP User Community」（29ページ）を参照）
- 専門誌『JMPer Cable』（「JMPer Cable」（29ページ）を参照）
- JMPに関する書籍（「JMP関連書籍」（29ページ）を参照）
- JMPスターター（「「JMPスターター」ウィンドウ」（30ページ）を参照）
- 教育用リソース（「サンプルデータテーブル」（28ページ）を参照）

チュートリアル

［**ヘルプ**］>［**チュートリアル**］を選択して、JMPのチュートリアルを表示できます。［**チュートリアル**］メニューの最初の項目は［**チュートリアルディレクトリ**］です。この項目を選択すると、すべてのチュートリアルをカテゴリ別に整理した新しいウィンドウが開きます。

JMPに慣れていない方は、まず［**初心者用チュートリアル**］を試してみてください。JMPのインターフェースおよび基本的な使用方法を学ぶことができます。

他のチュートリアルでは、実験の計画、標本平均と定数の比較など、JMPの具体的な活用法を学習できます。

サンプルデータテーブル

JMPのマニュアルで取り上げる例は、すべてサンプルデータを使用しています。サンプルデータディレクトリを開くには、［**ヘルプ**］>［**サンプルデータライブラリ**］を選択します。

サンプルデータテーブルを文字コード順に並べた一覧を表示する、またはカテゴリごとにサンプルデータを表示するには、［**ヘルプ**］>［**サンプルデータ**］を選択します。

サンプルデータテーブルは次のディレクトリにインストールされています。

Windowsの場合: C:¥Program Files¥SAS¥JMP¥13¥Samples¥Data

Macintoshの場合: \Library\Application Support\JMP\13\Samples\Data

JMP Proでは、サンプルデータが（JMPではなく）JMPPROディレクトリにインストールされています。シングルユーザーライセンス版のJMP（JMP シュリンクラップ）では、サンプルデータがJMPSWディレクトリにインストールされています。

サンプルデータの使用例を参照するには、［**ヘルプ**］>［**サンプルデータ**］を選択し、教育用セクションから検索してください。教育用リソースについては、http://jmp.com/tools にも情報があります。

統計用語とJSL用語の習得

［**ヘルプ**］メニューには、次の索引が用意されています。

統計の索引　統計用語が説明されています。

スクリプトの索引　JSL関数、オブジェクト、ディスプレイボックスに関する情報を検索できます。スクリプトの索引からサンプルスクリプトを編集して実行することもできます。

JMPを使用するためのヒント

JMPを最初に起動すると、「使い方ヒント」ウィンドウが表示されます。このウィンドウには、JMPを使う上でのヒントが表示されます。

「使い方ヒント」ウィンドウを表示しないようにするには、［**起動時にヒントを表示する**］のチェックを外します。再表示するには、［**ヘルプ**］>［**使い方ヒント**］を選択します。または、「環境設定」ウィンドウで非表示に設定することもできます。詳細については、『JMPの使用法』を参照してください。

ツールヒント

次のような項目の上にカーソルを置くと、その項目を説明するツールヒントが表示されます。

- メニューまたはツールバーのオプション
- グラフ内のラベル
- レポートウィンドウ内の結果（テキスト）（カーソルで円を描くと表示される）
- 「ホームウィンドウ」内のファイル名またはウィンドウ名
- スクリプトエディタ内のコード

ヒント： Windowsでは、JMP環境設定でツールヒントを表示しないよう設定できます。［**ファイル**］>［**環境設定**］>［**一般**］を選択し、［**メニューのヒントを表示**］の選択を解除します。このオプションは、Macintoshでは使用できません。

JMP User Community

JMP User Communityでは、さまざまな方法でJMPをさらに習得したり、他のSASユーザとのコミュニケーションを図ったりできます。ラーニングライブラリには1ページガイド、チュートリアル、デモなどが用意されており、JMPを使い始める上でとても便利です。また、JMPのさまざまなトレーニングコースに登録して、自己教育を進めることも可能です。

その他のリソースとして、ディスカッションフォーラム、サンプルデータやスクリプトファイルの交換、Webcastセミナー、ソーシャルネットワークグループなども利用できます。

WebサイトのJMPリソースにアクセスするには、［**ヘルプ**］>［**JMP User Community**］を選択するか、https://community.jmp.com/ をご覧ください。

JMPer Cable

JMPer Cableは、JMPユーザを対象とした年刊の専門誌です。JMPer Cableは次のJMP Webサイトで閲覧可能です。

http://www.jmp.com/about/newsletters/jmpercable/（英語）

JMP関連書籍

JMP関連書籍は、次のJMP Webページで紹介されています。

https://www.jmp.com/ja_jp/academic/books-for-jmp-users.html

「JMP スターター」ウィンドウ

JMP またはデータ分析にあまり慣れていないユーザは、「JMP スターター」ウィンドウから開始するとよいでしょう。カテゴリ分けされた項目には説明がついており、ボタンをクリックするだけで該当の機能を起動できます。「JMP スターター」ウィンドウには、［分析］、［グラフ］、［テーブル］、および［ファイル］メニュー内の多くのオプションがあります。また、JMP Pro の機能やプラットフォームのリストも含まれています。

- 「JMP スターター」ウィンドウを開くには、［**表示**］（Macintosh では［**ウィンドウ**］）**>**［**JMP スターター**］を選択します。
- Windows で JMP の起動時に自動的に「JMP スターター」を表示するには、［**ファイル**］**>**［**環境設定**］**>**［**一般**］を選び、「開始時の JMP ウィンドウ」リストから［**JMP スターター**］を選択します。Macintosh では、［**JMP**］**>**［**環境設定**］**>**［**起動時に JMP スターターウィンドウを表示する**］を選択します。

テクニカルサポート

JMP のテクニカルサポートは、JMP のエンジニアが担当し、その多くは、統計学などの技術的な分野の知識を有しています。

http://www.jmp.com/japan/support には、テクニカルサポートへの連絡方法などが記載されています。

第2章

予測モデルと発展的なモデルについて

モデル化手法の概要

この『予測モデルおよび発展的なモデル』では、「応答のスクリーニング」、「パーティション」、「ニューラル」などの高度な統計手法を説明します。

- モデル化に関するユーティリティとして、データクリーニングや前処理の機能が用意されています。各機能は、データを探索したり、理解を深めるのに役立ちます。第3章「モデル化ユーティリティ」を参照してください。
- 「ニューラル」プラットフォームは、入力層、出力層、および、1 ～ 2 層の隠れ層（中間層）をもつ多層パーセプトロンニューラルネットワークをあてはめます。ニューラルネットワークは、柔軟な関数によって、入力変数から1つまたは複数の応答変数を予測します。第4章「ニューラルネットワーク」を参照してください。
- 「パーティション」プラットフォームは、X と Y の関係に従ってデータを再帰的に分割し、ディシジョンツリー（決定木）を作成します。第5章「パーティション」を参照してください。
- JMP PRO「ブートストラップ森」プラットフォームは、学習データから何回もデータを無作為抽出し、その無作為抽出された各データにディシジョンツリー（決定木）をあてはめ、それらの結果を組み合わせて予測値を求めます。第6章「ブートストラップ森」を参照してください。
- JMP PRO「ブースティングツリー」プラットフォームは、小さなディシジョンツリーをいくつも積み重ねた加法モデルを作成します。各ツリーはそれぞれ少数（通常は5つ以下）の分岐で構成されます。各ツリーは残差に対して再帰的にあてはめられていきます。第7章「ブースティングツリー」を参照してください。
- JMP PRO「K近傍法」プラットフォームは、各データ行の予測値を、その近傍にあるデータから求めます。応答変数がカテゴリカルな場合は「分類」を行い、応答変数が連続尺度の場合には「予測」を行います。第8章「K近傍法」を参照してください。
- JMP PRO「単純 Bayes」プラットフォームは、応答変数がカテゴリカルなデータに対して、分類を行います。なお、説明変数（すなわち、因子）は、データマイニングの分野では「特徴（features）」とも呼ばれています。第9章「単純Bayes」を参照してください。
- JMP PRO「モデルの比較」プラットフォームでは、さまざまなモデルの予測能力を比較できます。各モデルの適合度や、診断プロットが表示されます。第10章「モデルの比較」を参照してください。
- JMP PRO「計算式デポ」プラットフォームは、モデルの整理や比較、プロファイル作成、スコア計算を行います。「計算式デポ」は、JMPデータテーブルとは切り離して候補となるモデルを保存できますので、モデルを比較するのに便利です。第11章「計算式デポ」を参照してください。
- 「曲線のあてはめ」プラットフォームには、多項式・ロジスティック曲線・Gompertz曲線・指数モデル・ピークモデル・薬物動態モデルなどの非線形モデルがあらかじめ用意されています。これらの非線形モデルをあてはめるのに、モデル式をユーザが設定する必要はありません。各種の分析機能やグラフ機能によって、グループ間の違いなどを検討できます。第12章「曲線のあてはめ」を参照してください。

- 「非線形回帰」プラットフォームでは、推定対象のモデル式やパラメータを指定し、ユーザ自身が定義した非線形モデルをあてはめることができます。第13章「非線形回帰」を参照してください。
- 「Gauss過程」プラットフォームは、複数の独立変数と、1つの応答変数との間の関係をモデル化します。独立変数と応答変数は、ともに連続尺度である必要があります。Gauss過程モデル（Gaussian process model）は、有限要素法のようなコンピュータによるシミュレーション実験などの分野で、データを完璧に補間するモデルとして広く利用されています。第14章「Gauss過程」を参照してください。
- 「時系列分析」プラットフォームでは、一変量の時系列データに対する分析や予測ができます。第15章「時系列分析」を参照してください。
- 「対応のあるペア」プラットフォームでは、相関がある2変数間の平均を比較し、その差を評価します。第16章「対応のあるペア分析」を参照してください。
- 「応答のスクリーニング」プラットフォームは、応答変数や説明変数が多数ある場合に、それらに対する検定の処理を一度に行います。検定結果や要約統計量は、データテーブルとしても出力されるため、それらの結果をさらに検討できます。第17章「応答のスクリーニング」を参照してください。
- 「工程のスクリーニング」プラットフォームは、経時的に測定された多数の工程データを探索的に調べます。このプラットフォームによって、管理図や安定性の指標や工程能力指数を求めたり、工程に生じた変化を検出したりできます。第18章「工程のスクリーニング」を参照してください。
- 「説明変数のスクリーニング」プラットフォームを使用すると、データセットから有意な説明変数をスクリーニングできます。第19章「説明変数のスクリーニング」を参照してください。
- JMP PRO「アソシエーション分析」プラットフォームは、一緒に登場する頻度が高いアイテム（項目）を探し出します。アソシエーション分析は、「マーケットバスケット分析」とも呼ばれています。同一のトランザクション内において頻繁に一緒に登場している商品を探し出すのによく用いられます。第20章「アソシエーション分析」を参照してください。

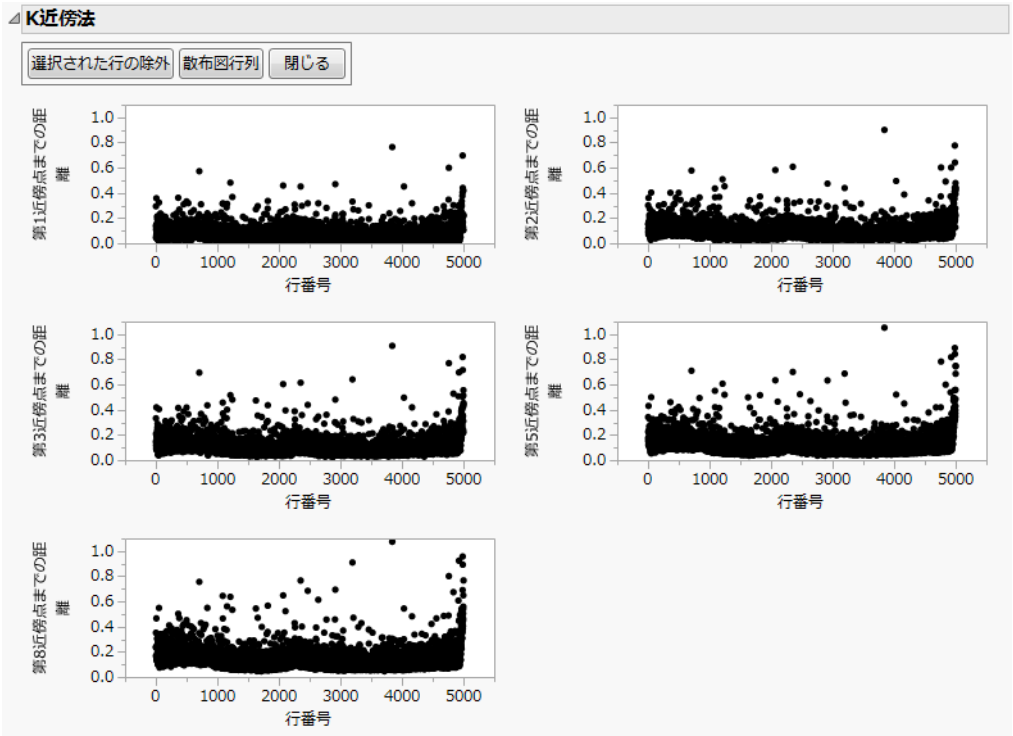
第3章

モデル化ユーティリティ
外れ値の除去、欠測値の補完、変数の選択

モデル化に関するユーティリティとして、データクリーニングや前処理の機能が用意されています。各機能は、データを探索したり、理解を深めるのに役立ちます。以下のような機能が用意されています。

- 一変量や多変量における外れ値を調べる。
- データの欠測値を探索し、補完する。
- JMP PRO 学習セット・検証セット・テストセットにデータ全体を分割する検証列を作成する。
- 説明変数が多いデータにおいて、強い効果をもつ説明変数を選び出す。

図3.1 多変量のk近傍法外れ値の例



「外れ値を調べる」ユーティリティ

データの外れ値を調べることは、重要です。外れ値が生じる理由としては、データの収集や記録における入力ミス、測定システムの不備、あるいは、「999」などの欠測値コードやエラーコードを生のデータ値として扱ってしまっている、などが考えられます。外れ値は推定に影響します。どのような統計分析でも、外れ値の方向にバイアス（偏り）がかかります。たとえば、外れ値があると、標本分散が過大に推定されてしまいます。しかし、場合によっては、外れ値を除去せずにそのまま残しておくべきかもしれません。なぜなら、外れ値を除去することにより、逆に、標本分散が過小に推定され、逆の方向にバイアスがかかるかもしれないからです。

外れ値を削除するにしても保持するにしても、まずそれらを見つけなければなりません。外れ値を視覚的に見つける方法はいくつかあります。たとえば、箱ひげ図、ヒストグラム、および散布図を描くことによって、極端な値を見つけられることはよくあります。詳細については、『はじめてのJMP』の「データの視覚化」章を参照してください。

「外れ値を調べる」ユーティリティには、一変量や多変量における外れ値を調べるための4つのオプションがあります。

分位点範囲の外れ値　一変量の分位点に基づいて、極端な値としての外れ値を識別します。このツールは、データ中の欠測値コードやエラーコードを見つけるのに便利です。まずこのツールから、外れ値の探索を始めると良いでしょう。「分位点範囲の外れ値」（38ページ）を参照してください。

ロバスト推定による外れ値　各列の中心と散らばりをロバストに推定し、それらの推定値に基づき、遠く離れているデータ値を外れ値として識別します。「ロバスト推定による外れ値」（41ページ）を参照してください。

多変量ロバスト推定による外れ値　「多変量の相関」プラットフォームで「ロバスト」オプションを指定したときに計算される平均と共分散行列から、Mahalanobisの距離を求め、外れ値を識別します。「多変量ロバスト推定による外れ値」（42ページ）を参照してください。

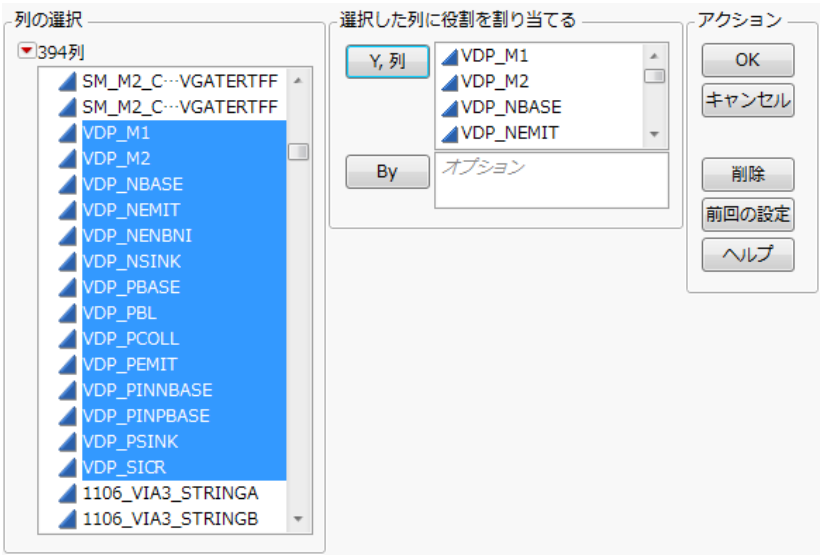
多変量のk近傍法外れ値　k番目の近傍点からの距離が遠いものを、外れ値として識別します。「多変量のk近傍法外れ値」（43ページ）を参照してください。

「外れ値を調べる」ユーティリティの例

「Probe.jmp」サンプルデータテーブルには、5800個の半導体ウエハについて測定された、387個の特徴（「**Responses**」列グループ）が含まれています。「**ロットID**」列と「**ウエハ番号**」列は、一意にウエハを識別します。ここでの目的は、データのいくつかの列を調べて、外れ値を識別することです。「外れ値を調べる」ユーティリティを使って外れ値を識別し、その後「一変量の分布」プラットフォームを使って分析してみましょう。

1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Probe.jmp」を開きます。
2. ［**分析**］>［**スクリーニング**］>［**外れ値を調べる**］を選択します。
3. 「**VDP_M1**」から「**VDP_SICR**」までの列を選択し、［**Y, 列**］をクリックします。これで、14個の列が選択されています（図3.2を参照）。

図3.2「外れ値を調べる」起動ウィンドウ



4. ［**OK**］をクリックします。
5. ［**分位点範囲の外れ値**］をクリックします。

   「分位点範囲の外れ値」レポートには、各列と、それぞれの外れ値の個数と値がリストされています。
6. 「分位点範囲の外れ値」レポートで［**外れ値のある列のみ表示**］チェックボックスをオンにします。これで、列のリストが外れ値のある列だけに制限されます。

   いくつかの列の外れ値の値が9999であることに注目してください。9は多くの業界で欠測値のコードとして使用されています。
7. 「「9」を含むデータ」レポートで、各列を選択します。
8. ［**「欠測値のコード」に最大「9」を追加**］をクリックします。

   元のデータを残したい場合には、［**名前を付けて保存**］コマンドを使って新しいファイル名で保存するように促す警告ダイアログが表示されます。
9. ［**OK**］をクリックします。
10. 「分位点範囲の外れ値」レポートで、［**再スキャン**］をクリックします。
11. ［**検索を整数に限定**］チェックボックスをオンにします。

    連続尺度のデータでは、エラーコードや欠測値コードなどを整数の値で表すことがよくあります。今、分析している列には、「9999」以外のエラーコードは含まれていないようです。
12. ［**検索を整数に限定**］の選択を解除します。

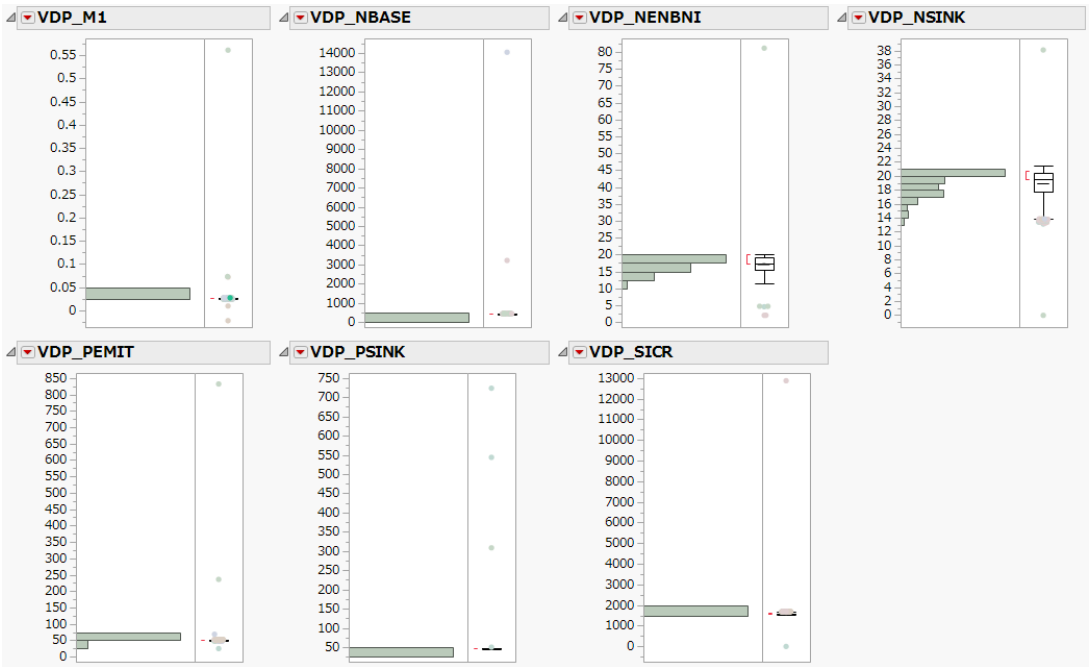
データを調べる

1. 「分位点範囲の外れ値」レポートで、残りのすべての列を選択します。

2. ［**行の選択**］をクリックします。
3. ［**分析**］>［**一変量の分布**］を選びます。
4. 先ほど選択した列を［**Y, 列**］の役割に割り当てます。「分位点範囲の外れ値」レポートでこれらの列を選択していたので、「一変量の分布」起動ウィンドウでもこれらの列がすでに選択されています。
5. ［**OK**］をクリックします。

   図3.3に、作成されるレポートの一部を示します。

図3.3 外れ値が選択された列の分布



「VDP_M1」列と**「VDP_PEMIT」**列では、選択された外れ値のいくつかは大半のデータに幾分近いものとなっています。その他の列では、選択された外れ値は十分に離れており、分析から除外しても良さそうです。

除外した外れ値の精査

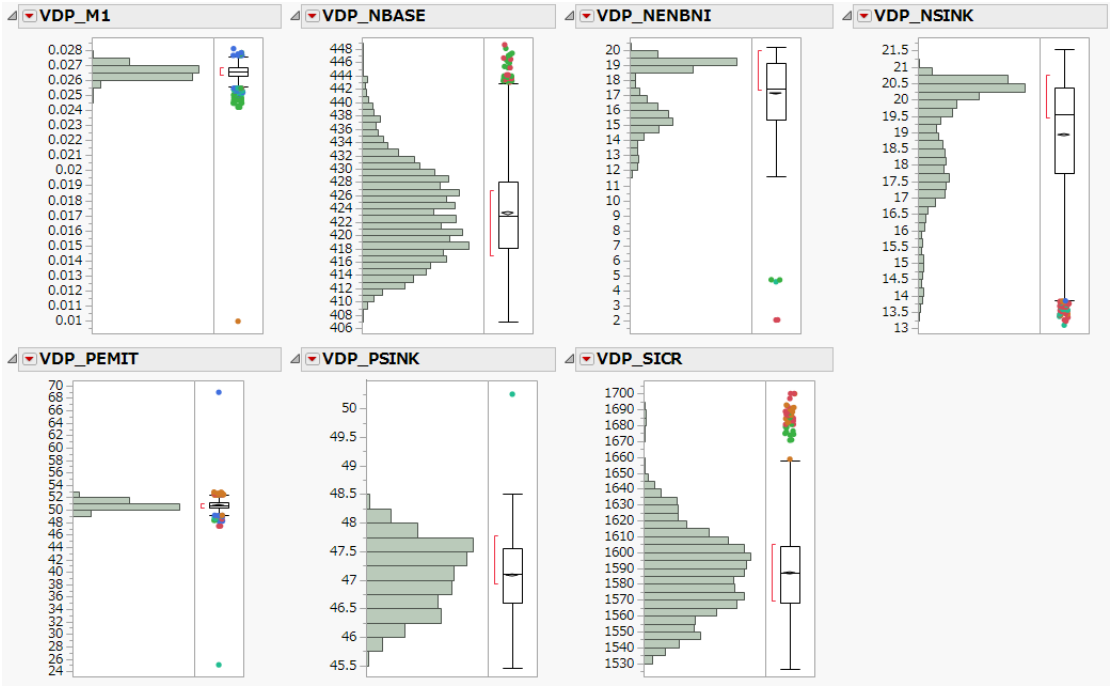
1. 「分位点範囲の外れ値」レポートで、Ctrl キーを押しながら**「VDP_M1」**と**「VDP_PEMIT」**の選択を解除します。
2. 残りの列は選択されたままにして、［**行の除外**］をクリックします。
3. 「Q」を20に変更します。
4. ［**再スキャン**］をクリックします。
5. レポートで**「VDP_M1」**と**「VDP_PEMIT」**を選択します。［**行の選択**］をクリックします。

データを再度調べる

1. 「一変量の分布」レポートを再度調べます。先ほどの操作で選択された外れ値は、大半のデータから離れているので、分析から除外しても良さそうです。
2. 「分位点範囲の外れ値」レポートで、［**行の除外**］をクリックします。
3. 「一変量の分布」レポートの赤い三角ボタンのメニューをクリックします。
4. ［**やり直し**］>［**分析のやり直し**］を選択します。

   図3.4に、作成されるレポートの一部を示します。

図3.4 外れ値が除外された列の分布



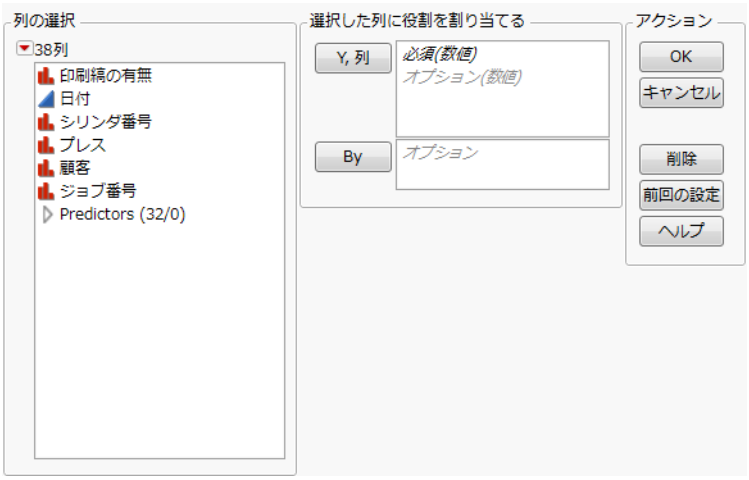
外れ値を除外したデータの分布は、より情報を把握しやすいものとなります。

「外れ値を調べる」ユーティリティの起動

メモ：「外れ値を調べる」のコマンドで分析できるのは連続尺度の列のみです。連続尺度以外の列も起動ウィンドウに指定できますが、無視されます。

「外れ値を調べる」を起動するには、［**分析**］>［**スクリーニング**］>［**外れ値を調べる**］を選択します。起動ウィンドウが表示されます。

図3.5「外れ値を調べる」起動ウィンドウ



起動ウィンドウで、分析列を選択し、[**Y, 列**] に指定します。また、**By** 変数を指定することもできます。[**OK**] をクリックすると、「外れ値を調べる」レポートが開きます。ここに、次の4つの外れ値分析コマンドが提示されます。

- 「分位点範囲の外れ値」(38ページ)
- 「ロバスト推定による外れ値」(41ページ)
- 「多変量ロバスト推定による外れ値」(42ページ)
- 「多変量のk近傍法外れ値」(43ページ)

分位点範囲の外れ値

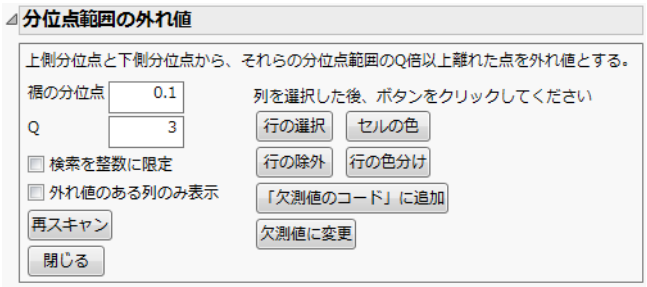
「分位点範囲の外れ値」ユーティリティでは、各列の分位点によって極端な値を識別します。分位点の計算には特定の確率分布を仮定する必要がないので、外れ値を検出するのに分位点を用いるのは便利です。分位点を求めるには、まず、データを小さい値から大きい値へと並べます。たとえば、20%の分位点は、データの20%がそれより小さい値です。外れ値は、分位点の範囲（2つの分位点の差）の定数倍によって定義されます。分位点の計算方法の詳細については、『基本的な統計分析』の「一変量の分布」章を参照してください。

「分位点範囲の外れ値」ユーティリティは、データにある欠測値コードを識別するのにも便利です。前述したように、一部の業界では、欠測値を「9」として入力しています（「999」や「9999」など）。このユーティリティでは、上側分位点以上のデータ値のなかで「9」が連続している整数をすべて、欠測値コードの候補として取り上げます。このユーティリティを使って、それらの候補をデータテーブルの「欠測値のコード」列プロパティに追加することもできます。

[分位点範囲の外れ値] のオプション

「分位点範囲の外れ値」パネルでは、外れ値の計算方法と管理方法を指定できます。図3.6は、デフォルトの「分位点範囲の外れ値」ウィンドウです。

図3.6「分位点範囲の外れ値」ウィンドウ



上側分位点もしくは下側分位点から、分位点範囲（上側分位点と下側分位点の差）をQ倍したもの以上離れている点がすべて外れ値とみなされます。Qの値と何パーセントの分位点にするかは、変更できます。

裾の分位点　分位点範囲を計算するために使われる下側分位点の累積確率。なお、上側分位点の累積確率は、1からこの値を引いたものとみなされます。たとえば、この「裾の分位点」に0.1が指定された場合、分位点範囲は、90%分位点から10%分位点を引いたものです。デフォルトの値は0.1です。

Q　外れ値を定義するのに使われる乗数。「裾の分位点」で定義された下側分位点もしくは上側分位点から、分位点範囲をQ倍したもの以上離れている値が外れ値とみなされます。Qの値を大きくするほど、外れ値の検出がより控えめになります。デフォルトの値は3です。

検索を整数に限定　外れ値の候補として取り上げるデータ値を整数だけに限定します。この機能は、欠測値コードやエラーコードを見つけるのに役立ちます。

外れ値のある列のみ表示　レポートに表示する列を、外れ値のある列だけに限定します。

特定の方法によって外れ値を探し出した後、レポートに表示されているこれらの外れ値に対して、さまざまな処理を行えます。ある列における外れ値に対して処理をしたい場合には、まず、「分位点範囲の外れ値」レポートでその列を選択してください。

行の選択　選択した列で外れ値を含む行が、データテーブルで選択されます。

行の除外　選択した行の「行の除外」属性を有効にします。この処理が終わった後、「分位点範囲の外れ値」レポートを更新するには、[**再スキャン**]をクリックしてください。

セルの色　データテーブルにおいて、外れ値を含むセルが色分けされます。

行の色分け　データテーブルにおいて、外れ値を含む行が色分けされます。

「欠測値のコード」に追加　選択した列の外れ値が、その列の「欠測値のコード」列プロパティに追加されます。このオプションを使うと、欠測値コードやエラーコードを、欠測値として定義できます。欠測値コードやエラーコードは、整数であることが多く、また、9がいくつか並んだ正または負の整数で表されることが多いです。この処理が終わった後、「分位点範囲の外れ値」レポートを更新するには、[**再スキャン**]をクリックしてください。

欠測値に変更 データテーブルにおいて、外れ値が欠測値に置換されます。データ値を欠測値に置換するには注意が必要です。データが無効か、正確でないとわかっている場合にのみ、データ値を欠測値に置換してください。この処理が終わった後、「分位点範囲の外れ値」レポートを更新するには、[**再スキャン**] をクリックしてください。

再スキャン 何らかの外れ値の処理をした後に、レポートを更新したい場合には、この［再スキャン］を行ってください。

閉じる 「分位点範囲の外れ値」パネルを閉じます。

「分位点範囲の外れ値」レポート

「分位点範囲の外れ値」レポートには、すべての列が、指定のオプションを使って検出された外れ値とともに表示されます。また、上側分位点と下側分位点、そして上側閾値と下側閾値が表示されます。これらの閾値の外側にある値は、外れ値とみなされます。各列の外れ値の個数も表示されます。各外れ値の値は、レポートの最後の列に表示されます。1つの列で複数回現れる外れ値については、その回数が括弧内に示されます。外れ値のない列はレポートに表示しない場合には、[**外れ値のある列のみ表示**] を選択します。

このレポートを読み取る際には、以下の点に注目してください。

- エラーコード。連続尺度のデータの場合、整数の値が大きいほど疑わしく、エラーコードと考えられます。たとえば、上側分位点が0.5程度で、下側分位点が-0.5程度である場合、1049や-777といった極端に大きな整数値はエラーコードである可能性が高いでしょう。
- ゼロ。欠測値を「0」とコーディングしている場合もあるでしょう。たとえば、データの大半が比較的大きな値で、0が外れ値として存在している場合、その0は欠測値である可能性が高いでしょう。

「9」を含むデータ

「分位点範囲の外れ値」ウィンドウのおける「「9」を含むデータ」レポートには、欠測値コードかもしれないデータ値を含む列が表示されます。欠測値コードの候補として取り上げられるデータ値は、上側分位点よりも大きな値で、かつ、9が連続しているもの（たとえば9999）のなかで、最大となっているものです。それらが頻出している場合、それらの外れ値は実際には欠測値コードであると考えられます。それらの度数が少ない場合は、単なる外れ値であるのか、それとも、欠測値コードであるのかを、さらに調べる必要があります。「「9」を含むデータ」レポートには、上側分位点も表示されます。

このレポートは、欠測値コードと類推されるデータ値が存在している場合にのみ表示されます。

「欠測値のコード」に最大「9」を追加 選択した外れ値が「欠測値のコード」列プロパティに追加されます。この処理の後、「分位点範囲の外れ値」レポートを更新するには、[**再スキャン**] をクリックしてください。

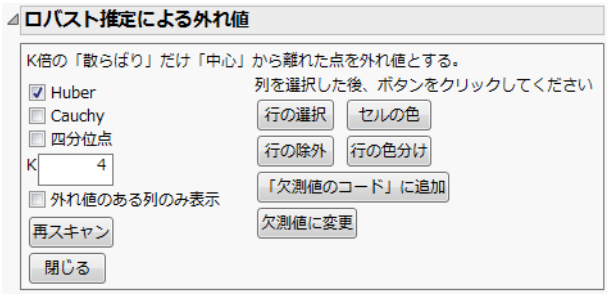
最大「9」を欠測値に変更 選択された外れ値が、データテーブルにおいて欠測値に置換されます。

メモ: データを変更する処理（［**欠測値に変更**］や［**行の除外**］など）を最初に選択した際、元のデータを保持するために［**名前を付けて保存**］コマンドを使ってデータテーブルを新しいファイルとして保存するよう求める警告ウィンドウが表示されます。このウィンドウが表示されたら、［**OK**］をクリックしてください。また、新しいデータを保存する場合は、［**ファイル**］>［**名前を付けて保存**］を選択し、新しい名前でファイルを保存してください。

ロバスト推定による外れ値

パラメータをロバストに推定する方法は、ロバストでない方法よりも、外れ値の影響を受けにくくなっています。［ロバスト推定による外れ値］には、中心とちらばりを推定するいくつかの推定法があり、中心から大きく離れた値が外れ値とみなされます。図3.7は、デフォルトの「ロバスト推定による外れ値」ウィンドウです。

図3.7「ロバスト推定による外れ値」ウィンドウ



「ロバスト推定による外れ値」オプション

指定されたロバストな方法によって中心とちらばりが推定され、中心から*k*倍のちらばりだけ離れた値が外れ値とみなされます。「ロバスト推定による外れ値」ウィンドウには、ロバストな推定法を選択したり、*k*を指定したりするオプションがあり、また、検出された外れ値を処理するためのツールがあります。

Huber　Huberの M 推定を使用して、中心とちらばりを求めます。これがデフォルトのオプションです。Huber and Ronchetti（2009）を参照してください。

Cauchy　Cauchy分布に従うと仮定して、中心とちらばりを推定します。Cauchy分布を仮定した推定は、破綻点（breakpoints）が高く、通常、Huber推定よりもロバストです。ただし、複数のクラスターにデータが分かれている場合、ある1つのクラスターに近い半分のデータだけしか考慮せず、残りのデータを無視する傾向があります。

四分位点　四分位範囲（IQR; interquartile range）に基づいてちらばりを推定します。中央値が中心の推定値として使われます。また、IQRを1.34898で割った値がちらばりの推定値として使われます。正規分布においては、IQRを1.34898で割った値は、標準偏差です。

K　中心からちらばりの*k*倍以上離れているデータ値を外れ値とみなします。*K*の値を大きくするほど、外れ値の検出がより控えめになります。デフォルトの値は4です。

外れ値のある列のみ表示 レポートに表示する列を、外れ値のある列だけに限定します。

特定の方法によって外れ値を探し出した後、レポートに表示されているこれらの外れ値に対して、さまざまな処理を行えます。ある列における外れ値に対して処理をしたい場合には、まず、「分位点範囲の外れ値」レポートでその列を選択してください。

行の選択 選択した列で外れ値を含む行が、データテーブルで選択されます。

行の除外 外れ値を含む行に、「行の除外」属性を割り当てます。この処理が終わった後、「ロバスト推定と外れ値」レポートを更新するには、[**再スキャン**] をクリックしてください。

セルの色 データテーブルにおいて、外れ値を含むセルが色分けされます。

行の色分け データテーブルにおいて、外れ値を含む行が色分けされます。

「欠測値のコード」に追加 選択した列の外れ値が、その列の「欠測値のコード」列プロパティに追加されます。このオプションを使うと、欠測値コードやエラーコードを、欠測値として定義できます。欠測値コードやエラーコードは、整数であることが多く、また、9がいくつか並んだ正または負の整数で表されることが多いです。この処理が終わった後、「ロバスト推定と外れ値」レポートを更新するには、[**再スキャン**] をクリックしてください。

欠測値に変更 データテーブルにおいて、外れ値が欠測値に置換されます。この処理が終わった後、「ロバスト推定と外れ値」レポートを更新するには、[**再スキャン**] をクリックしてください。

再スキャン 何らかの外れ値の処理をした後に、レポートを更新したい場合には、この［再スキャン］を行ってください。

閉じる 「ロバスト推定と外れ値」パネルを閉じます。

多変量ロバスト推定による外れ値

「多変量ロバスト推定による外れ値」ツールは、「多変量の相関」プラットフォームの［ロバスト］オプションを使用して、多変量間の関係を調べます。「多変量の相関」プラットフォームの詳細については、『多変量分析』の「相関と多変量の手法」に関する章を参照してください。

外れ値分析

［外れ値分析］では、各点から多変量正規分布の中心までのMahalanobisの距離を計算します。この距離は、特定の相関行列をもつ多変量正規分布の確率密度を表す等高線に対応しています。中心から離れるほど、外れ値である確率が高くなります。Mahalanobisの距離などの詳細については、『多変量分析』の「多変量の相関」章を参照してください。

行を除外した後、分析を再実行するか、またはユーティリティを閉じます。分析を再実行すると、除外されていない行だけで多変量分布の中心が再計算されます。なお、除外された行をデータテーブルにて非表示にしない限り、それらの行はグラフに表示されたままです。

「Mahalanobisの距離」の赤い三角ボタンのメニューから、[**Mahalanobisの距離を保存**] を選択すると、Mahalanobisの距離をデータテーブルに保存できます。

図3.8「多変量ロバスト推定による外れ値」の「Mahalanobisの距離」プロット

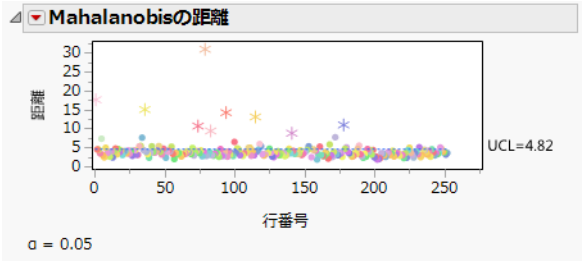


図3.8は、16列から計算されたMahalanobisの距離です。プロットには、4.82の位置に上側管理限界（UCL; upper control limit）も描かれています。このUCLは、どの程度のデータを外れ値として扱うかを考えるのに役立ちます。ただし、どれを外れ値とするかは、各状況に応じて判断してください。UCLの詳細については、Mason and Young（2002）を参照してください。

［多変量のロバスト推定による外れ値分析］のオプション

［多変量のロバスト推定による外れ値分析］の赤い三角ボタンのメニューには、多変量のデータを分析するための数々のオプションがあります。これらのオプションとそれらの説明については、『多変量分析』の「多変量の相関」章を参照してください。

多変量のk近傍法外れ値

外れ値を検出する方法の1つとして、他のほとんどの点から離れている点を選ぶことが考えられます。それには、近傍点からの距離を用いることが考えられます。「多変量のk近傍法外れ値」ユーティリティは、各点から、そのK番目に近い点までのユークリッド距離のプロットを描きます。ユーザは近傍点の個数Kに対する最大数（ここでは、それをkとします）を指定します。このとき、プロット数を少なくするためにフィボナッチ数列によっていくつかの値をスキップしながら、$K = 1, 2, 3, \ldots, k$に対してプロットが描かれます。

このアプローチは、指定したkの値による影響を受けます。kの値が小さいと点を外れ値として識別できないことがあり、また、kの値が大きいと点が間違って外れ値に分類されることがあります。

- 2 ～ 3 個の近傍点だけを調べるように、k に小さい値を指定したとしましょう。k 個より多い点のクラスターがあり、そのクラスターが残りの点から離れている場合、クラスター内の点から近傍点までの距離は小さくなります。その場合は、外れ値のクラスターを検出できない可能性があります。
- 逆に、kに大きな値を指定したとしましょう。k個より少ないデータ点からなるクラスターがある場合、そのクラスターに属するすべての点が外れ値とみなされます。つまり、kに大きな値を指定すると、「それらの点がクラスター内の点である」という事実を無視して、そのクラスター全体を外れ値としてみなすことになります。

「K近傍法」レポート

コマンドのリストから［多変量のk近傍法外れ値］を選択すると、考慮する最遠端の近傍点、つまり上限として使用するkの値を指定するよう求められます。デフォルトでは8に設定されています。

レポートには、指定された*k*までの、*K*の各値に対するプロットが描かれます。各プロットの*K*の値は、縦軸のラベルとして「第a近傍点までの距離」という形式で表示されます。ここで、aは*a*番目に近い近傍点を表す整数です。各プロットには、*i*行目の点から*a*番目に近い近傍点までの距離が示されます。*K*の複数の値にわたって近傍点からの距離が大きい点は、概して外れ値だと言えます。

プロットの上には以下のボタンがあります。

選択された行の除外　選択した点に対応する行を、今後の分析から除外します。データテーブルでは、「除外」の行属性がこれらの行に割り当てられます。「K近傍法」レポートを再実行するか閉じるかを尋ねられます。分析を再実行すると、新しい近傍点が識別されます。プロットが更新され、除外された点は非表示になります。

散布図行列　分析対象のすべての列に対する散布図行列を含む別のウィンドウを開きます。「k近傍法」プロットで点を選択し、散布図行列でそれらの点を見ることで、外れ値であるかもしれない値を探索できます。

閉じる　「K近傍法」レポートを閉じます。

「外れ値を調べる」ユーティリティのオプション

以下のオプションについて詳しくは、『JMPの使用法』の「JMPのレポート」章を参照してください。

やり直し　分析を繰り返すオプションや、やり直すオプションを含みます。また、［自動再計算］オプションを選択すると、このオプションに対応しているプラットフォームにおいて、データテーブルに加えた変更が、該当するレポートに即座に反映されるようになります。

スクリプトの保存　レポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。

Byグループのスクリプトを保存　By変数の全水準に対するレポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。起動ウィンドウでBy変数を指定した場合のみ使用可能です。

「外れ値を調べる」ユーティリティの別例

「多変量のk近傍法外れ値」の例

「Water Treatment.jmp」サンプルデータには、都市の排水処理工場内の38個のセンサーで毎日計測された値が含まれています。これらのデータから外れ値の可能性がある値を検出しましょう。外れ値の原因としては、センサーの故障、嵐、その他の状態が考えられます。

1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Water Treatment.jmp」を開きます。
2. ［**分析**］>［**スクリーニング**］>［**外れ値を調べる**］を選択します。
3. 「**Sensor Measurements**」列グループを選択し、［**Y, 列**］をクリックします。
4. ［**OK**］をクリックします。
5. ［**多変量のk近傍法外れ値**］を選択します。

6. k近傍法のKとして「13」を入力します。
7. ［**OK**］をクリックします。

図3.9 多変量のk近傍法外れ値の例での外れ値

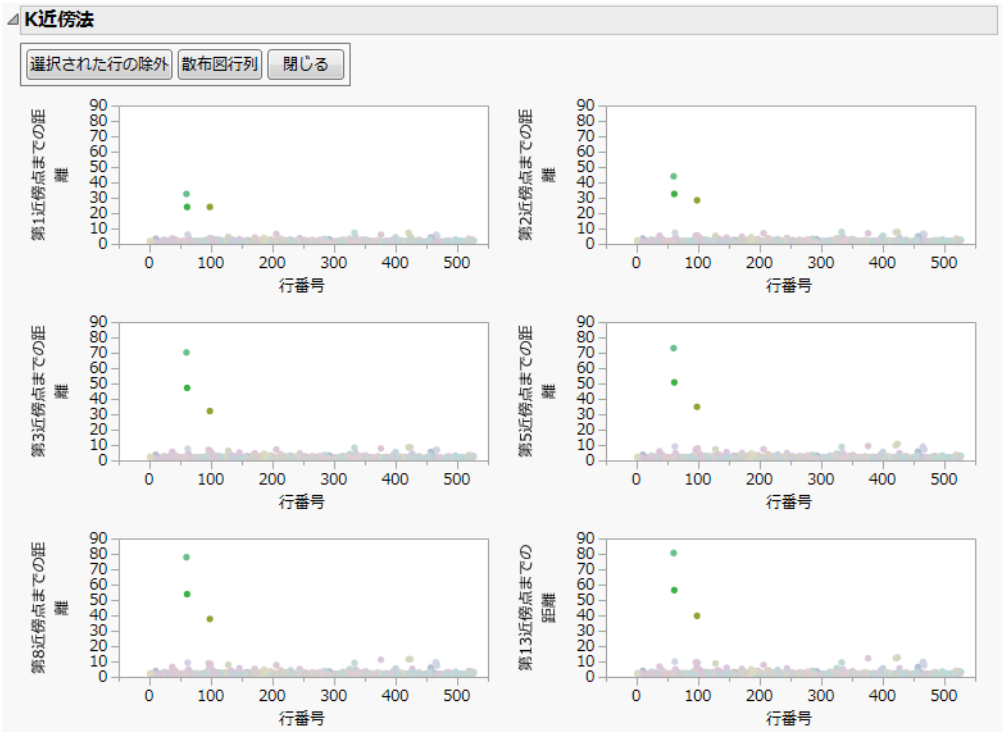
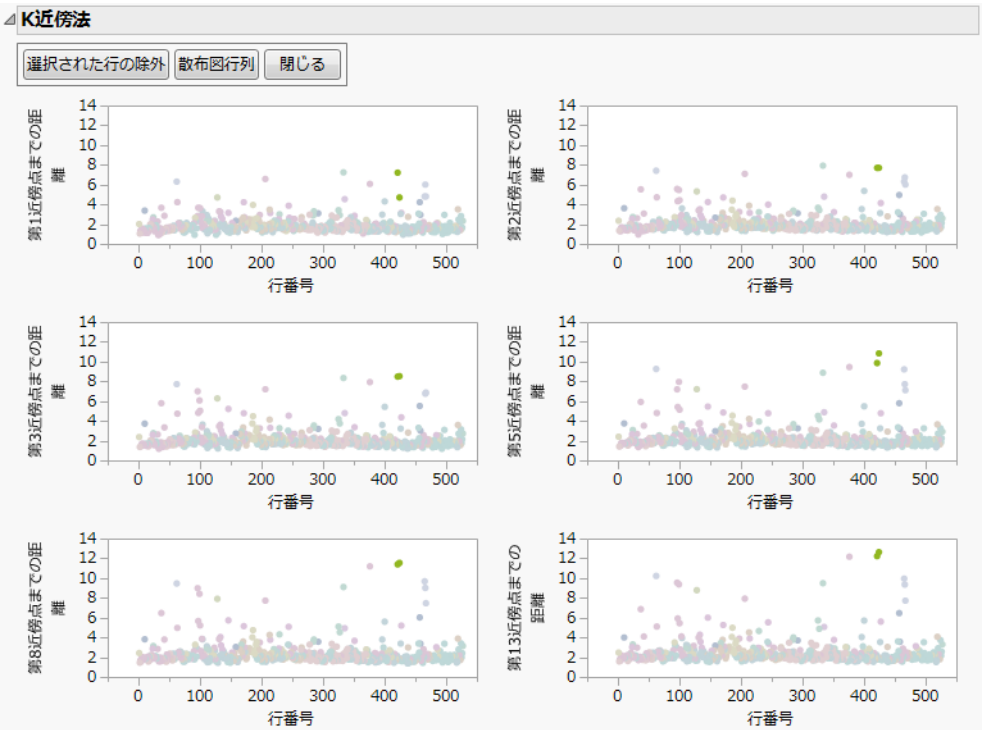


図3.9の「K近傍法」プロット内では、3つの外れ値が選択されています。これら3つのデータ点それぞれが、排水処理工場内の二次沈殿槽が正常に動作していないとされる行に対応しています。これら3つのデータ点は不良な機器によるものであるため、これらを今後の分析から除外します。

8. 3つの外れ値を選択して、［**選択された行の除外**］をクリックします。

   ユーティリティを再実行するかウィンドウを閉じるかを尋ねるダイアログが表示されます。
9. ［**再実行**］をクリックします。
10. k近傍法のKとして「13」を入力します。
11. ［**OK**］をクリックします。

図3.10 多変量のk近傍法の例での外れ値



次に、行400に近い薄緑色の2つの外れ値に注目してください。この2つの点は、*k*が大きくなるにつれ、距離はどのように変化しているでしょうか。これらの2つの行は、排水処理工場による固形物負荷が高かったときのものです。これらのデータ点の第13近傍点までの距離は比較的高いものですが、調査に含めたいものなので、除外はしません。その代り、この後の分析ではそのことを念頭に置いておく必要があります。

「欠測値を調べる」ユーティリティ

欠測値は、統計分析の結果に影響する場合があります。たとえば、もし寿命の調査において、多くの健康な人々のデータが欠測していると、それらを考慮せずに分析した結果は、寿命を短く見積もってしまう方向にバイアスがかかります。統計分析を行う前には、欠測値があるかどうかだけではなく、どのような欠測が生じているかを理解しておかなければなりません。

「欠測値を調べる」ユーティリティでは、欠測値を調べる、いくつかの方法が用意されています。また、多変量正規分布に基づいて欠測値を補完することもできます。この補完は、**ランダムな欠測**（MAR; Missing At Random）、つまり、「データが欠測する確率は、観測されているデータにのみ依存している」という状態を前提にしています。ランダムな欠測（MAR）**ではない**と思われる場合は、多くのプラットフォームに備わっている［欠測値をカテゴリとして扱う］オプションの使用などを検討してください。このオプションの詳細については、『基本的な回帰モデル』の「モデルの指定」章を参照してください。

「欠測値を調べる」ユーティリティの例

「Arrhythmia.jmp」サンプルデータには、452人の患者の心電図（ECG）データが含まれています。このデータは、元々、心電図のさまざまなパターンから不整脈を分類するために収集されたものです。ただし、このデータテーブルには欠測値があります。ここでは、主にこれらの欠測値を調べて、必要に応じて値を補完してみましょう。欠測値の補完は、連続尺度の列に対してのみ行うので、ここでは2つの段階に分けて分析を行います。

欠測値を調べる

1. ［ヘルプ］>［サンプルデータライブラリ］を選択し、「Arrhythmia.jmp」を開きます。
2. ［分析］>［スクリーニング］>［欠測値を調べる］を選択します。
3. すべての列（280個）を選択し、［Y, 列］をクリックします。
4. ［OK］をクリックします。［欠測値のある列のみ表示］チェックボックスを選択します。

図3.11 欠測値レポート

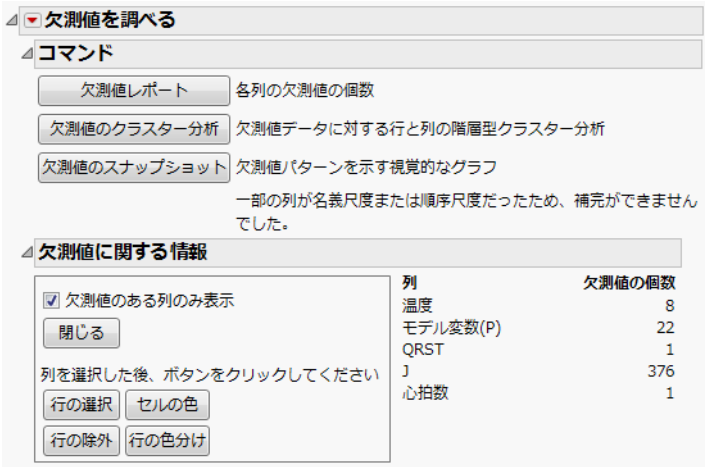


図3.11に示す「欠測値に関する情報」から、欠測値のある列は5つだけであることがわかります。合計452行のうち、「J」列には376個の欠測値があります。大量のデータが欠測しているため、たとえそれらの欠測値を補完したとしても分析には役立ちません。ただし、ランダムな欠測（MAR）ではないと判断できる場合には、各プラットフォームの［欠測値をカテゴリとして扱う］オプションを使って「J」列をモデルに含めると、それらの欠測情報が役立つこともあります。

ここには、［多変量正規分布による補完］と［多変量の特異値分解補完］という2つの補完オプションが表示されていません。分析に含まれている一部の列はカテゴリカルであることを示すメッセージが表示されます。このデータテーブルには、数値データタイプで、かつ名義尺度の列がいくつかあります。これらの列は補完に使用できません。

欠測値の補完

欠測値のある5つの列は連続尺度です。データテーブル内の連続尺度の列に対し、欠測値補完を使用して「**J**」**列**以外の4つの列の欠測値を保管することにします。この場合、値の欠測している確率は、除外された名義尺度の変数ではなく、連続尺度の変数にのみ依存していると暗黙的に仮定します。この新しい分析を行うには、「欠測値を調べる」ユーティリティをもう一度起動する必要があります。

1. ［**分析**］>［**スクリーニング**］>［**欠測値を調べる**］を選択します。
2. 起動ウィンドウで、「**280列**」の横にある赤い三角ボタンをクリックします。

   列フィルタメニューを使用して、「列の選択」リストに連続尺度の列だけを表示します。
3. ［**尺度**］>［**すべて選択解除**］を選択します。

   すべての列が「列を選択」リストから削除されます。
4. ［**尺度**］>［**連続尺度**］を選択します。

   これで、「列を選択」リストには、連続尺度である207個の列だけが表示されます。
5. 207個の列をすべて選択します。Ctrlキーを押しながら「**J**」列をクリックしてこの列を選択解除し、［**Y, 列**］をクリックします。
6. ［**OK**］をクリックします。
7. ［**多変量正規分布による補完**］をクリックします。

   共分散の推定値として縮小させた推定値を用いるかどうかを尋ねるウィンドウが表示されます。
8. ［**はい**］をクリックします。

   ［**名前を付けて保存**］コマンドを使って元のデータを保存するよう求める警告ダイアログが表示されます。
9. ［**OK**］をクリックします。

図3.12 欠測値補完に関する情報

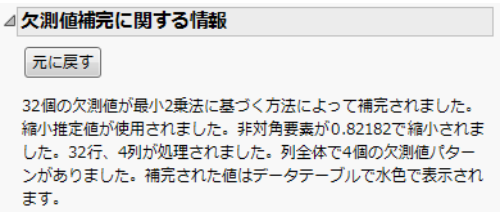


図3.12に示す「欠測値補完に関する情報」には、補完された欠測値の個数などの具体的な情報が表示されます。4つの列において欠測値が補完されました。

「欠測値を調べる」ユーティリティの起動

「欠測値を調べる」モデル化ユーティリティを起動するには、［**分析**］>［**スクリーニング**］>［**欠測値を調べる**］を選択します。関心のある列を［Y, 列］リストに入力します。

メモ:「欠測値を調べる」ユーティリティに入力できるのは、数値データタイプの列だけです。

欠測値レポート

起動ウィンドウで［OK］をクリックした後、レポートが開き、「コマンド」アウトラインと「欠測値に関する情報」が表示されます。以下のコマンドがあります。

- 「欠測値レポート」（49ページ）
- 「欠測値のクラスター分析」（49ページ）
- 「欠測値のスナップショット」（50ページ）
- 「多変量正規分布による補完」（50 ページ）（起動ウィンドウに指定した列に、名義尺度や順序尺度のものがある場合は使用できません。）
- 「多変量の特異値分解補完」（51 ページ）（起動ウィンドウに指定した列に、名義尺度や順序尺度のものがある場合は使用できません。）

欠測値レポート

［欠測値レポート］をクリックすると、各列の名前と、その列の欠測値の個数をリストした「欠測値に関する情報」が開きます。

欠測値のある列のみ表示 欠測値がない列をリストから除外します。

閉じる 「欠測値に関する情報」を閉じます。

行の選択 データテーブル内で、「欠測値に関する情報」において選択した列で欠測値を含む行を選択します。

行の除外 データテーブル内で、「欠測値に関する情報」においいて選択した列で欠測値を含む行に対し、「除外」の行属性を割り当てます。

セルの色 データテーブル内で、「欠測値に関する情報」において選択した列で欠測値を含むセルに色をつけます。

行の色分け データテーブル内で、「欠測値に関する情報」において選択した列で欠測値を含む行を選択します。

欠測値のクラスター分析

［欠測値のクラスター分析］は、欠測値のパターンに対して階層型クラスター分析を行います。

- プロット右側の樹形図は、欠測値のパターンに対するクラスター分析の結果です。各行は、［テーブル］>［欠測値パターン表示］を使用して得られる行と同じものです。
- プロット下側の樹形図は、変数に対するクラスター分析の結果です。

このレポートを使用して、特定グループの列に類似した欠測値パターンの傾向があるかどうかを判断できます。

プロットの行は、欠測値のパターンによって定義されます。各パターンに対してそれぞれ1つの行があります。列は変数に対応しています。赤いセルは、プロットの下に表示されている列が欠測値になっていることを示しています。セルの上にカーソルを置くと、そのセルが表す行番号のリストが表示されます。プロット内をクリックすると、その欠測パターンのデータ行が選択されます。また、プロットには選択されていることを示す縦の棒が描かれます。

欠測値のスナップショット

［欠測値のスナップショット］をクリックすると、欠測値のセルプロットが表示されます。列は変数を表します。黒いセルは欠測値を示します。このプロットは、データ収集期間が終わる前に被験者が調査から外れることができるような経時測定データの欠測を理解する上で特に便利です。

多変量正規分布による補完

「多変量正規分布による補完」ユーティリティは、多変量正規分布に基づいて欠測値を補完します。この手順では、すべての変数が連続尺度でなければなりません。アルゴリズムは最小2乗法に基づく補完法を使用します。共分散行列はペアごとの共分散を使用して構築されます。対角要素（分散）は、各変数に対する欠測値以外のすべての値を用いて計算されます。任意の2つの変数に対する対角以外の要素は、両変数に対して欠測値でないすべてのオブザベーションを用いて計算されます。共分散行列が特異な場合、アルゴリズムはMoore-Penrose疑似逆行列に基づく最小2乗最小ノルム法を使って補完を行います。

［多変量正規分布による補完］では、共分散の推定値として縮小させた推定値（shrinkage estimation）を使用できます。縮小させた推定値を使用すると、共分散行列の推定値を改善できます。縮小させた推定値の詳細については、Schafer and Strimmer（2005）を参照してください。

メモ: 検証列が指定されている場合、共分散行列は学習セットだけを用いて計算されます。

欠測値補完に関する情報

「欠測値補完に関する情報」には、欠測値補完に関する次のような情報が表示されます。

- 補完の手法（最小2乗法または最小2乗最小ノルム法）
- 補完された個数
- 縮小推定値の使用の有無
- 非対角要素を収縮した係数
- 処理された行数および列数
- 欠測値パターンの数

補完が完了すると、データテーブル内で補完された値に対応するセルが明るい青色で表示されます。「欠測値に関する情報」が開いている場合は、表示内容が更新され、欠測値がなくなったことが示されます。

補完を取り消すには［**元に戻す**］をクリックします。すると、補完された値が元の欠測値に戻ります。

多変量の特異値分解補完

「多変量の特異値分解補完」ユーティリティは、特異値分解（SVD; Singular Value Decomposition）を使って欠測値を補完します。このユーティリティは、変数の数が数百を超えるようなデータに便利です。特異値分解の計算には共分散行列の計算が必要ないため、多数の変数を含むデータテーブルの場合は、特異値分解の手法をお勧めします。この手順では、すべての変数が連続尺度でなければなりません。

特異値分解はオブザベーションの行列**X**を**X = UDV'**で表します。ここで、**U**と**V**は直交行列、**D**は対角行列です。

「多変量の特異値分解補完」ユーティリティのデフォルトで使用される特異値分解アルゴリズムは疎なLanczos法であり、これはIRLBA（**implicitly restarted Lanczos bidiagonalization method**）とも呼ばれます。Baglama and Reichel(2005)を参照してください。このアルゴリズムにより、以下のことが行われます。

1. 初期値として、各欠測値を、その列の平均に置き換えます。
2. 欠測部分が置き換えられた行列**X**を、特異値分解します。
3. 欠測値を、特異値分解から得られた**UDV'**行列の対応する要素に置き換えます。
4. 行列**X**が変化しなくなるまで、手順2と3の特異値分解を繰り返します。

補完法

［多変量の特異値分解補完］をクリックすると、「補完法」ウィンドウが開き、推奨される設定が表示されます。

特異ベクトルの数　補完で計算および使用される特異ベクトルの数です。

> **メモ：**特異ベクトルの次元を大きくしすぎないことが重要です。次元が大きいと、特異値分解と補完の各反復において、数値が変化しません。

最大反復回数　欠測値の補完に使用される反復回数です。

反復ログの表示　反復回数を示す「詳細」レポートが開き、条件の詳細が表示されます。

データが大きい場合は、進捗バーが表示されて、特異値分解が完了した次元が示されます。いつでも補完を停止し、その次元の数を使用できます。

欠測値補完に関する情報

「欠測値補完に関する情報」には、欠測値補完に関する次のような情報が表示されます。

- 補完の手法
- 補完された個数
- 処理された行数および列数

補完が完了すると、「欠測値に関する情報」にはどの列にも欠測値が存在しないことが自動的に示されます。補完された値はデータテーブル内で明るい青色で表示されます。

補完を取り消すには［**元に戻す**］をクリックします。すると、補完された値が元の欠測値に戻ります。

「欠測値を調べる」ユーティリティのオプション

以下のオプションについて詳しくは、『JMPの使用法』の「JMPのレポート」章を参照してください。

やり直し　分析を繰り返すオプションや、やり直すオプションを含みます。また、［自動再計算］オプションを選択すると、このオプションに対応しているプラットフォームにおいて、データテーブルに加えた変更が、該当するレポートに即座に反映されるようになります。

スクリプトの保存　レポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。

Byグループのスクリプトを保存　By変数の全水準に対するレポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。起動ウィンドウでBy変数を指定した場合のみ使用可能です。

JMP PRO「検証列の作成」ユーティリティ

検証（validation）とは、データの一部をモデルパラメータの推定に使用し、残りのデータでモデルの予測能力を評価することを指します。このようにデータを分割することにより、モデルがオーバーフィット（過学習）することを回避できます。

検証列は、データを2つまたは3つの部分に分けます。

- モデルパラメータの推定に使うデータを、学習セットといいます。
- 予測能力が高いモデルを選ぶのに用いるデータを、検証セットといいます。
- モデルが選択された後、モデルの予測能力をチェックするデータを、テストセットといいます。

検証列は、［モデルのあてはめ］プラットフォームで検証の手法の1つとして使用できます。

JMP PRO「検証列の作成」ユーティリティの例

「Lipid Data.jmp」データテーブルは、カリフォルニア州の、ある病院で収集された、95人の患者に関する血液測定値・身体測定値・質問票データです。この例では、モデルの検証に用いることができる検証列を作成してみます。

1. ［ヘルプ］>［サンプルデータライブラリ］を選択し、「Lipid Data.jmp」を開きます。
2. ［分析］>［一変量の分布］を選びます。
3. 「性別」に［Y, 列］を割り当てます。［OK］をクリックします。

図3.13「Lipid Data.jmp」での「**性別**」の分布

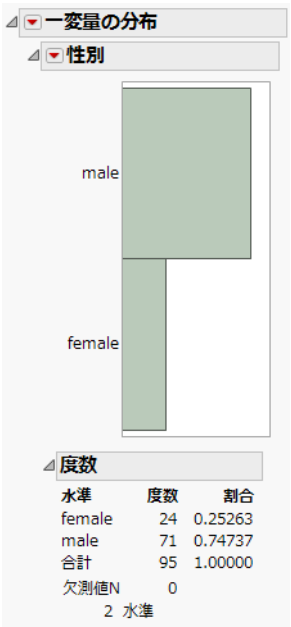


図3.13は、データ内の「**性別**」の分布を示しています。被験者の男性と女性の割合は同じではないことに注目してください。女性のデータのほうが少ないので、検証セットと学習セット全体では性別のバランスをとる必要があります。

4. ［**分析**］>［**予測モデル**］>［**検証列の作成**］を選択します。
5. ［**層化無作為抽出**］をクリックします。
6. 検証データを層化抽出する際に層とする列として、「**性別**」を選択します。
7. ［**OK**］をクリックします。

   データテーブルに「検証」列が追加されます。モザイク図を作成すると、検証セットと学習セットの分布を確認できます。
8. ［**分析**］>［**二変量の関係**］を選びます。
9. 「**検証**」を［Y, 目的変数］に、「**性別**」を［X, 説明変数］に割り当てます。
10. ［**OK**］をクリックします。

図3.14 検証セットと学習セットにおける性別の分布

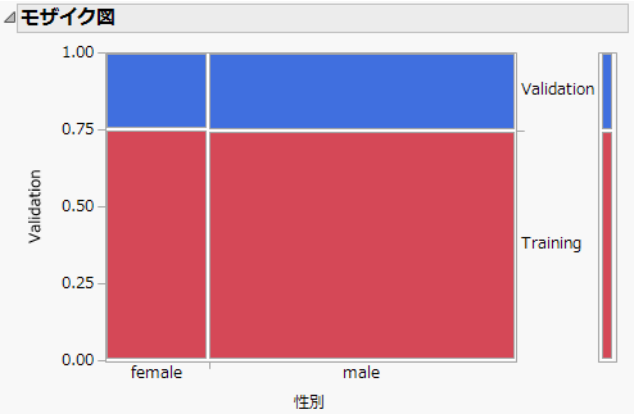


図3.14は、検証セットと学習セットにおける**「性別」**の分布を示しています。男性と女性それぞれの約75%が学習セットに、男性と女性それぞれの約25%が検証セットに含まれていることがわかります。

JMP PRO「検証列の作成」ユーティリティの起動

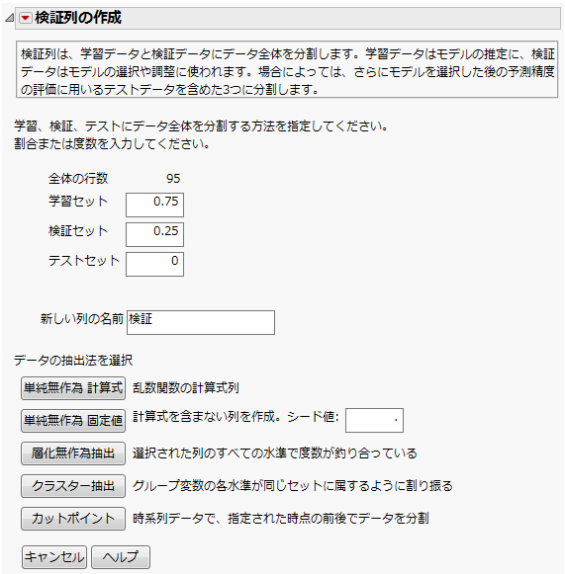
「検証列の作成」ユーティリティを起動するには、2通りの方法があります。

- ［分析］>［予測モデル］>［検証列の作成］を選択します。「「検証列の作成」ウィンドウ」（54ページ）を参照してください。
- プラットフォームの起動ウィンドウで［検証］をクリックします。「プラットフォームの起動ウィンドウで［検証］をクリックする」（56ページ）を参照してください。

JMP PRO「検証列の作成」ウィンドウ

「検証列の作成」ウィンドウでは、各セットに割り振る行の割合（データ全体に対する割合）、もしくは行数を指定した後、データの抽出法を選択します。

図3.15「検証列の作成」ウィンドウ



- 「学習セット」・「検証セット」・「テストセット」の横には、これらのセットのそれぞれに含めたい行の割合または行数を入力します。デフォルト値は、学習セットが行の約75%で、検証セットが行の25%です。
- 「新しい列の名前」の横には、検証列の名前を入力します。

データの抽出には5通りの方法があります。

単純無作為 計算式　入力された配分に基づいてデータをセットに分けます。たとえば、デフォルトの値が入力されている場合、各行は学習セットに含まれる確率が0.75、検証セットに含まれる確率が0.25となります。この方法では、無作為抽出を行う計算式が列に保存されます。計算式を表示するには、「列」パネルで列名の右側にあるプラスアイコンをクリックします。

単純無作為 固定値　入力された配分に基づいてデータをセットに分けます。たとえば、デフォルトの値が入力されている場合、各行は学習セットに含まれる確率が0.75、検証セットに含まれる確率が0.25となります。今後、この配分を再現するための乱数シード値を入力できます。この方法では、計算式ではなくてデータ値が列に保存されます。

層化無作為抽出　指定した列の水準に基づいてデータをセットに分けます。学習セット、検証セット、およびテストセットのそれぞれで、列の水準のバランスをとりたい場合には、このオプションを使用します。

［層化無作為抽出］をクリックすると、層化無作為抽出の対象となる列を1つ以上選択できるウィンドウが表示されます。［OK］をクリックすると、検証列がデータテーブルに追加されます。［単純無作為 固定値］の場合と同様に、行は指定された配分に基づいて無作為にデータセットに割り当てられます。ただし、単純無作為抽出ではなくて、指定された列の水準または水準の組み合わせを層として、層化無作為抽出が行われます。

データテーブルには、層別変数が記された「ノート」プロパティのある列が追加されます。

クラスター抽出 指定した列の水準全体が、または、2つ以上の列の水準の組み合わせ全体が同じセットに割り振られるように、データを各セットに分けます。このオプションは、ある列における水準を別々のセットに分割するのが適切でない場合に使用します。

［クラスター抽出］をクリックすると、1つまたは複数のグループ列を指定するためのウィンドウが表示されます。［OK］をクリックすると、グループ列の水準が各セットに無作為に割り振られます。まず、検証セットやテストセットに割り振るのですが、ある水準を検証セットやテストセットに割り振ったときに、指定した割合や行数より大きくなっても、その水準はそのセットに割り振ります。そのため、学習セットに割り振られる行数は、指定されたものよりも少なくなります。結果的に得られる各セットの大きさは、指定したものとわずかに異なります。

カットポイント 時系列のカットポイント（データを分割する時点）に基づいてデータをセットに分けます。このオプションは、データを期間に基づいてセットに割り当てたい場合に使用します。

［カットポイント］をクリックすると、1つまたは複数の列を選択して期間を定義するためのウィンドウが表示されます。［OK］をクリックすると、データを分割するカットポイントの情報を知らせるJMP警告が表示されます。そして、そのカットポイントで分割した列がデータテーブルに追加されます。学習セットは、最初のカットポイントから2番目のカットポイントまでの行で構成されます。検証セットは、2番目のカットポイントから3番目のカットポイントまでの行で構成されます。テストセットは、残りの行で構成されます。これらのセットは、指定した行の割合または数を反映するように選択されます。

JMP PRO プラットフォームの起動ウィンドウで［検証］をクリックする

プラットフォームの起動ウィンドウをすでに開いていて、その起動ウィンドウで検証列を作成したい場合には、本節で説明する方法を使用してください。ただし、次の点に注意してください。

- プラットフォームが「検証」列に対応している必要があります。
- どの列も「列の選択」リストで選択されていない状態にて、以下の操作を行って下さい。

プラットフォームの起動ウィンドウで［検証］ボタンをクリックします。「学習セット」に0.7、「検証セット」に0.3、「テストセット」に0.0というデフォルト設定が入力された、「検証列の作成」ウィンドウが開きます。

1. 「学習セット」・「検証セット」・「テストセット」の横に目的の割合または行数を入力します。
2. 「新しい列の名前」の横に新しい列の名前を入力します。
3. ［OK］をクリックします。

計算式を含む新しい列がデータテーブルに表示されます。また、起動ウィンドウでは、追加された新しい列が「検証」の役割に割り当てられます。

メモ: プラットフォームの起動ウィンドウから「検証列の作成」ユーティリティを起動することは、［分析］>［予測モデル］>［検証列の作成］から［単純無作為 計算式］を選ぶのと同じことです。［単純無作為 固定値］、［層化無作為抽出］、［クラスター抽出］、［カットポイント］は使用できません。

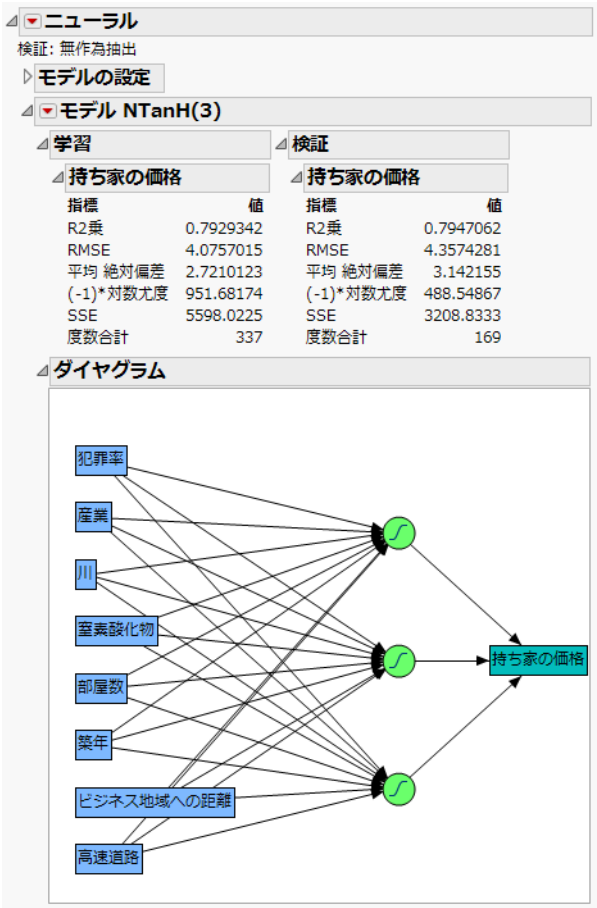
第4章

ニューラルネットワーク
複雑な非線形関係のための多層モデル

JMP PRO このプラットフォームのほとんどの機能は JMP Pro 専用であり、このアイコンで示されています。

「ニューラル」プラットフォームは、入力層、出力層、および、1～2層の隠れ層（中間層）をもつ多層パーセプトロンニューラルネットワークをあてはめます。ニューラルネットワークは、柔軟な関数によって、入力変数から1つまたは複数の応答変数を予測します。ニューラルネットワークは、応答曲面の関数式を求めないでよい場合や、入力と応答との関係を明らかにする必要がない場合に、優れた予測を行います。

図4.1 ニューラルネットワークの例

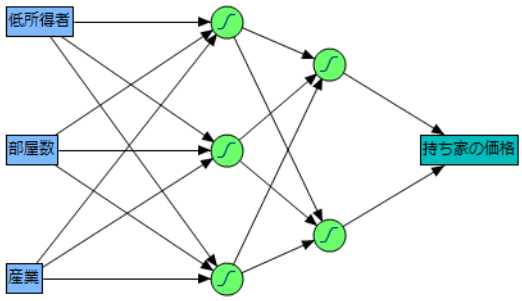


ニューラルネットワークの概要

ニューラルネットワークの入力は、まず、隠れ層（中間層）の各ノードに定義された関数によって変換されます。そして、その変換された関数の結果が、さらに変換されて、出力となります。JMPでは、隠れ層を最大で2層、指定することができます。各隠れ層には、任意の数のノードを含めることができます。

図4.2は、2つの隠れ層をもつニューラルネットワークです。X変数が3つ、Y変数が1つあります。まず、X変数が、左側の隠れ層（第2層）によって変換されます。次に、その結果が、右側の隠れ層（第1層）によって変換されます。そして、さらに、その結果を変換して、出力が求められます。

図4.2 ニューラルネットワークのダイヤグラム



隠れ層の各ノードで使われる関数を、「活性化関数」といいます。活性化関数は、X変数の線形結合を変換したものです。活性化関数の詳細は、「隠れ層の構造」（62ページ）を参照してください。

応答で適用される関数は、連続尺度の応答に対しては単なる線形結合で、名義／順序尺度の応答に対してはロジスティック変換です。

ニューラルネットワークの長所は、応答曲面を柔軟にモデル化できることです。隠れノードと層の数が十分であれば、どんな曲面でも任意の精度で近似することができます。ニューラルネットワークの短所は、通常の回帰のようにX変数がY変数に直結しているのではなく、中間に層が挟まっているために結果の解釈が難しいことです。

「ニューラル」プラットフォームの起動

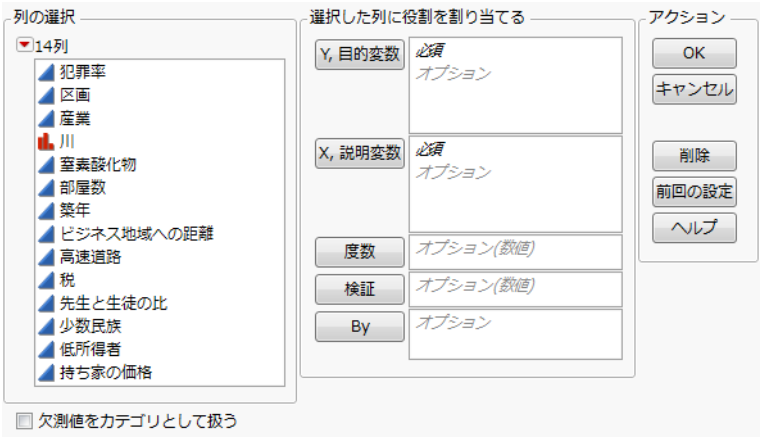
「ニューラル」プラットフォームを起動するには、［**分析**］>［**予測モデル**］>［**ニューラル**］を選択します。

「ニューラル」プラットフォームの分析方法は、2段階で設定します。まず、「ニューラル」起動ウィンドウに、分析に用いる変数を指定します。次に、「モデルの設定」パネルで分析のオプションを指定します。

JMP PRO「ニューラル」起動ウィンドウ

「ニューラル」起動ウィンドウでは、X変数とY変数を指定します。また、オプションで、検証列を指定したり、欠測値をカテゴリとして扱うかどうかを選択します。

図4.3「ニューラル」起動ウィンドウ



Y, 目的変数　応答変数を指定します。なお、複数の応答変数に対するニューラルネットワークでは、（応答変数へのパラメータを除く）すべてのモデルパラメータが共有されています。

X, 説明変数　入力変数を指定します。

度数　度数変数を指定します。

JMP PRO **検証**　検証列を指定します。詳細は、「検証法」（61ページ）を参照してください。「列の選択」リストで列を選択せず、［検証］ボタンをクリックすると、データテーブル内に検証列を作成することができます。「検証列の作成」ユーティリティの詳細については、『予測モデルおよび発展的なモデル』の「モデル化ユーティリティ」章を参照してください。

By　変数の水準ごとに個別にモデルをあてはめる場合に、変数を指定します。

JMP PRO **欠測値をカテゴリとして扱う**　1つの意味があるカテゴリとして欠測値を扱うには、このチェックボックスをオンにします。この機能を使うと、説明変数に欠測値があるデータも、除外されずに、計算に使われます。欠測値にも意味がある場合に便利です。このチェックボックスをオフにした場合は、欠測値のある行は分析から除外されます。

連続尺度の説明変数では、まず、非欠測値の平均で欠測値が置換されます。そして、欠測値に対する指示変数が作成され、説明変数としてモデルに追加されます。なお、「モデルの設定」パネルで［説明変数の変換］オプションを使用して説明変数を変換した場合は、変換した後の変数の平均で欠測値が置換されます。

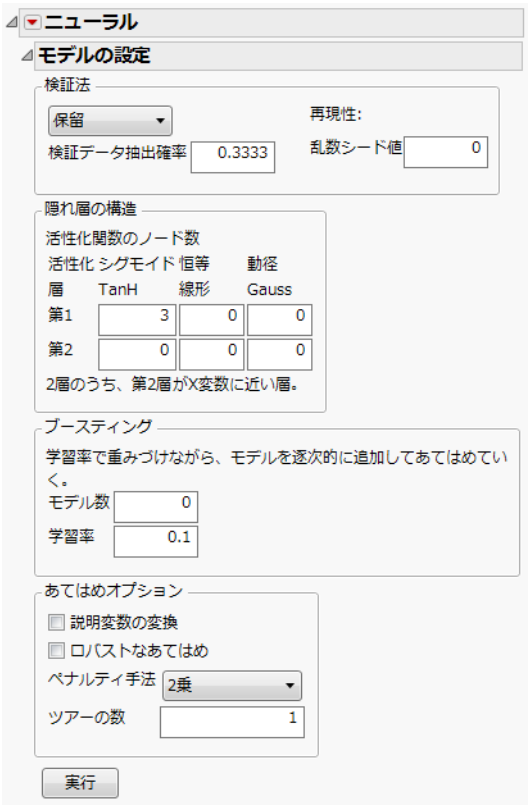
カテゴリカルな説明変数では、欠測値は1つのカテゴリとして扱われます。

「モデルの設定」パネル

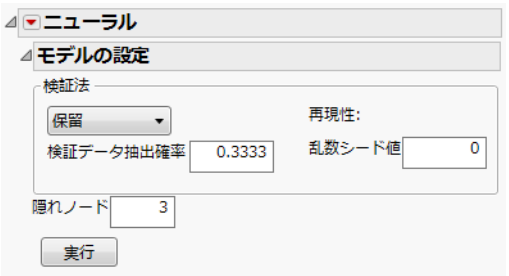
「モデルの設定」パネルでは、検証法や隠れ層の構造を指定します。また、勾配ブースティングを採用するかどうかなど、あてはめに関するオプションを設定します。

図4.4「モデルの設定」パネル

JMP Pro



JMP



検証法　モデルの検証に使う手法を選択します。詳細は、「検証法」（61ページ）を参照してください。

乱数シード値　将来に「ニューラル」プラットフォームを再実行したときに検証セットの無作為抽出などを再現したい場合は、ゼロ以外のシード値を指定してください。「乱数シード値」はデフォルトでゼロ、つまり同じ結果を再現しないように設定されています。分析をスクリプトに保存するとき、ここに入力した乱数シード値がスクリプトに保存されます。

隠れ層の構造　各隠れ層に含めるノードの種類とその個数を指定します。詳細は、「隠れ層の構造」（62ページ）を参照してください。

メモ：JMP の標準バージョンでは、活性化関数として TanH 関数しか使用できず、また、隠れ層が 1 つのニューラルネットワークしかあてはめられません。

JMP PRO **ブースティング** 勾配ブースティングのオプションを指定します。詳細は、「ブースティング」（63ページ）を参照してください。

JMP PRO **あてはめに関するオプション** 変数変換とモデルのあてはめに関するオプションを設定します。詳細は、「あてはめに関するオプション」（63ページ）を参照してください。

実行 ニューラルネットワークモデルをあてはめ、モデルのレポートを表示します。

[**実行**] をクリックしてモデルをあてはめた後も、「モデルの設定」パネルを開いて設定を変更すれば、別のモデルをあてはめることができます。

検証法

ニューラルネットワークは、とても柔軟なモデルであり、オーバーフィットしやすい傾向があります。オーバーフィットすると、推定に用いたデータでの予測は精確でも、将来のデータに対する予測精度は悪くなります。「ニューラル」プラットフォームでは、オーバーフィットを防ぐために次のような処理が行われます。

- モデルのパラメータにペナルティを課す
- 独立したデータセットを使ってモデルの予測能力を評価する

検証（validation）とは、データの一部をモデルパラメータの推定に使用し、残りのデータでモデルの予測能力を評価する方法を指します。

- モデルパラメータの推定に使うデータを、**学習**セットといいます。
- ペナルティの最適値を探したり、モデルの予測能力を評価したりするのに用いるデータを、**検証**セットといいます。
- 予測能力の最終評価に使うデータを、**テスト**セットといいます。検証列を指定した場合のみ、テストセットを使用できます。

学習セット、検証セット、テストセットは、1つのデータテーブルを分割して作成します。以下のいずれかの方法を選択してデータセットを分割します。

除外行の保留 行の属性によって、データを分割します。除外されていない行を学習セット、除外されている行を検証セットとして用います。

行の属性の使用と行の除外方法の詳細については、『JMP の使用法』の「データの入力と編集」章を参照してください。

保留 データを無作為に学習セットと検証セットに分割します。データのうち、検証セットとして使用する部分の割合（保留する割合）を指定することができます。

K分割 データをK個に分割します。順番に、(K-1) 個分のデータにモデルがあてはめられ、残っているデータでモデルが検証されます。全部でK個のモデルがあてはめられます。検証に用いた統計量が最良のものが、最終的なモデルとして選ばれます。

この方法は、少ないデータを効果的に利用するので、小規模なデータセットに適しています。

JMP PRO **検証列** 検証列の値に基づいて、データを分割します。検証列の指定は「ニューラル」起動ウィンドウで行います。図4.3を参照してください。

列の値によって、分割方法および検証方法が決まります。

- 検証列に一意な値が3つある場合は

  最小の値が学習セットを示します。

  中間の値が検証セットを示します。

  最大の値がテストセットを示します。
- 検証列に一意な値が2つある場合は、データが学習セットと検証セットに分かれます。
- 検証列に一意な値が4つ以上ある場合は、K分割検証法が行われます。

隠れ層の構造

メモ: JMPの標準バージョンでは、活性化関数としてTanH関数しか使用できず、また、隠れ層が1つのニューラルネットワークしかあてはめられません。

「ニューラル」プラットフォームでは、1層または2層のニューラルネットワークをあてはめることができます。第1層のノード数を増やすか、第2層を追加すると、ニューラルネットワークの柔軟性が増します。各層に含めるノードの数には制限がありません。第2層のノードはX変数の関数です。第1層のノードは第2層のノードの関数です。Y変数は第1層のノードの関数です。

隠れ層の各ノードで使われる関数を、「活性化関数」といいます。活性化関数は、X変数の線形結合を変換したものです。以下の活性化関数を使用できます。

TanH TanH関数（双曲正接関数）は、シグモイド関数です。TanH関数は、入力値が-1～1の範囲に変換されるように尺度化されたロジスティック関数です。TanH関数は次式で表されます。

$$\frac{e^{2x}-1}{e^{2x}+1}$$

xはX変数の線形結合です。

線形 恒等関数です。X変数の線形結合を、そのまま変換しないで用います。

この線形関数は、非線形関数と併用するのが普通です。線形関数を第2層の活性化関数とし、非線形関数を第1層の活性化関数として配置すると、X変数の次元を減らしてからY変数の非線形モデルを作成できます。

線形の活性化関数だけを使用すると、Y変数が連続尺度の場合には、Y変数に対するモデルは、X変数の線形結合で表されるので、通常の線形回帰モデルとなります。また、Y変数が順序尺度や名義尺度の場合には、通常のロジスティック回帰モデルとなります。

Gauss Gauss関数です。動径基底関数ネットワークにする場合、または応答曲面の形状がGauss分布（正規分布）の密度関数である場合にこのオプションを使用します。Gauss関数は次式で表されます。

e^{-x^2}

xはX変数の線形結合です。

ブースティング

JMP PRO ブースティングは、複数の小さいモデル（基底モデル）を逐次的にあてはめていき、それらの結果を合わせて、加法的な大きなモデルを構築する手法です。ブースティングでは、小さなモデルをあてはめて、その残差（尺度化した残差）が計算されます。その残差に対して、また、小さなモデルをあてはめます。この処理を繰り返します。最後に、小さなモデルを組み合わせて、最終的なモデルを作成します。検証セットによってモデルを評価して、小さいモデルをあてはめる回数を決めます（指定した回数を超えた場合は、そこで終了します）。

ブースティングは、1つの大きなモデルをあてはめるよりも、通常、計算時間が短くてすみます。ただし、あてはめる基底モデルを1～2個のノードから成る1層のモデルにしないと、あてはめるモデル数が多い場合には計算時間は短くなりません。

「モデルの設定」パネルの「ブースティング」で、あてはめる基底モデルの数と学習率を指定します。「モデルの設定」パネルの「隠れ層の構造」で基底モデルの構造を指定します。

学習率は$0 < r \leq 1$の範囲で設定します。学習率が1に近い値だと、最終モデルへの収束が速くなりますが、データにオーバーフィットしやすくなります。「モデル数」に少ない数を指定した場合には、学習率を1に近い値に設定してください。

ブースティングの仕組みを理解するために、1層・2ノードのニューラルネットを基底モデルとし、モデル数を8とした例を想定してみましょう。最初に、1層・2ノードのモデルをデータにあてはめます。このモデルの予測値を学習率によって尺度化し、実測値から引いて、尺度化した残差を求めます。次のステップで、先ほどの尺度化した残差に、再び、1層・2ノードのモデルをあてはめます。この手順を繰り返して8つのモデルをあてはめます。これ以上モデルを追加しても、検証セットの適合度が改善されないと判断された場合は、モデル数が8に満たなくても処理を終了します。こうして作成した一連の基底モデルを組み合わせ、最終的なモデルを作成します。もし、この例で、モデルを6回あてはめた時に処理が終了したとすると、最終的なモデルは1層、2 x 6 = 12ノードで構成されます。

あてはめに関するオプション

JMP PRO モデルのあてはめに関する以下のオプションを使用できます。

説明変数の変換　連続尺度の説明変数をすべて、Johnson Su分布またはJohnson Sb分布で変換し、正規分布に近づけます。この変数変換によって、外れ値や歪んだ分布が与える悪影響を抑えることが期待できます。

「モデルに関するオプション」（66ページ）の［変換した共変量の保存］オプションを参照してください。

ロバストなあてはめ　最小2乗ではなく、最小絶対偏差によってモデルの学習を行います。応答変数に外れ値がある場合に、このオプションは外れ値の影響を小さくします。連続尺度の応答にのみ使用できます。

ペナルティの手法 ペナルティ（罰則）の種類を選択します。ニューラルネットワークには、データにオーバーフィットする傾向があります。尤度にペナルティを課すことにより、その傾向を軽減できます。「ペナルティの手法」（64ページ）を参照してください。

ツアーの数 最適化計算を繰り返す回数を指定します。指定されたツアーの回数だけ、初期値を乱数で決めて、パラメータを推定します。そのなかで検証セットの適合度統計量が最良のものを最終的なモデルとして選びます。

ペナルティの手法

ペナルティは$\lambda p(\beta_i)$という式で表されます。λはペナルティパラメータ、p( )はパラメータ推定値の関数（ペナルティ関数）です。ペナルティパラメータの最適値は、検証セットで評価することにより推定されます。

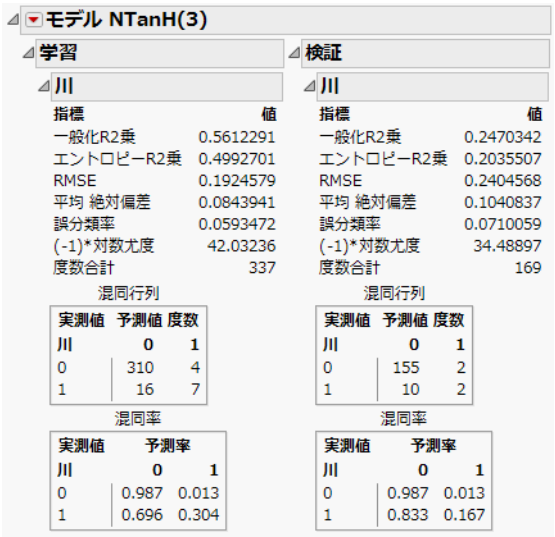
表4.1 ペナルティ手法の説明

方法	ペナルティ関数	説明
2乗	$\sum \beta_i^2$	X変数のほとんどが応答に影響していると考えられる場合に使用してください。
絶対	$\sum \lvert\beta_i\rvert$	X変数の数が多く、そのうちの少数だけが応答につよく影響していると考えられる場合には、この2つの手法のいずれかを使用してください。
重み減衰	$\sum \frac{\beta_i^2}{1+\beta_i^2}$	
ペナルティなし	なし	ペナルティを課しません。データが大規模で、処理時間を短くしたい場合には、このオプションを使用します。ただし、ペナルティを課した場合より、モデルの予測能力は低くなる可能性があります。

モデルのレポート

ニューラルネットワークをあてはめると、レポートが作成されます。学習セットと検証セットに対し、適合度指標が計算されます。また、応答が名義尺度または順序尺度の場合は、混同行列が表示されます。

図4.5 ニューラルモデルのレポートの例



学習と検証における適合度

学習セットと検証セットに対し、適合度指標が計算されます。図4.5を参照してください。

一般化R2乗　この指標は、一般的な回帰モデルに適用できるものです。一般化R2乗は、尤度Lから算出され、最大が1となるように尺度化されています。完全にモデルがデータにあてはまっている場合は1、切片だけのモデルと同等なあてはまりの場合には0になります。一般化R2乗は、通常のR2乗（正規分布に従う連続尺度の応答変数に対する標準最小2乗法のR2乗）を一般化したものです。この一般化R2乗は、「NagelkerkeのR^2」、または「Craig and UhlerのR^2」とも呼ばれており、Cox and Snellの疑似R^2を最大が1になるように尺度化したものです。詳細は、Nagelkerke（1991）を参照してください。

エントロピーR2乗　現在のモデルと、切片だけのモデルの対数尤度を比較します。応答が名義尺度または順序尺度の場合のみ表示されます。

R2乗　通常の決定係数です。

RMSE　誤差の平均平方の平方根。応答が名義尺度や順序尺度の場合は、誤差は(1-p)で計算されます。ここで、pは、実際に生じた応答水準に対する予測確率です。

平均 絶対偏差　誤差の絶対値の平均。応答が名義尺度や順序尺度の場合は、誤差は(1-p)で計算されます。ここで、pは、実際に生じた応答水準に対する予測確率です。

誤分類率　予測確率が最も大きい応答の水準が、観測された水準と一致しない割合。応答が名義尺度または順序尺度の場合のみ表示されます。

(-1)*対数尤度　対数尤度の符号を逆にしたもの。『基本的な回帰モデル』を参照してください。

SSE　誤差平方和。応答が連続尺度の場合のみ表示されます。

度数合計　使用されたオブザベーションの数。「ニューラル」起動ウィンドウで［度数］変数を指定した場合は、度数列の合計が「度数合計」になります。

応答が複数ある場合、応答ごとに適合度統計量が計算されます。その他に、「一般化R2乗」および「(-1)*対数尤度」は全体に対しても計算されます。

混同行列

応答が名義尺度または順序尺度の場合、混同行列と混同率も表示されます。図4.5を参照してください。混同行列は、応答水準の実測値と予測値を2元度数表にまとめたものです。名義尺度や順序尺度の応答においては、予測確率の最も高いカテゴリが予測値となります。混同率は、混同行列の度数を行合計で割ったものです。

モデルに関するオプション

モデルのレポートのタイトルバーにある赤い三角ボタンをクリックすると、レポートに出力を追加したり、計算結果を保存したりするオプションが表示されます。モデルのレポートの赤い三角ボタンをクリックして表示されるメニューには以下のオプションがあります。

ダイヤグラム　隠れ層の構造を示したダイヤグラムを表示します。

推定値の表示　レポートにパラメータ推定値を表示します。

プロファイル　予測プロファイルを起動します。名義尺度または順序尺度の場合は、予測プロファイルの各行がそれぞれ応答の水準を示します。赤い三角ボタンのメニューのオプションの詳細については、『プロファイル機能』の「プロファイル」章を参照してください。

カテゴリカルプロファイル　予測プロファイルを起動します。［プロファイル］とほぼ同じグラフですが、すべてのカテゴリの確率が1行のプロファイルにまとめられています。名義尺度または順序尺度の応答でのみ作成できます。赤い三角ボタンのメニューのオプションの詳細については、『プロファイル機能』の「プロファイル」章を参照してください。

等高線プロファイル　等高線プロファイルを起動します。モデルに連続尺度の因子が2つ以上含まれている場合に使用できます。赤い三角ボタンのメニューのオプションの詳細については、『プロファイル機能』の「等高線プロファイル」章を参照してください。

曲面プロファイル　曲面プロファイルを起動します。モデルに連続尺度の因子が2つ以上含まれている場合に使用できます。赤い三角ボタンのメニューのオプションの詳細については、『プロファイル機能』の「局面プロット」章を参照してください。

ROC曲線　ROC曲線を作成します。名義尺度または順序尺度の応答でのみ作成できます。ROC曲線の詳細は、「パーティション」章の「ROC曲線」（90ページ）を参照してください。

リフトチャート　リフトチャートを作成します。名義尺度または順序尺度の応答でのみ作成できます。リフトチャートの詳細は、「パーティション」章の「リフトチャート」（91ページ）を参照してください。

予測値と実測値のプロット　予測値と実測値のプロットを作成します。連続尺度の応答でのみ使用できます。

予測値と残差のプロット　予測値と残差のプロットを作成します。連続尺度の応答でのみ使用できます。

計算式の保存　データテーブルに新しい列を作成し、応答の予測値と隠れ層のノードの計算式を保存します。

プロファイル式の保存　データテーブルに新しい列を作成し、応答の予測値の計算式を保存します。計算式の中には、隠れ層のノードの計算式も組み込まれています。このオプションで保存される計算式は、Flash形式のプロファイルで使用できます。

高速計算式の保存　データテーブルに新しい列を作成し、応答の予測値の計算式を保存します。計算式の中には、隠れ層のノードの計算式も組み込まれています。このオプションで保存される計算式は、他のオプションのものに比べ、計算が高速です。ただし、Flash形式のプロファイルで使用することはできません。

JMP PRO **予測式を発行**　予測式を作成し、それらを「計算式デポ」プラットフォームの計算式列のスクリプトとして保存します。「計算式デポ」レポートが開いていない場合は、このオプションによって「計算式デポ」レポートが作成されます。「計算式デポ」章（167ページ）を参照してください。

SASデータステップの作成　新しいデータセットのスコアを計算するのに使えるSASコードを作成します。

検証の保存　学習セットと検証セットのいずれに各行が使われたかを示す新しい列を、データテーブルに作成します。「ニューラル」起動ウィンドウで検証列を指定した場合、このオプションは使用できません。「「ニューラル」起動ウィンドウ」（59ページ）を参照してください。

JMP PRO **変換した共変量の保存**　データテーブルに新しい列を作成し、変換した説明変数を保存します。列には、変換を示す計算式が含まれます。このオプションは、「モデルの設定」パネルで［説明変数の変換］オプションをオンにした場合のみ使用できます。「あてはめに関するオプション」（63ページ）を参照してください。

あてはめの削除　モデルのレポート全体を削除します。

ニューラルネットワークの例

この例では、「Boston Housing.jmp」データテーブルを使用します。地域ごとの住宅価格の中央値を、その地域の地理的特性によって予測するモデルを作成してみましょう。次の手順に従ってニューラルネットワークモデルを作成します。

1. ［**分析**］>［**予測モデル**］>［**ニューラル**］を選択して、「ニューラル」プラットフォームを起動します。
2. **「持ち家の価格」**を［**Y, 目的変数**］に指定します。
3. その他の列（**「犯罪率」**から**「低所得者」**まで）を［**X, 説明変数**］に指定します。
4. ［**OK**］をクリックします。
5. **「検証データ抽出確率」**に「0.2」と入力します。
6. **「乱数シード値」**に「1234」と入力します。

メモ: 検証セットが無作為に抽出されるため、一般に、結果は図とは異なります。上記のシード値を入力すると、この例で示す結果を再現できるようになります。

7. 第1層のノード数を「3」に設定します。
8. **「説明変数の変換」**オプションをオンにします。
9. ［**実行**］をクリックします。

図4.6のようなレポートが作成されます。

図4.6「ニューラル」レポート

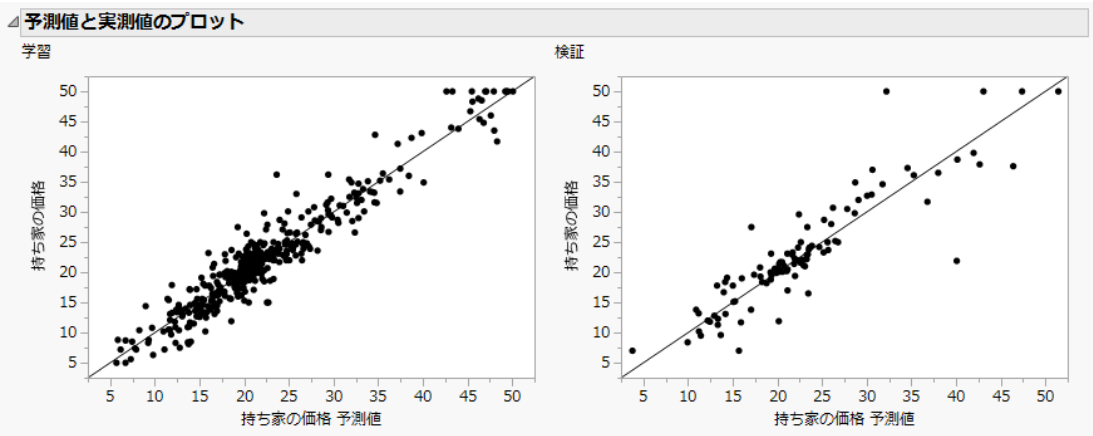
モデル NTanH(3)

学習		検証	
持ち家の価格		**持ち家の価格**	
指標	値	指標	値
R2乗	0.8990555	R2乗	0.8084408
RMSE	2.9354078	RMSE	3.9068845
平均 絶対偏差	2.1869167	平均 絶対偏差	2.7010518
(-1)*対数尤度	1008.2971	(-1)*対数尤度	283.73124
SSE	3481.114	SSE	1556.9022
度数合計	404	度数合計	102

学習セットと検証セットの両方に対し、結果が表示されます。検証セットの結果を見ると、得られたニューラルネットワークが将来のデータをどれぐらい予測できるかが分かります。

検証セットのR2乗は、0.819と高い値になっています。このことは、学習に使用されなかったデータについても、モデルの予測能力が高いことを示します。モデルのあてはめを示すもう一つの指標として、「モデル」の赤い三角ボタンのメニューから［**予測値と実測値のプロット**］を選択します。図4.7のようなプロットが作成されます。

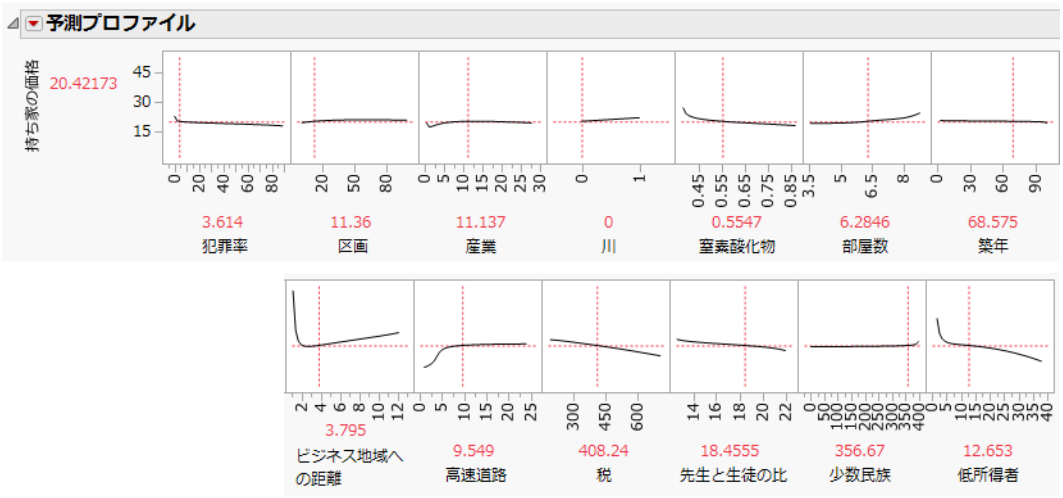
図4.7 予測値と実測値のプロット



点が回帰直線上に並んでいることから、予測値が実測値に近いことがわかります。

X変数が予測値に与える影響を把握するため、「モデル」の赤い三角ボタンのメニューから［**プロファイル**］を選択します。図4.8のようなプロファイルが表示されます。

図4.8 プロファイル



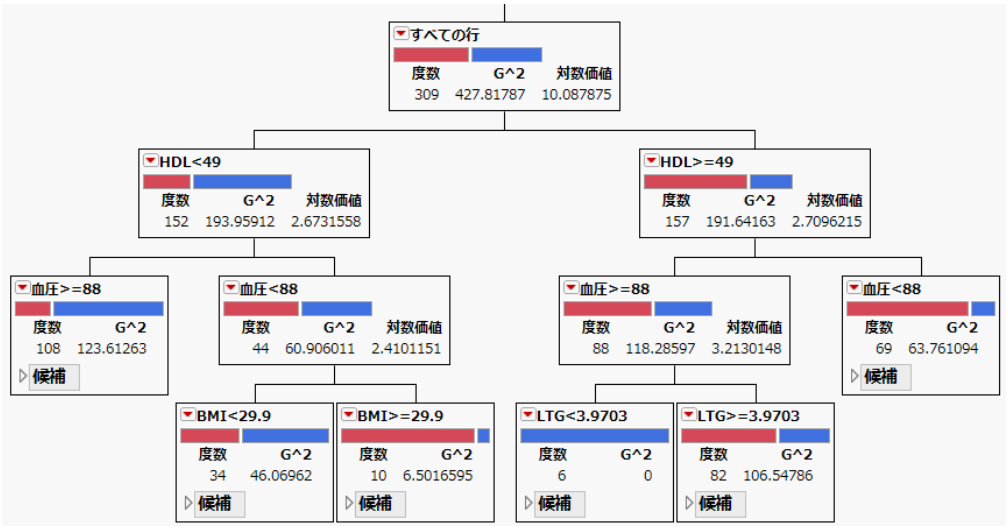
変数の中には、プロファイルの傾きが正のものと負のものがあります。たとえば、**「部屋数」**は傾きが正です。地域において部屋数が多い家が多いほど、住宅価格の中央値は高くなることを示します。「先生と生徒の比」は、町ごとに求められた、先生1人あたりの生徒数です。この変数の傾きは負です。これは、先生1人あたりの生徒数が多いほど、その町の住宅価格が低くなる傾向があることを意味します。

第5章

パーティション
ディシジョンツリーによるモデル化

「パーティション」プラットフォームは、説明変数と目的変数の関係に従ってデータを再帰的に分割し、ディシジョンツリー（決定木）を作成します。そのアルゴリズムは、最も効果的に応答を予測するような分岐を、説明変数の可能なすべての分岐を検索して探し出します。データの分岐（**パーティション**）は繰り返され、分割のルールを示すディシジョンツリーが最終的に形成されます。分岐は、適合度が適度になるまで続けられます。このアルゴリズムでは、多数の可能な分岐から最適なものを選び出します。この手法は、モデル化や発見を行うのに非常に役立ちます。

図5.1 ディシジョンツリーの例



「パーティション」プラットフォームの概要

「パーティション」プラットフォームは、説明変数と目的変数の関係に従ってデータを再帰的に分割し、ディシジョンツリーを作成します。パーティションにはいろいろなバリエーションがあり、ディシジョンツリー（決定木）、CART™、CHAID™、C4.5、C5などの名前で呼ばれています。パーティションはよく、以下のような理由からデータマイニング手法とみなされています。

- 事前にモデルを用意しなくても変数の関係が検討できる
- 膨大なデータを容易に処理することができる
- 結果が解釈しやすい

パーティションのよく知られた利用としては、病気を診断する上での発見的モデルを作成することです。多数の患者に対する症状と診断結果が与えられた場合には、パーティションを使って新しい患者の診断に役立つ階層的な質問を生成できます。

説明変数にも、連続尺度とカテゴリカルの両方を使用できます。**説明変数**が連続尺度の場合は、分岐値に基づいて分岐が行われ、分岐値を境として上と下に標本が分かれます。説明変数がカテゴリカルの場合は、標本が2つの水準グループに分かれます。

また、目的変数は、連続尺度とカテゴリカル（名義／順序尺度）のどちらでもかまいません。目的変数が連続尺度の場合は、応答値の平均があてはめられます。目的変数がカテゴリカルな場合は、あてはめた値が目的変数の水準の確率になります。どちらの場合も、データは2つのグループの応答の差が最大になるように分岐します。

分岐基準の詳細は、「統計的詳細」（103ページ）を参照してください。

対話的パーティショニングの詳細については、Hawkins, D.M., and Kass, G.V.(1982) and Kass, G.B (1980)を参照してください。

「パーティション」プラットフォームの例

この例では、「パーティション」プラットフォームを使用して、糖尿病患者の1年間にわたる症状進行（Lowまたは High）を予測するディシジョンツリーを作成します。

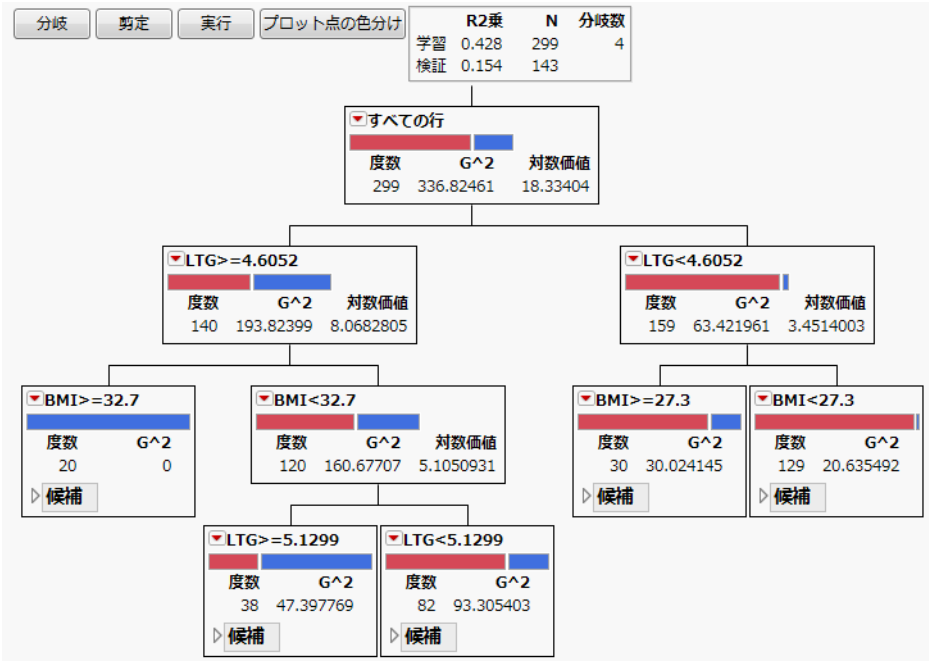
1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Diabetes.jmp」を開きます。
2. ［**分析**］>［**予測モデル**］>［**パーティション**］を選択します。
3. 「**Y 2値**」を選択し、［**Y, 目的変数**］をクリックします。
4. 「**年齢**」から「**グルコース**」までを選択し、［**X, 説明変数**］をクリックします。
5. 「**検証データの割合**」に「0.33」と入力します。

メモ：JMP Pro では、分析対象のデータテーブルにて検証セットを示す列がある場合、その列を指定することもできます。この例では、**「検証」**列を選択し、**［検証］**ボタンをクリックしてください。その場合、**「検証データの割合」**は「0」に設定してください。

6. **［OK］**をクリックします。
7. プラットフォームのレポートウィンドウで**［実行］**をクリックし、自動分岐を実行します。

メモ：学習セットと検証セットがランダムに決められるため、実際の結果は図 5.2 とは異なります。

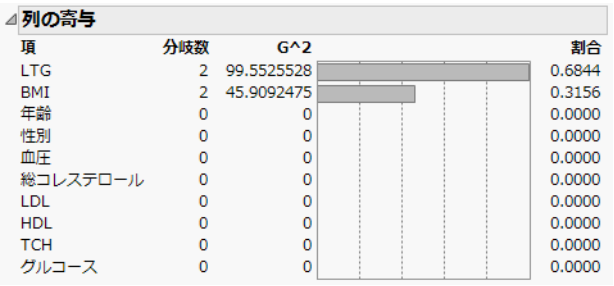
図5.2 糖尿病の「パーティション」レポート



今回の実行では、分岐数が 4 となり、「検証」セットの最終的な「R2 乗」は 0.154 となりました。このディシジョンツリーには、4 つの分岐と、各分岐におけるオブザベーション数が示されています。

8. 「Y 2 値のパーティション」の横にある赤い三角ボタンをクリックし、**［列の寄与］**をクリックします。

図5.3「列の寄与」レポート



列の寄与

項	分岐数	G^2	割合
LTG	2	99.5525528	0.6844
BMI	2	45.9092475	0.3156
年齢	0	0	0.0000
性別	0	0	0.0000
血圧	0	0	0.0000
総コレステロール	0	0	0.0000
LDL	0	0	0.0000
HDL	0	0	0.0000
TCH	0	0	0.0000
グルコース	0	0	0.0000

この「列の寄与」レポートには、ディシジョンツリーモデルの説明変数として**「LTG」**と**「BMI」**だけが示されています。これらの2つ列は、それぞれ2つの分岐に使用されています。なお、実際の分析結果は、ここでの結果と異なる可能性があります。なぜなら、「検証データの割合」を使用すると、検証セットは分析対象のデータテーブルから無作為に選択されます。分析を実行するたびに、新しい検証セットが無作為に選択さるため、それぞれの結果は異なったものになります。

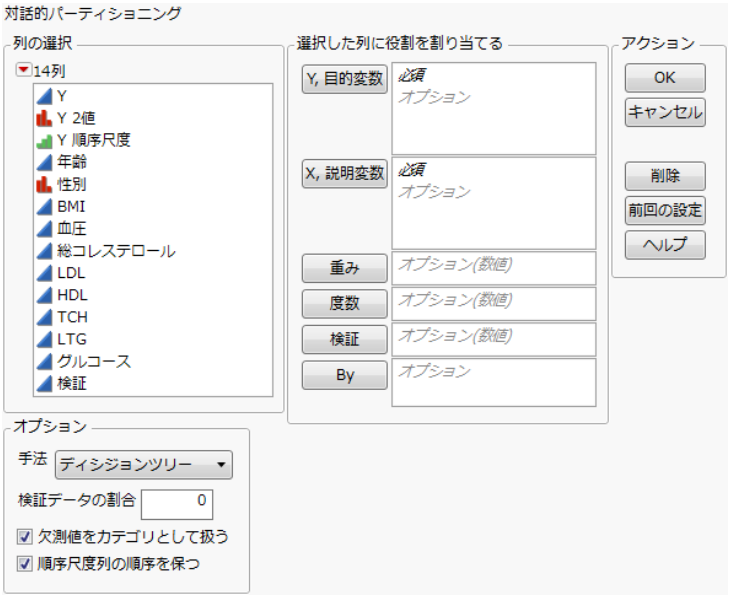
9. 「Y 2値のパーティション」の横にある赤い三角ボタンをクリックし、**［列の保存］>［予測式の保存］**を選択します。

「Diabetes.jmp」データテーブルに、**「確率(Y 2値==Low)」**、**「確率(Y 2値==High)」**、および**「最尤 Y 2値」**という列が追加されます。これらの応答確率の計算方法を確認するには、「列」パネルで各列の横にある計算式アイコン をダブルクリックします。

「パーティション」プラットフォームの起動

「パーティション」プラットフォームを起動するには、[分析]>[予測モデル]>[パーティション]を選択します。

図5.4「パーティション」起動ウィンドウ



Y, 目的変数　分析する目的変数です。

X, 説明変数　説明変数です。

重み　分析において各行の重みとして使用される数値を含む列です。

度数　分析において各行の度数として使用される数値を含む列です。

JMP PRO **検証**　多くとも3つの数値を含む数値列。「検証」（93ページ）を参照してください。

By　別々に分析を行いたいときに、そのグループ分けをする変数を指定します。指定された列の各水準ごとに、別々に分析が行われます。各水準の結果は別々のレポートに表示されます。複数のBy変数を割り当てた場合、それらのBy変数の水準の組み合わせごとに別々のレポートが作成されます。

JMP PRO **手法**　パーティションの手法としてディシジョンツリー、ブートストラップ森、ブースティングツリー、K近傍法、単純Bayesを選択できます。

「ディシジョンツリー」以外の手法については、第6章「ブートストラップ森」、第7章「ブースティングツリー」、第8章「K近傍法」、および第9章「単純Bayes」を参照してください。

検証データの割合　データ全体のうち検証に用いるデータの割合です。「検証」（93ページ）を参照してください。

欠測値をカテゴリとして扱う　説明変数がカテゴリカルな場合、このチェックボックスをオンにすると、分析において、欠測値が1つのカテゴリとして扱われます。説明変数が連続尺度の場合は、欠測値が同一の数値を持つものとして扱われます。「欠測値をカテゴリとして扱う」（88ページ）を参照してください。

順序尺度列の順序を保つ　このチェックボックスをオンにすると、順序尺度の列において、順序を保つ分岐だけが考慮されるようになります。

「パーティション」レポート

初期状態の「パーティション」レポートには、パーティションの状態を示すグラフ、分岐などを行うコントロールボタン、要約統計量、およびディシジョンツリー（決定木）が表示されます。プロットやツリーは、初期状態では分岐をまったく行っていない状態になっています。レポートの詳細は、カテゴリカルな応答と連続尺度の応答とで異なります。

コントロールボタン

コントロールボタンを使用して、ディシジョンツリーとの対話的な操作を行います。

分岐　最適な分岐を使用してデータのパーティションを作成します。複数の分岐を指定するには、Shiftキーを押したまま［**分岐**］をクリックします。

剪定　最も新しい分岐を削除します。

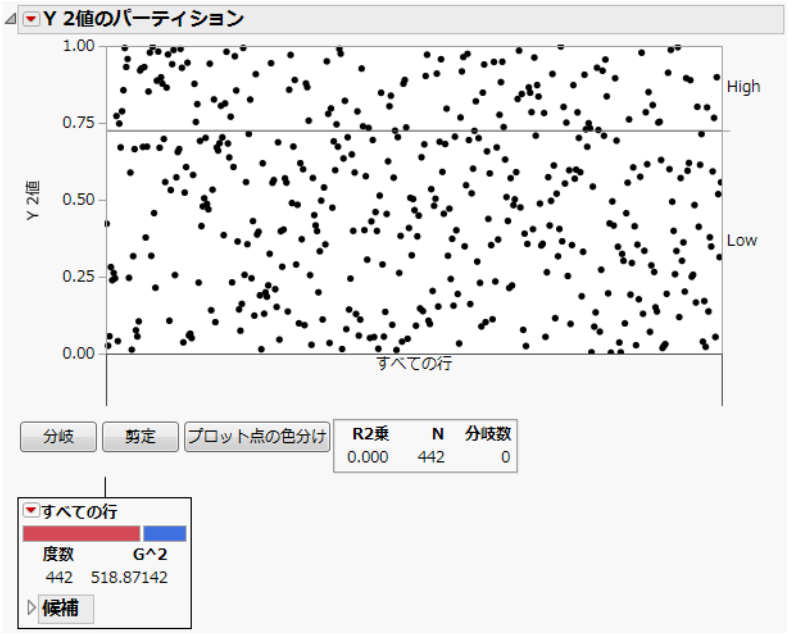
実行　（検証セットを用いている場合に使用可能。）検証セットにおける適合度統計量が最適になるまで、分岐を自動的に行います。「検証」（93ページ）を参照してください。なお、検証セットを用いていない場合は、どれだけの分岐を行うかは分析者自身が決定しなければいけません。

プロット点の色分け　カテゴリカルな応答の場合は、応答水準に応じてオブザベーションに色がつけられます。これらの色はデータテーブルに追加されます。

カテゴリカルな応答のオプション

次図は、「Diabetes.jmp」サンプルデータテーブルのカテゴリカルな応答［**Y 2値**］に対するプロットです。

図5.5 カテゴリカルな応答の「パーティション」レポート



パーティションのグラフ

「パーティション」のプロットにおいて、各点はそれぞれデータの各行を表しています。なお、検証セットを用いた場合、学習セットだけがプロットされます。初期状態のプロットでは、分岐は行われていません。

次の点に注意してください。

- 左側の縦軸は、各応答の割合を示しています。
- 右側の縦軸は、応答値を示しています。
- プロット中の横線は、各分割ごとの応答水準の割合を示しています。初期状態での横線は、分岐が1回も行われていないので、データ全体における応答水準の割合を示しています。
- どのような分岐が行われたかは、X軸の下に、テキストによる説明によって示されます。また、プロット内においては、各データ点が縦線により分けられます。これらの縦線によって分けられた領域は、ツリーの各ノードに対応しています。X軸のテキストにおいて、最も新しい分岐は最上に（つまり、X軸のすぐ下に）表示されます。プロットは、分岐や剪定のたびに更新されます。

要約レポート

図5.6 カテゴリカルな応答の要約レポート

	R2乗	N	分岐数
学習	0.428	299	4
検証	0.154	143	

要約レポートには、学習データの適合度統計量が表示されます（検証データやテストデータを使用した場合には、それらから計算された適合度統計量も表示されます）。要約レポートの適合度統計量は、分岐や剪定のたびに更新されます。

R2 乗　現在の R^2 の値。

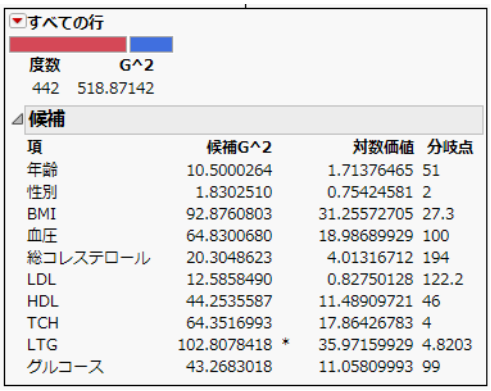
N　オブザベーションの数。

分岐数　ディシジョンツリー内の現在の分岐数。

各ノードに関するレポート

ツリー内の各ノードには、それらのノードに関する情報と、赤い三角ボタン（これをクリックすると、追加のオプションを選べます）があります。また、終端ノードには、「候補」レポートも表示されます。

図5.7 カテゴリカルな応答の終端ノードレポート


すべての行

度数	G^2
442	518.87142

候補

項	候補G^2		対数価値	分岐点
年齢	10.5000264		1.71376465	51
性別	1.8302510		0.75424581	2
BMI	92.8760803		31.25572705	27.3
血圧	64.8300680		18.98689929	100
総コレステロール	20.3048623		4.01316712	194
LDL	12.5858490		0.82750128	122.2
HDL	44.2535587		11.48909721	46
TCH	64.3516993		17.86426783	4
LTG	102.8078418	*	35.97159929	4.8203
グルコース	43.2683018		11.05809993	99

度数　そのノードに属する学習データのオブザベーション数。

G^2（連続尺度の応答に使用される平方和ではなく）カテゴリカルな応答に使用される適合度統計量。値が小さいほど、適合度が良いことを示します。「統計的詳細」（103ページ）を参照してください。

候補　各列の「候補」レポートに、その列の最適な分岐に関する詳細が示されます。すべての列の分岐のなかで最適なものにアスタリスク（*）が付いています。

項　候補列を表示します。

候補 G^2　最適な分岐点の尤度比カイ 2 乗。尤度比カイ 2 乗値が最大である説明変数で分岐すると、モデルの対数尤度が、分岐の前後で最も大きく増加します。

LogWorth　$-\log_{10}(p\text{ 値})$ で定義される、対数価値統計量。対数価値が最大となる分岐点が、最適な分岐点とみなされます。詳細については、「統計的詳細」（103 ページ）を参照してください。

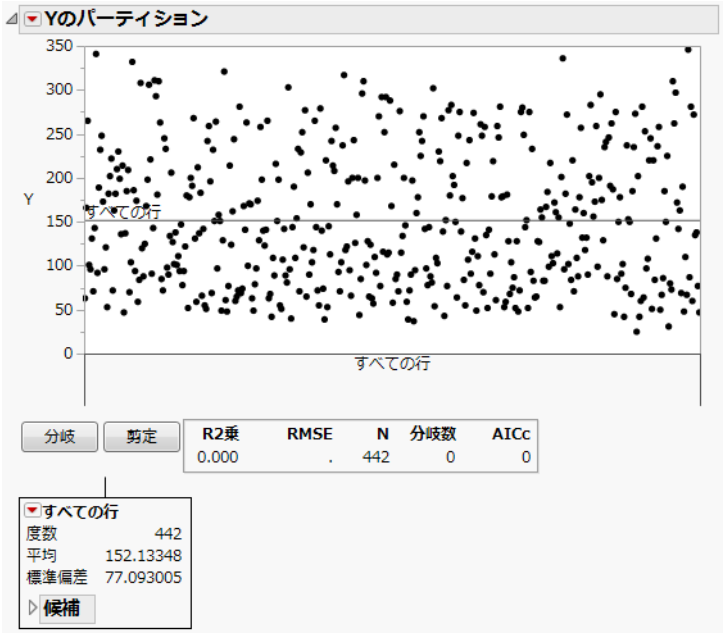
分岐点　分岐を決定する説明変数の値。カテゴリカルな説明変数については、左側に分岐されるカテゴリが表示されます。

最適な分岐点にはアスタリスク（*）がついています。しかし、「候補G^2」が大きい変数と「対数価値」が大きい変数が同じでない場合もあります。検定統計量が最大となる分岐には「<」、対数価値が最大となる分岐には「>」を表示することによって、最大値を持つ変数を別々に示しています。アスタリスクがついている場合は、検定統計量が最大の変数と、対数価値が最大の変数が一致しているケースです。詳細は、「統計的詳細」（103ページ）の節を参照してください。

連続尺度の応答のレポート

次図は、「Diabetes.jmp」サンプルデータテーブルの連続尺度の応答「Y」に対するプロットです。

図5.8 連続尺度の応答の「パーティション」レポート



パーティションのグラフ

「パーティション」のプロットは、初期状態は分岐が何も行われていない状態のものです。各点はそれぞれデータの各行を表しています。なお、検証セットを用いた場合、プロットには学習セットの行だけが描かれます。

次の点に注意してください。

- 縦軸は、オブザベーションの応答値を表しています。
- 横線は、各ノードにおける応答の平均値を表しています。初期状態では、横線は、応答の全体平均を表しています。
- 縦線で分割されている領域は、ディシジョンツリーの分岐を表しています。これらの縦線によって分けられた領域は、ツリーの各ノードに対応しています。X軸のテキストにおいて、最も新しい分岐は最上に（つまり、X軸のすぐ下に）表示されます。プロットは、分岐や剪定のたびに更新されます。

ヒント：狭い領域において、分岐の情報を見るには、横軸のラベルの上にカーソルを置いてください。ツールヒントが表示されます。

要約レポート

図5.9 連続尺度の応答の要約レポート

	R2乗	RMSE	N	分岐数	AICc
学習	0.490	54.690663	299	4	3253.83
検証	0.366	61.16064	143		

要約レポートには、学習データの適合度統計量が表示されます（検証データやテストデータを使用した場合には、それらから計算された適合度統計量も表示されます）。要約レポートの適合度統計量は、分岐や剪定のたびに更新されます。

R2乗　現在のR^2の値。

RMSE　誤差の標準偏差。

N　オブザベーションの数。

分岐数　ディシジョンツリー内の現在の分岐数。

AICc　修正済みの赤池の情報量規準。詳細については、『基本的な回帰モデル』の付録「統計的詳細」を参照してください。

各ノードに関するレポート

ツリー内の各ノードには、それらのノードに関する情報と、赤い三角ボタン（これをクリックすると、追加のオプションを選べます）があります。また、終端ノードには、「候補」レポートも表示されます。

図5.10 連続尺度の応答の終端ノードレポート

分岐　剪定

R2乗	RMSE	N	分岐数	AICc
0.000	.	442	0	0

すべての行

度数	442
平均	152.13348
標準偏差	77.093005

候補

項	候補SS		対数価値	分岐点
年齢	101593.5850		3.61765555	51
性別	4860.2308		0.40460420	1
BMI	729618.2987		46.76510648	27.3
血圧	446708.6991		23.89250262	102
総コレステロール	157877.7151		6.39470855	194
LDL	120014.5872		4.49760240	126.6
HDL	390514.6338		20.07867675	46
TCH	470204.7358		25.55525064	3.73
LTG	764133.3264	*	50.04935713	4.6052
グルコース	341244.3856		16.90517792	100

度数 そのノードに属する学習データのオブザベーション数。

平均 そのノードに属する学習データの、応答の平均値。

標準偏差 そのノードに属する学習データの、応答の標準偏差。

候補 各列の「候補」レポートに、その列の最適な分岐に関する詳細が示されます。すべての列の分岐のなかで最適なものにアスタリスク（*）が付いています。

項 候補列を表示します。

候補 SS 最適な分岐点の平方和。

LogWorth $-\log_{10}(p\ 値)$ で定義される、対数価値統計量。対数価値が最大となる分岐点が、最適な分岐点とみなされます。詳細については、「統計的詳細」（103 ページ）を参照してください。

分岐点 分岐を決定する説明変数の値。カテゴリカルな説明変数については、左側に分岐されるカテゴリが表示されます。

最適な分岐点にはアスタリスク（*）がついています。しかし、「候補 SS」が大きい変数と「対数価値」が大きい変数が同じでない場合もあります。検定統計量が最大となる分岐には「<」、対数価値が最大となる分岐には「>」を表示することによって、最大値を持つ変数を別々に示しています。アスタリスクがついている場合は、検定統計量が最大の変数と、対数価値が最大の変数が一致しているケースです。詳細は、「統計的詳細」（103 ページ）の節を参照してください。

「パーティション」プラットフォームのオプション

「パーティション」の赤い三角ボタンのメニューには、必要に応じてレポートをカスタマイズするためのオプションがあります。使用可能なオプションは、分析に使用するデータのタイプによって決まります。

表示オプション レポート要素の表示と非表示を切り替えるオプションがあります。

点の表示 点を表示します。カテゴリカルな応答の場合、点を表示するか、または色のついた矩形領域を表示するかを切り替えます。

ツリーの表示 パーティションツリー（大きいツリー）を表示します。

グラフの表示 データをプロットしたグラフを表示します。

カラーバーの表示 （カテゴリカルな応答のみ）各ノードにおいて、分岐の割合を示す色つきのバーを表示します。

分岐統計量の表示 分岐統計量を表示します。カテゴリカルの分岐統計量である G^2 については、「統計的詳細」（103 ページ）を参照してください。

割合を表示 （カテゴリカルな応答のみ）各ノードにおいて、割合と確率を表示します。

なお、[**度数を表示**] を選択すると、度数とともに、割合と確率も一緒に表示されます。「割合」と「確率」の詳細については、「統計的詳細」（103 ページ）を参照してください。

度数を表示 （カテゴリカルな応答のみ）各ノードにおいて、度数を表示します。このオプションを選択すると、[**割合を表示**]が自動的に選択されます。[**割合を表示**]の選択を解除すると、度数も表示されなくなります。

分岐の候補を表示 「候補」レポートを表示します。

分岐の候補を並べ替え 「候補」レポートにおける列を、統計量と対数価値のいずれかに基づいて並べ替えます。

最良分岐 ツリーを最適な分岐点で分岐します。これは[**分岐**]ボタンをクリックするのと同じことです。

最悪分岐を剪定 応答変数を識別する能力が最も弱い分岐が削除されます。これは[**剪定**]ボタンをクリックするのと同じです。

分岐の最小サイズ 度数または割合を入力して、分岐の最小許容サイズを指定します。度数（標本サイズ）として指定する場合は、1以上の値を入力してください。全体に対する割合として指定する場合は、1未満の値を入力してください。デフォルト値は、5と、行数を10,000で割った値以下の最大の整数のうち、いずれか大きいほうの値に設定されます。

列のロック 分岐の対象とならないように、列をロックします。なお、列はロックされているか否かとは無関係に、その表示と非表示を切り替えることができます。

予測値と実測値のプロット （連続尺度の応答のみ）予測値と実測値のプロットを表示します。「予測値と実測値のプロット」（89ページ）を参照してください。

小さいツリー表示 プロットの右側に小さなツリーを表示します。

三次元ツリー 三次元に描かれたツリーを表示します。このオプションを表示するには、Shiftキーを押しながら赤い三角ボタンをクリックします。

葉のレポート 最下位のノードについて、それぞれの平均値と度数を表示します。

列の寄与 各列があてはめにどれだけ寄与したかを示すレポートを表示します。レポートには、その列が分岐した回数や、その列に起因するG^2の合計、または平方和の合計が表示されます。

分岐履歴 分岐の回数に対するR2乗のプロットを表示します。検証セットを利用した場合（除外した行で検証を行う場合、保留によって検証を行う場合、または、検証列を使用した場合）は、学習セットと検証セットの両方に関して、R2乗値の曲線が描かれます。曲線は、学習セットは青色、検証セットは赤色で描かれます。また、[K分割交差検証]を選択した場合は、すべてのデータの曲線が青色で、交差検証のR2乗曲線が緑色で表示されます。

K分割交差検証 学習セットと分割セットの適合度統計量を示した「交差検証法」レポートを表示します。検証については、「K分割交差検証」（94ページ）を参照してください。

ROC曲線 （カテゴリカルな応答のみ）ROC（Receiver Operating Characteristic: 受診者動作特性）曲線を描きます。ROC曲線は、応答水準を予測確率で並べ替えて、モデルの予測精度を見るものです。「ROC曲線」（90ページ）を参照してください。

リフトチャート （カテゴリカルな応答のみ）リフトチャートを描きます。リフトチャートは、モデルの予測精度を見るものです。「リフトチャート」（91ページ）を参照してください。

あてはめの詳細の表示 （カテゴリカルな応答のみ）「あてはめの詳細」レポートには、あてはめに関するいくつかの指標と「混同行列」レポートが表示されます。「「あてはめの詳細」の表示」（84ページ）を参照してください。

列の保存 モデルやツリーの結果を保存するオプション、およびSASコードを作成するオプションがあります。

残差の保存 モデルの残差をデータテーブルに保存します。

予測値の保存 モデルの予測値をデータテーブルに保存します。

葉の番号を保存 ツリーの葉の番号をデータテーブルに保存します。

葉のラベルを保存 ツリーの葉のラベルをデータテーブルに保存します。ラベルを見ると、ツリーのどの枝をたどればその行が見つかるかがわかります。枝は「&」で区切られています。たとえば、「サイズ (小型 , 中型)& サイズ (小型)」のようになります。ただし、カテゴリを示すラベルが繰り返しになるときは、省略されます。葉のラベルがツリーの上位にあるノードのカテゴリを含む場合、繰り返しになる部分が「^」で置き換えられます。たとえば「サイズ (小型 , 中型)& サイズ (小型)」は、「^& サイズ (小型)」と表示されます。

予測式の保存 予測式をデータテーブルの列に保存します。予測式は、条件節が枝分かれした形で、ツリーの構造を反映しています。応答変数が連続尺度の場合、予測式の列には、［予測対象］プロパティが割り当てられます。カテゴリカルな場合は、［応答確率］プロパティが割り当てられます。

欠測処理予測式の保存 データに欠測値があり、かつ［欠測値をカテゴリとして扱う］チェックボックスがオフの場合に、欠測値をランダムに処理する予測式を保存します。この予測式では、説明変数が欠測値の場合、どちらに分岐するかがランダムに決められます。応答変数が連続尺度の場合、予測式の列には、［予測対象］プロパティが割り当てられます。カテゴリカルな場合は、［応答確率］プロパティが割り当てられます。［欠測値をカテゴリとして扱う］チェックボックスをオンにした場合は、Shift キーを押しながらレポートの赤い三角ボタンをクリックすると、［欠測処理予測式の保存］を選択できます。

葉の番号の式を保存 葉の番号を計算する計算式をデータテーブルの列に保存します。

葉のラベルの式を保存 葉のラベルを計算する計算式をデータテーブルの列に保存します。

SAS DATA ステップの作成 データセットのスコア計算に使用できる SAS コードを作成します。

JMP PRO **予測式を発行** 予測式を作成し、それを「計算式デポ」プラットフォームの計算式列のスクリプトとして保存します。「計算式デポ」レポートが開いていない場合は、このオプションによって「計算式デポ」レポートが作成されます。「計算式デポ」章（167ページ）を参照してください。

JMP PRO 欠測処理予測式を発行 欠測処理予測式を作成し、それを「計算式デポ」プラットフォームの計算式列のスクリプトとして保存します。「計算式デポ」レポートが開いていない場合は、このオプションによって「計算式デポ」レポートが作成されます。「計算式デポ」章（167 ページ）を参照してください。［欠測値をカテゴリとして扱う］チェックボックスをオンにした場合は、Shift キーを押しながらレポートの赤い三角ボタンをクリックすると、このオプションを使用できます。

利益行列の指定 （応答がカテゴリカルな場合のみ）分類の判定が正しいときや正しくないときの利益や損失を指定できます。名義尺度の応答の場合は、確率の閾値を使用して利益行列の各要素を指定できます。「「あてはめの詳細」の表示」（84 ページ）を参照してください。

プロファイル 対話式のプロファイルを表示します。因子の値を変更すると、それに応じた応答の予測値の曲線が描かれます。『プロファイル機能』の「プロファイル」章を参照してください。

プロット点の色分け （カテゴリカルな応答のみ）応答のグループごとに、点に色をつけます。これは［プロット点の色分け］ボタンをクリックするのと同じことです。「カテゴリカルな応答のオプション」（76 ページ）を参照してください。

以下のオプションについて詳しくは、『JMP の使用法』の「JMP のレポート」章を参照してください。

やり直し 分析を繰り返すオプションや、やり直すオプションを含みます。また、［自動再計算］オプションを選択すると、このオプションに対応しているプラットフォームにおいて、データテーブルに加えた変更が、該当するレポートに即座に反映されるようになります。

スクリプトの保存 レポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。

By グループのスクリプトを保存 By 変数の全水準に対するレポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。起動ウィンドウで By 変数を指定した場合のみ使用可能です。

「あてはめの詳細」の表示

図 5.11 カテゴリカルな応答の「あてはめの詳細」（「Diabetes.jmp」の「**Y 2 値**」）

あてはめの詳細

指標	学習	検証	定義
エントロピーR2乗	0.4276	0.1543	1-Loglike(model)/Loglike(0)
一般化R2乗	0.5656	0.2464	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
平均 -Log p	0.3224	0.5312	Σ -Log(ρ[j])/n
RMSE	0.3201	0.4080	√ Σ(y[j]-ρ[j])²/n
平均 絶対偏差	0.2083	0.2917	Σ |y[j]-ρ[j]|/n
誤分類率	0.1371	0.2238	Σ (ρ[j]≠ρMax)/n
N	299	143	n

混同行列

学習

実測値 Y 2値	予測値 Low	予測値 High
Low	212	12
High	29	46

検証

実測値 Y 2値	予測値 Low	予測値 High
Low	84	13
High	19	27

エントロピー R2 乗 現在のモデルと、切片だけのモデルの対数尤度を比較します。値が 1 に近いほど、適合度が良いことを示します。

一般化 R2 乗 この指標は、一般的な回帰モデルに適用できるものです。一般化 R2 乗は、尤度 L から算出され、最大が 1 となるように尺度化されています。完全にモデルがデータにあてはまっている場合は 1、切片だけのモデルと同等なあてはまりの場合には 0 になります。一般化 R2 乗は、通常の R2 乗（正規分布に従う連続尺度の応答変数に対する標準最小 2 乗法の R2 乗）を一般化したものです。この一般化 R2 乗は、「Nagelkerke の R^2」、または「Craig and Uhler の R^2」とも呼ばれており、Cox and Snell の疑似 R^2 を最大が 1 になるように尺度化したものです。詳細は、Nagelkerke（1991）を参照してください。値が 1 に近いほど、適合度が良いことを示します。

平均 -Log p -log(p) の平均です。p は、実際に生じた応答水準に対する予測確率です。値が小さいほど、適合度が良いことを示します。

RMSE 誤差の標準偏差。応答と p（実際に発生したイベントの確率の予測値）の差を示します。値が小さいほど、適合度が良いことを示します。

平均 絶対偏差 応答と p（実際に発生したイベントの確率の予測値）の差の絶対値の平均。値が小さいほど、適合度が良いことを示します。

誤分類率 予測確率が最も大きい応答の水準が、観測された水準と一致しない割合。値が小さいほど、適合度が良いことを示します。

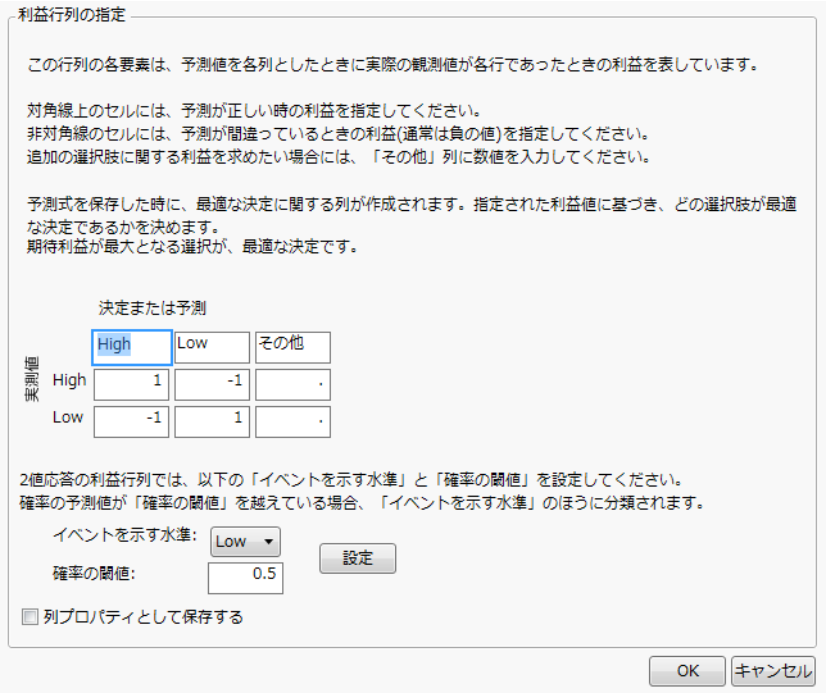
「混同行列」レポートには、学習セットの混同行列、および、検証セットとテストセットが使われている場合には、それらの混同行列が表示されます。「混同行列」は、応答の実測値と予測値を2元度数表にまとめたものです。

応答に「利益行列」列プロパティがある場合、または［利益行列の指定］オプションを使用して損失を指定した場合は、「決定行列」レポートが表示されます。「「決定行列」レポート」（87ページ）を参照してください。

利益行列の指定

カテゴリカルな応答には、利益行列（profit matrix）を使用できます。利益行列では、望ましくない結果に損失（cost）を、望ましい結果には利益（profit）を割り当てます。

図5.12「利益行列の指定」ウィンドウ



利益行列では、カテゴリカルな応答変数の実測値と予測値の組み合わせのそれぞれに対して、利益や損失を入力します。データには存在しないカテゴリを選んだ場合の損失を指定するには、「その他」列に値を入力します。指定した利益行列を列プロパティとして応答列に保存するには、**[列プロパティとして保存する]** チェックボックスをオンにします。このチェックボックスをオフにしておくと、利益行列は現在の「パーティション」レポートだけに適用されます。

利益行列の確率の閾値指定

応答がバイナリのときは、利益行列に重みを入力する代わりに、「利益行列の指定」ウィンドウで確率の閾値を指定できます。値が利益行列に対してどのように計算されるかについては、『JMP の使用法』の「列情報ウィンドウ」章を参照してください。

イベントを示す水準　確率をモデル化するときに、興味があるほうの水準を指定します。

確率の閾値　「イベントを示す水準」が生じる確率に対する閾値を指定します。「イベントを示す水準」が生じる確率が、この閾値を超える場合、そのオブザベーションは「イベント」として分類されます。

[利益行列の指定] オプションを使用して損失を定義した後、[あてはめの詳細の表示] を選択すると、「決定行列」レポートが表示されます。「「決定行列」レポート」（87ページ）を参照してください。

利益行列を指定し、モデルの予測式を保存すると、データテーブルに保存された計算式列には以下のものが含まれます。

- **<水準>の利益**: データの各行、および応答の各水準ごとに、その水準に分類した場合に生じる利益の期待値が表示されます。
- **利益が最大となる<列名>の予測値**: データの各行ごとに、期待利益が最大となる応答の水準が表示されます。
- **<列名>の期待利益**: データの各行ごとに、「利益が最大となる<列名>の予測値」列の値に分類したときの利益の期待値が表示されます。
- **<列名>の実測利益**: データの各行ごとに、「利益が最大となる<列名>の予測値」列の値に分類したときに、実際には現在のデータ値であった場合の利益が表示されます。

「決定行列」レポート

図5.13「決定行列」レポートを含む「あてはめの詳細」レポート

あてはめの詳細

指標	学習	検証	定義
エントロピーR2乗	0.3788	0.0367	1-Loglike(model)/Loglike(0)
一般化R2乗	0.5382	0.0647	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
平均 -Log p	0.4134	0.6402	Σ -Log(ρ[j])/n
RMSE	0.3649	0.4694	√ Σ(y[j]-ρ[j])²/n
平均 絶対偏差	0.2854	0.3861	Σ |y[j]-ρ[j]|/n
誤分類率	0.1915	0.3810	Σ (ρ[j]≠ρMax)/n
N	94	42	n

混同行列

学習

実測値	予測値 High	予測値 Low
重症		
High	46	12
Low	6	30

検証

実測値	予測値 High	予測値 Low
重症		
High	15	11
Low	5	11

決定行列

学習

実測値	決定行列 High	決定行列 Low
重症		
High	51	7
Low	12	24

検証

実測値	決定行列 High	決定行列 Low
重症		
High	20	6
Low	6	10

指定された利益行列

実測値	決定 High	決定 Low
High	1	-5
Low	-3	1

実測値	決定行列 High	決定行列 Low
重症		
High	0.879	0.121
Low	0.333	0.667

実測値	決定行列 High	決定行列 Low
重症		
High	0.769	0.231
Low	0.375	0.625

誤分類率
0.2021

誤分類率
0.2857

メモ: 決定行列が表示されるのは、応答に「利益行列」列プロパティがあるか、または［利益行列の指定］オプションによって利益行列を指定した場合だけです。決定行列は「あてはめの詳細」レポートの一部として表示されます。

利益行列を指定すると、その利益行列に基づいて各決定の利益が計算されます。そして、［あてはめの詳細の表示］を選択すると、「決定行列」レポートが表示されます。

「決定行列」レポートの「決定行列 度数」は、指定された利益行列から計算される利益が最大となるような予測を反映しています。このレポートには、学習セット、および検証セットとテストセット（それらが定義されている場合）に対する「決定行列 度数」行列と「決定行列 割合」行列が表示されます。また、参考として利益行列も表示されます。

メモ：［利益行列の指定］オプションを使って利益行列の重みを変更した場合は、その変更を反映するように「決定行列」レポートが自動的に更新されます。

「決定行列 度数」行列　行が実測されたカテゴリ、列が分類されたカテゴリで、各セルに度数を含んだ2元表を表示します。

指定された利益行列　利益行列を定義する重みです。

「決定行列 割合」行列　各行ごとに、分類されているカテゴリの割合を示した行列を表示します。この行列は、すべてのデータが正しく分類されている場合、対角線上の割合がすべて1になります。

ヒント：「決定行列 割合」行列などを求めるのに、デフォルトの利益行列を用いることもできます。このデフォルトの利益行列における利益と損失は、1 と -1 から構成されています。赤い三角ボタンのメニューから［**利益行列の指定**］を選択し、デフォルト値に何も変更を加えないで［**OK**］をクリックすると、このデフォルトの利益行列が使われます。

行列は2列に表示されます。

- 「決定行列 度数」行列は最初の行に表示されます。
- 「指定された利益行列」は最初の行の右側に表示されます。
- 「決定行列 割合」行列は2行目に表示されます。

欠測値をカテゴリとして扱う

［欠測値をカテゴリとして扱う］オプションを使うと、説明変数における欠測値を何らかの情報をもつ値として取り扱います。なお、このオプションを選択すると、分岐のときに、欠測値はランダムに割り振られるのではなく、確定的に取り扱われます。［欠測値をカテゴリとして扱う］オプションは起動ウィンドウにあり、デフォルトで選択されています。このオプションを選択すると、欠測値は以下のように取り扱われます。

- カテゴリカルな説明変数における欠測値は、その変数における1つのカテゴリとして分析に使われます。
- 連続尺度の説明変数における欠測値の処理では、まず、説明変数の非欠測値が並べ替えられます。そして、並べ替えられた値の最後尾に欠測値があるものとみなされ、最適な分岐が探索されます。次に、並べ替えられた値の先頭に欠測値があるものとみなされ、最適な分岐が探索されます。そして、いずれの分岐がより良いかが、対数価値によって判断されます。なお、最初の分岐での処理と同じものが、後続の分岐でも使われます。つまり、ある変数の欠測値が最後尾の値として最初に処理されたら、その変数の欠測値は後

続の分岐でも最後尾の値として処理されます。先頭の値として最初に処理されていたら、後続の分岐でも先頭の値として処理されます。

［欠測値をカテゴリとして扱う］オプションが選択されていない場合、欠測値は以下のように取り扱われます。

- 欠測値がある変数を分岐変数として使用すると、その変数の欠測値のある各行はそれぞれ分岐のいずれかの側に無作為に割り当てられます。
- 欠測値のある変数が分岐変数として初めて使用されると、**「インピュート」**列が要約レポートに追加され、補完の回数が示されます。その後、補完が行われるにつれ、「インピュート」列が更新されます。図5.14を参照してください。ここでは5回の補完が行われています。

メモ：補完の回数は、欠測値を含む行の数より大きくなることがあります。補完は各分岐で発生し、欠測値のある行が無作為に複数回割り当てられることがあるためです。行が無作為に割り当てられるたびに、補完の回数が増加します。

図5.14 要約レポートの「インピュート」の値

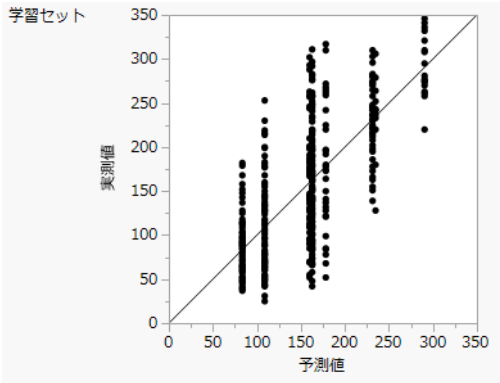
R2乗	RMSE	N	分岐数	インピュート	AICc
0.443	6.8588963	506	1	5	3390.67

予測値と実測値のプロット

応答が連続尺度の場合は、「予測値と実測値のプロット」が実測値と予測値の関係を示す典型的なプロットとして用いられます。ディシジョンツリーをあてはめた場合、各葉のすべてのオブザベーションで予測値は同じになるので、葉の数がn枚の場合、「予測値と実測値のプロット」には、多くてn個の異なる予測値が表示されます。よって、実測値は、n本の縦線上で各葉の平均を中心に散らばります。

対角線はY＝Xの直線です。あてはまりが完全なら、すべての点が対角線上に並びます。検証セットが使用されている場合は、学習セットと検証セットの両方におけるグラフが描かれます。図5.15を参照してください。

図5.15 応答が連続尺度の場合の「予測値と実測値のプロット」



ROC曲線

［ROC曲線］オプションはカテゴリカルな応答に対してのみ使用できます。受診者動作特性（ROC; Receiver Operating Characteristic）曲線は、応答水準を予測確率で並べ替えて、モデルの予測精度を見るものです。ROC曲線の概要については、『基本的な統計分析』の「ロジスティック分析」章を参照してください。

応答変数がカテゴリカルな場合のパーティションでは、その予測値は0～1です。その予測値に対して、特定の**閾値**を設定して、各オブザベーションを陽性／陰性に分類することを考えます。たとえば、閾値を0.5として、予測値が0.5以上ならば「陽性」、0.5未満ならば「陰性」に分類します。この閾値を変化すると陽性／陰性に正しく分類される個数が変化しますが、その分類にはトレードオフがあります。

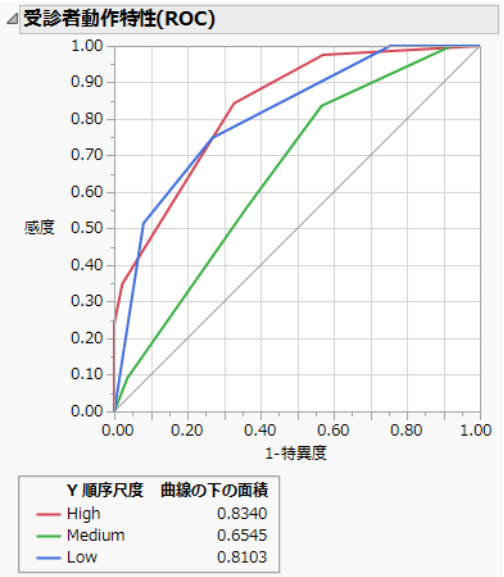
すべての閾値に対して、次のような統計量を求めることにより、ROC曲線は描かれます。

- **感度**（sensitivity）は、真陽性の割合です。「真陽性」とは、イベントが生じているものが、「陽性」と正しく分類されることを指します。
- **特異度**（specifity）は、真陰性の割合です。「真陰性」とは、イベントが生じていないものが、「陰性」と正しく分類されることを指します。

ROC曲線は「1-特異度」に対して「感度」をプロットしたものです。n個の分岐を持つパーティションモデルにはn+1個の予測値があります。そのとき、ROC曲線はn+1本の線分で構成されます。

応答が3水準以上の場合には、1つの水準と、その水準以外のすべての水準とを比較したROC曲線が描かれます。各ROC曲線は、該当の応答水準を「陽性」としたときのものです。なお、2水準の場合には、ある水準を「陽性」と考えたROC曲線は、もう一方の水準を「陽性」と考えたROC曲線と、左上から右下への対角線で対称なものになっています。

図5.16 3つの応答水準のROC曲線



応答の実測値が予測値の大きさの順番どおりに並んでいれば、まず、「陽性」の応答値が先に位置して、その後に「陰性」の応答値が続きます。この場合、ROC曲線はまず上方向へ直進し、そして、右方向へ直進します。逆に、モデルが応答をうまく予測できていない場合には、左下から右上にかけての対角線上にROC曲線はプロットされます。

実際のデータを用いた場合、ROC曲線は対角線の上側にプロットされます。なお、曲線の下側の領域（AUC; Area Under Curve）はモデルの適合度を示します。AUCが1の場合は完全に適合していることを意味します。AUCが0.5に近い値の場合は、モデルによる予測がうまくいかないことを示します。

応答に3水準以上の場合は、ROC曲線を描くと、曲線の下側の領域（AUC）が最も大きい応答水準を知ることができます。

リフトチャート

［リフトチャート］は、ROC曲線とは別の角度から、モデルの予測精度を描いたものです。リフトチャートは、データの部分ごとのリフト値をプロットしたものです。まず、応答の予測値でデータを並び替えて、ある閾値以上の予測値のものでグループを構成します。**リフト**値は、そのグループでのイベントが生じているものの割合が、データ全体でのイベントが生じてるものの割合の何倍になっているかを示したものです。

図5.17 リフトチャート

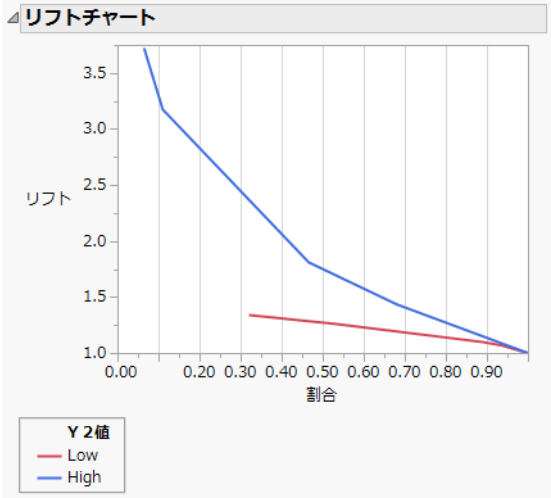


図5.18 リフトチャートのリフト表

Prob High	N > Prob High	Portion	N High in Portion	Portion High	Lift = portion high/ overall high of .27
0.97	20	0.06	20	1.00	3.72
0.77	44	0.14	39	0.89	3.30
0.33	68	0.22	47	0.69	2.57
0.31	168	0.54	78	0.46	1.73
0.04	309	1.00	83	0.27	1.00

図5.18は、図5.17のリフトチャートのうち「High」に対するものの「リフト」と「割合」の計算を示したものです。あてはめられたパーティションのモデルは、応答**「Y 2値」**を予測するためのもので、5つの分岐を持ちます。その**「Y 2値」**列には、「Low」と「High」の2水準があります。このリフトチャートは、309行のデータに基づいています。応答が「High」のデータは83行あり、それは全体の27%（= 83/309）となっています。

- **Prob High:** モデルから求められた5つの予測値。
- N > Prob High: **「Prob High」**にリストされている数値以上の予測値であるオブザベーション数。
- **Portion:** 「N > Prob High」を309で割った値。
- **N High in Portion:** 該当グループにおいて、応答が「High」となっているオブザベーション数。
- **Portion High:** **「N > Prob High」**を**「N High in Portion」**で割った値。
- Lift: **「Portion High」**を0.27で割った値。

リフト値は、モデルの予測値が大きいもので構成されたグループの「High」の割合が、データ全体での「High」の割合と比べてどれだけ大きくなっているかを示しています。データの最初の6%に対するリフト値は3.72です。つまり、このモデルの予測値が大きい上位6%を選び出すと、その6%のグループを無作為抽出したときよりも、応答が「High」になっている人の人数は3.72倍になっています。

ノードのオプション

ここでは、各ノードの赤い三角ボタンをクリックしたときに表示されるオプションを紹介します。

最良分岐　そのノードより下に位置するノードにおいて、最適な分岐を行います。

ここを分岐　そのノードで、最適な列によって分岐します。

分岐変数の指定　分岐点を指定します。どのような基準で分岐点が決まるかがわかるだけでなく、独自の分岐点を設定することもできます。分岐列を指定する際には、分岐する位置を指定するための以下のオプションを選択できます。

最適の値　選択した変数の最適値で分岐が行われます。

指定した値　ユーザが指定した値で分岐が行われます。

分岐テーブルの出力　分岐点の候補とその値を示すデータテーブルが開きます。

下を剪定　そのノードより下の分岐を削除します。

最悪分岐を剪定　そのノードより下に位置する最悪の分岐を削除します。

行の選択　その葉に該当する行をデータテーブル内で強調表示します。複数のノードの行を選択するには、Shiftキーを押しながら他のノードからこのコマンドを選択します。

詳細の表示　選択した変数の分岐基準を示すデータテーブルが作成されます。データテーブルには、分岐点とその基準値が保存され、また、分岐点と基準値のグラフを作成するスクリプトも含まれます。

ロック　ロックしたノードとそのサブノードは、分岐の対象になりません。チェックマークをつけると、ノードのタイトルに鍵のアイコンが表示されます。

検証

分岐の数が多いツリーを作成すると、データにオーバーフィット（過学習）する可能性があります。オーバーフィットすると、モデルの作成に用いたデータでの予測は精確でも、将来のデータに対する予測精度は悪くなります。検証（validation）とは、データの一部をモデルパラメータの推定に使用し、残りのデータでモデルの予測能力を評価する方法を指します。

- モデルパラメータの推定に使うデータを、**学習**セットといいます。
- モデルの予測能力を評価するのに用いるデータを、**検証**セットといいます。
- 予測能力の最終評価に使うデータを、**テスト**セットといいます。検証列を指定した場合のみ、テストセットを使用できます。「「パーティション」プラットフォームの起動」（75ページ）を参照してください。

検証セットを指定した場合には、［**実行**］ボタンが表示されます。この［**実行**］ボタンは、［**分岐**］ボタンを手動で繰り返し押すことなしに、分岐処理を一度に実行したいときに使います。［**実行**］ボタンをクリックすると、現時点より後の10回のどの分岐においても、検証セットのR2乗が改善されない時点まで、一度に処理が実行されます。この方法では、解釈しにくい複雑なツリーになるかもしれませんが、求められたツリーの予測精度は高いでしょう。

［**実行**］ボタンを使ったときには、［**分岐履歴**］コマンドがオンになります。なお、［**実行**］ボタンを使ったときにノードの数が40を超えるツリーができた場合は、［**ツリーの表示**］コマンドはオフになります。

学習セット、検証セット、テストセットは、分析対象のデータを分割することにより作成します。データを分割する方法としては、以下のような方法が用意されています。

除外された行　行の属性によって、データを分割します。除外されていない行を学習セット、除外されている行を検証セットとして用います。

行の属性と行の除外の詳細については、『JMPの使用法』の「データの入力と編集」章を参照してください。

保留　データを無作為に学習セットと検証セットに分割します。プラットフォームの起動ウィンドウにある「検証データの割合」で、検証セットとして使用する部分の割合（保留する割合）を指定することができます。検証データの割合の詳細については、「「パーティション」プラットフォームの起動」（75ページ）を参照してください。

K分割交差検証　データをランダムに*K*個に分割します。順番に、*(K-1)*個分のデータにモデルがあてはめられ、残っているデータでモデルが検証されます。全部で*K*個のモデルがあてはめられます。最終的なモデルは交差検証のR2乗により選択されます。なお、JMPの「パーティション」では、オーバーフィットを防ぐために、ある停止ルールが使われています。この方法は、少ないデータを効率的に利用するので、小規模なデータセットに適しています。「K分割交差検証」（94ページ）を参照してください。

JMP PRO **検証列** 検証列の値に基づいて、データを分割します。検証列には、多くとも3つの数値が含まれている必要があります。検証列の指定は「パーティション」起動ウィンドウで行います。「「パーティション」プラットフォームの起動」（75ページ）を参照してください。

この列の値によって、データの分割方法が決まります。

- 検証列の水準が2つの場合は、小さい方の値をもつ行が学習セット、大きい方の値をもつ行が検証セットとして使われます。
- 水準が3つの場合は、値が小さいものから順に、学習セット、検証セット、テストセットとして使われます。

「列の選択」リストで列を選択せず、［検証］ボタンをクリックすると、データテーブル内に検証列を作成することができます。「検証列の作成」機能の詳細については、「「検証列の作成」ユーティリティ」を参照してください。

K分割交差検証

K分割交差検証では、データが*K*個に分割されます。その1つの部分が検証セットとして、残りの部分が学習セットとして扱われます。

R2乗値による交差検証法は、何も制約がない場合、オーバーフィットしがちです。この傾向に対処するために、交差検証R2乗の増加量が少なくなった時点で処理を停止します。具体的には、現時点から10個のモデルにおいていずれのモデルも交差検証R2乗値の増加量が0.005より大きくなっていないものが選択されます。

［K分割交差検証］オプションを選択すると、レポートが表示されます。このレポートの結果は、ディシジョンツリーを分割するたびに更新されます。または、［実行］をクリックすると、アウトラインに最終的なモデルの結果が表示されます。

K分割交差検証のレポート

K分割交差検証のレポートには次の情報が表示されます。

K分割 分割数。

(-2)*対数尤度またはSSE 応答がカテゴリカルなとき、(-1)*対数尤度の2倍、つまり(-2)*対数尤度値が計算されます。応答が連続尺度の場合は、誤差平方和（SSE）が計算されます。最初の行には、各分割にわたって平均化された結果が表示されます。2行目には、すべてのオブザベーションに対する1つのモデルのあてはまりの結果が表示されます。

R2乗 最初の行には、各分割にわたって平均化されたR2乗値が表示されます。2行目には、すべてのオブザベーションに対する1つのモデルのあてはまりのR2乗値が表示されます。

パーティションの別例

以下、連続尺度の応答の例、説明変数における欠測値の例、および利益行列を使用する例を紹介します。

連続尺度の応答変数を用いた例

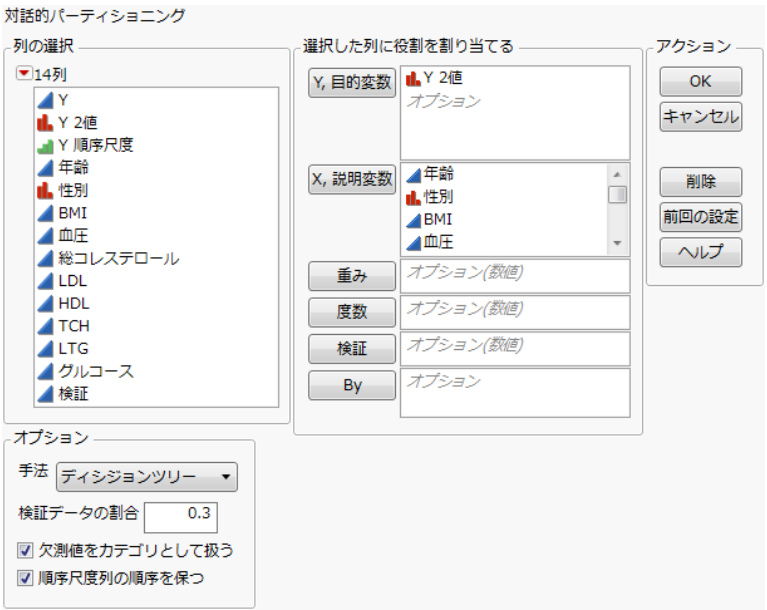
この例では、「パーティション」プラットフォームを使用して、糖尿病患者における 1 年間の症状進行を予測するディシジョンツリーを構築します。この症状進行は、数値で測定されています。

1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Diabetes.jmp」を開きます。
2. ［**分析**］>［**予測モデル**］>［**パーティション**］を選択します。
3. 「**Y**」を選択し、［**Y, 目的変数**］をクリックします。
4. 「**年齢**」から「**グルコース**」までを選択し、［**X, 説明変数**］をクリックします。
5. お使いのJMPに基づいて検証手順を選択します。
   - JMP Proの場合は、「**検証**」を選択し、［**検証**］をクリックします。
   - JMPの場合は、「**検証データの割合**」に「0.3」と入力します。

   指定が完了すると、JMPユーザの起動ウィンドウは図5.19のようになります。

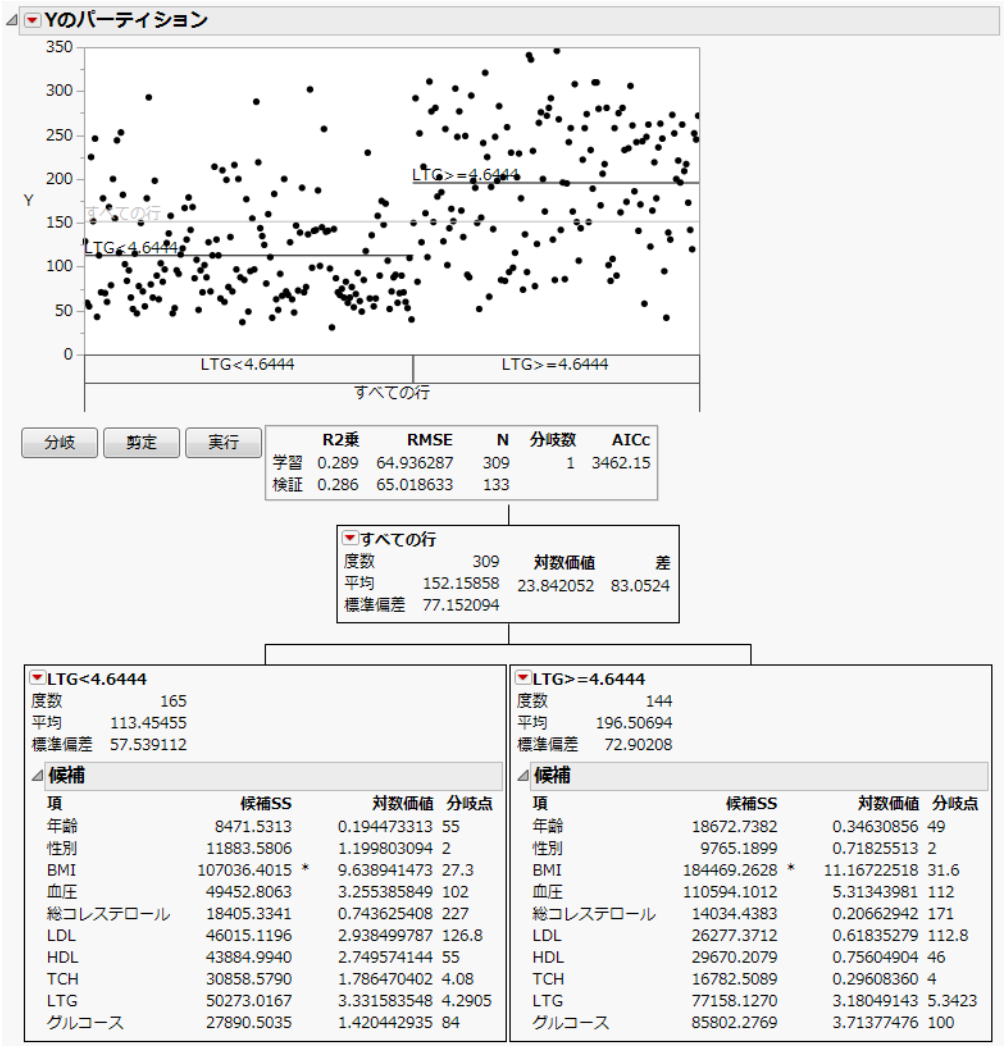
メモ：「検証データの割合」を使用した場合、検証セットが無作為に選択されるために、本節で示すものと結果が異なってきます。

図5.19 指定が完了した起動ウィンドウ、「検証データの割合」を「0.3」に指定



6. ［OK］をクリックします。
7. プラットフォームのレポートウィンドウで、［分岐］を1回クリックして分岐を実行します。

図5.20 ディシジョンツリーを隠した状態の最初の分岐後のレポート



学習セットには309行がありますが、それらが2つに分割されます。

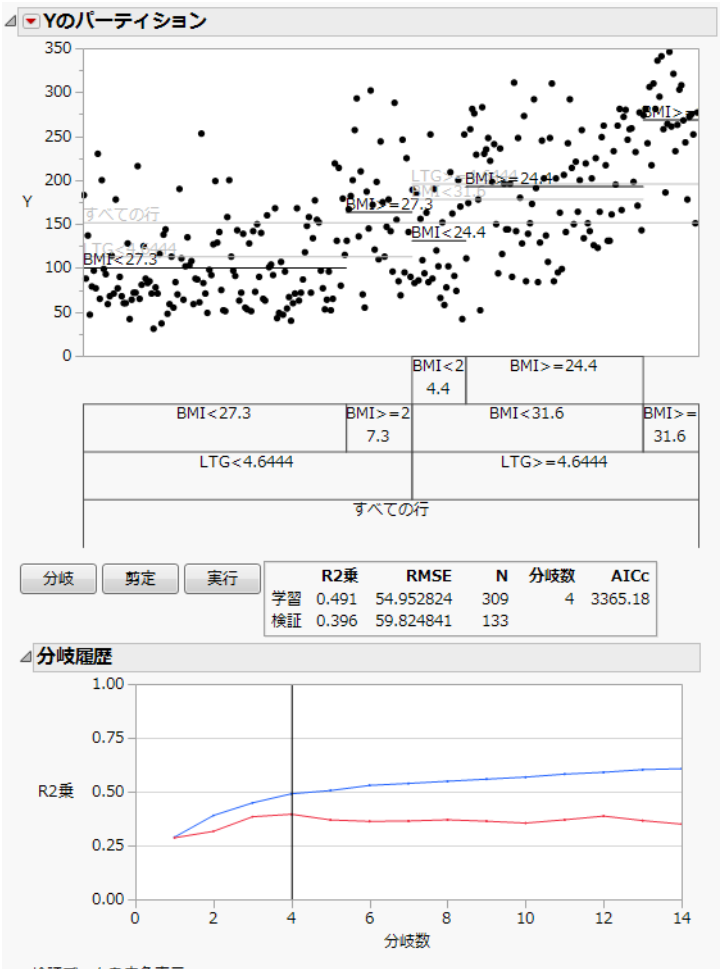
– 左の葉は「LTG」が4.6444未満のデータで、165行あります。

– 右の葉は「LTG」が4.6444以上のデータで、144行あります。

左の葉も右の葉も、次の分岐は「BMI」で行われます。右の葉の「BMI」の「候補SS」は、左の葉の「BMI」の「候補SS」より高くなっています。したがって、次の分岐は右の葉で行われます。

8. ［実行］をクリックして自動分岐を使用します。

図5.21 検証を使った自動分岐後のレポート



最終的なツリーでは、4回の分岐が行われています。「分岐履歴」プロットを見ると、その4回の分岐のあとには、検証セットにおけるR2乗値が改善していません。検証セットのR2乗値は0.39ですので、このモデルでは症状進行を精確には予測できないようです。また、先頭のプロットからも、各ノードにおいて応答変数のばらつきが大きいことから、応答変数を精確に予測できていないことが分かります。

欠測値をカテゴリとして扱う例

この例では、顧客の信用リスクを予想するためのディシジョンツリーを構築します。データには欠測値が含まれているため、［欠測値をカテゴリとして扱う］オプションが役立つかどうかも見てみます。

「パーティション」プラットフォームの起動

1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Equity.jmp」を開きます。
2. ［**分析**］>［**予測モデル**］>［**パーティション**］を選択します。
3. 「**BAD**」を選択し、［**Y, 目的変数**］をクリックします。
4. 「**LOAN**」から「**DEBTINC**」までを選択し、［**X, 説明変数**］をクリックします。
5. ［**OK**］をクリックします。

［欠測値をカテゴリとして扱う］を使ったディシジョンツリーとROC曲線の作成

1. Shiftキーを押しながら［**分岐**］をクリックします。
2. 「分岐数を入力」に「5」と入力し、［**OK**］をクリックします。
3. 「BADのパーティション」の赤い三角ボタンをクリックし、［**ROC曲線**］をクリックします。
4. 「BADのパーティション」の赤い三角ボタンをクリックし、［**列の保存**］>［**予測式の保存**］を選択します。

   「**確率(BAD==Good Risk)**」列と「**確率(BAD==Bad Risk)**」列には、「欠測値をカテゴリとして扱う」オプションに基づいた予測式によって、ローン申込者の信用リスクを分類する計算式が含まれています。この予測式を、［欠測値をカテゴリとして扱う］を使わなかった場合の予測式と比較してみましょう。

［欠測値をカテゴリとして扱う］を使わないディシジョンツリーとROC曲線の作成

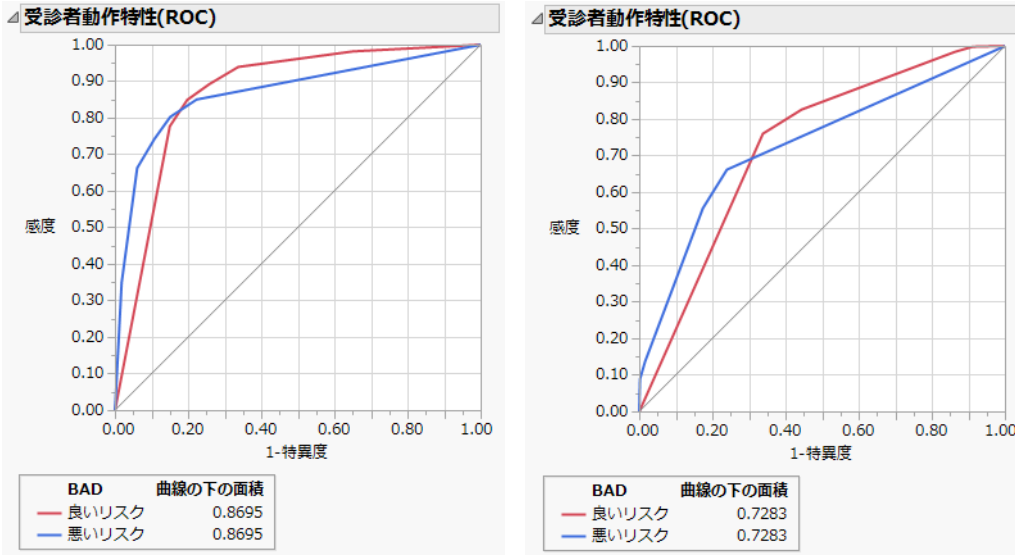
1. 「BADのパーティション」の赤い三角ボタンをクリックし、［**やり直し**］>［**分析の再起動**］を選択します。
2. ［**欠測値をカテゴリとして扱う**］の選択を解除します。
3. ［OK］をクリックし、「［欠測値をカテゴリとして扱う］を使ったディシジョンツリーとROC曲線の作成」の手順を繰り返します。

   「**確率(BAD==Good Risk) 2**」列と「**確率(BAD==Bad Risk) 2**」列には、［欠測値をカテゴリとして扱う］を使わない計算式が含まれています。

ROC曲線の比較

2つのモデルのROC曲線を比較してみてください。左側は［欠測値をカテゴリとして扱う］を使ったモデルで、右側は［欠測値をカテゴリとして扱う］を使わなかったモデルです。

図5.22［欠測値をカテゴリとして扱う］を使ったモデル（左）と［欠測値をカテゴリとして扱う］を使わなかったモデルのROC曲線



［欠測値をカテゴリとして扱う］を使ったモデルの曲線の下側の領域（AUC）（0.8695）は、［欠測値をカテゴリとして扱う］を使わなかったモデルのAUC（0.7283）より大きくなっています。この例の応答変数は2水準なので、一方の水準のROC曲線は他方の水準のROC曲線と対角線で対称となっています。また、それら両者のAUCは等しいです。

メモ：［欠測値をカテゴリとして扱う］を使わなかった場合、得られるAUCは分析ごとに変化します。［欠測値をカテゴリとして扱う］を使わなかった場合、欠測値を含む行は、分岐のいずれかに無作為に割り当てられます。そのため、分析を再実行すると、結果がわずかに異なります。

「モデルの比較」プラットフォームの使用

次に、「モデルの比較」プラットフォームを用いてこれらのモデルを比較してみましょう。先ほどの手順で作成した2つの予測式の違いを見てみましょう。

1. **［分析］＞［予測モデル］＞［モデルの比較］**を選択します。
2. **「確率(BAD==Good Risk)」**、**「確率(BAD==Bad Risk)」**、**「確率(BAD==Good Risk) 2」**、および**「確率(BAD==Bad Risk) 2」**を選択し、**［Y, 予測子］**をクリックします。

   最初の計算式列ペアには［欠測値をカテゴリとして扱う］を使ったモデルの計算式が含まれています。2つ目の計算式列ペアには［欠測値をカテゴリとして扱う］を使わなかったモデルの計算式が含まれています。
3. **［OK］**をクリックします。

図5.23「モデルの比較」で得られた適合度指標

BADの適合度指標

作成方法	.2.4.6.8	エントロピーR2乗	一般化R2乗	平均 -Log p	RMSE	平均 絶対偏差	誤分類率	N
パーティション		0.3813	0.5015	0.3092	0.3013	0.1817	0.1158	5960
パーティション		0.1226	0.1825	0.4384	0.3689	0.2755	0.1864	5960

「適合度指標」レポートを見ると、［欠測値をカテゴリとして扱う］を使った最初のモデルの方が、［欠測値をカテゴリとして扱う］を使わなかった2つ目のモデルより予測精度が良いことがわかります。最初のモデルの方が、R2乗値が大きく、RMSE値と誤分類率が小さくなっています。これらの傾向はROCを比較しても分かります。

メモ： 前述したとおり、［欠測値をカテゴリとして扱う］を使わなかったときの結果は、乱数に依存しているために異なります。

利益行列と決定行列の例

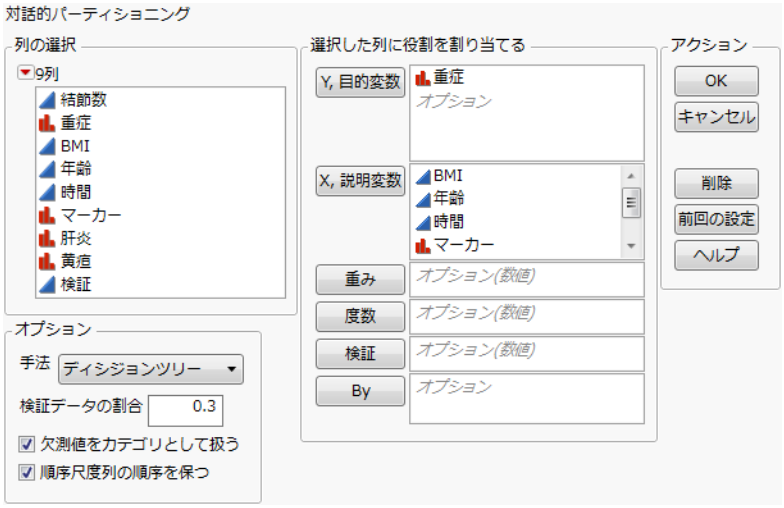
この例では、肝臓がんを患う患者の調査を見てみましょう。あなたは、さまざまな測定値やマーカーに基づき、病気の重症度（HighまたはLow）のいずれかに患者を分類したいとします。患者を分類する上で考えられる間違いには2つあります。つまり、重症度がHighの患者をLowグループに分類してしまうこと、および重症度がLowの患者をHighグループに分類してしまうことです。臨床上、重症度が実際にはHighである患者をLowに分類してしまうことはコストが大きいと考えられます。なぜなら、その患者は必要な治療を受けることができない可能性があるからです。それに比べ、重症度が実際にはLowの患者をHighに分類してしまっても、それほどコストは大きくないと考えられます。その患者は必要とされる以上の治療を受けるかもしれませんが、それが大きな問題になることはないと考えられます。

次の例では、肝臓がんに関して利益行列を定義します。分析すると、「決定行列」レポートを得ます。この「決定行列」レポートを見ることで、定義した利益行列のコストに応じた分類を評価できます。

1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Liver Cancer.jmp」を開きます。
2. ［**分析**］>［**予測モデル**］>［**パーティション**］を選択します。
3. 「**重症**」を選択し、［**Y, 目的変数**］をクリックします。
4. 「**BMI**」から「**黄疸**」までを選択し、［**X, 説明変数**］をクリックします。
5. お使いのJMPに基づいて検証手順を選択します。
   - JMP Proの場合は、「**検証**」を選択し、［**検証**］をクリックします。
   - JMPの場合は、「**検証データの割合**」に「0.3」と入力します。

メモ：「検証データの割合」を使用した場合、検証セットが無作為に選択されるために、ここで示すものと結果が異なってきます。

図5.24 指定が完了した起動ウィンドウ、「検証データの割合」を「0.3」に指定



6. ［**OK**］をクリックします。
7. Shiftキーを押しながら［**分岐**］をクリックします。
8. 「分岐数を入力」に「10」と入力し、［**OK**］をクリックします。
   プロットの下のパネルで「分岐数」が10になっていることを確認してください。
9. 「重症のパーティション」の横にある赤い三角ボタンをクリックし、［**利益行列の指定**］を選択します。
10. 入力内容を次のように変更します。
    - High行High列のボックスに「1」と入力します。
    - High行Low列のボックスに「-5」と入力します。
    - Low行High列のボックスに「-3」と入力します。
    - Low行Low列のボックスに「1」と入力します。

図5.25 指定後の利益行列

決定または予測

実測値	High	Low	その他
High	1	-5	.
Low	-3	1	.

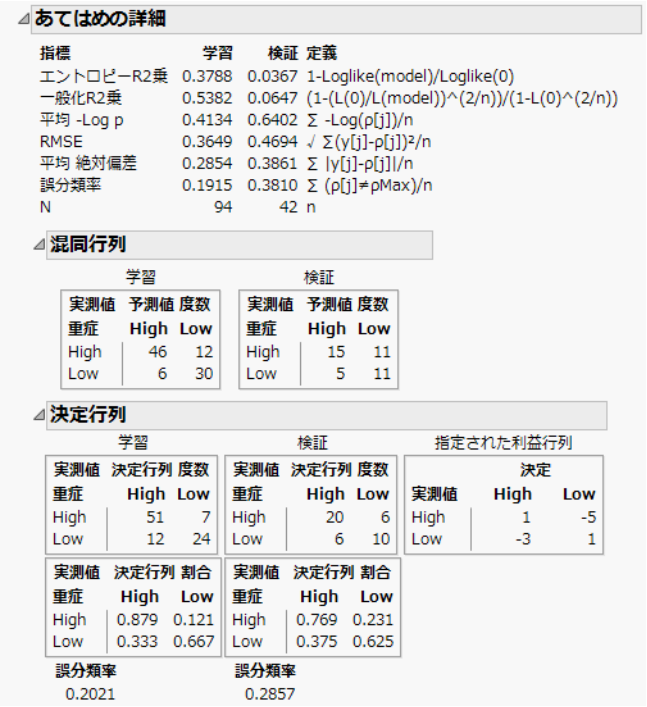
ヒント：この利益行列は今後の分析で使用できるように列プロパティとして保存できます。その場合は、利益行列ウィンドウの下部にある［列プロパティとして保存する］チェックボックスをオンにします。

次の点を念頭に置いてください。

- 値1はいずれも、正しい決定を下したときの利益を反映しています。
- 値-3は、重症度が実際にはLowの患者をHighとして分類してしまった場合に、損失が正しい決定を下したときに得られる利益の3倍であることを意味します。
- 値-5は、重症度が実際にはHighの患者をLowとして分類してしまった場合に、損失が正しい決定を下したときに得られる利益の5倍であることを意味します。

11. [**OK**] をクリックします。
12.「重症のパーティション」の横にある赤い三角ボタンをクリックし、[**あてはめの詳細の表示**] を選択します。

図5.26「混同行列」レポートと「決定行列」レポート



あてはめの詳細

指標	学習	検証	定義
エントロピーR2乗	0.3788	0.0367	1-Loglike(model)/Loglike(0)
一般化R2乗	0.5382	0.0647	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
平均 -Log p	0.4134	0.6402	Σ -Log(ρ[j])/n
RMSE	0.3649	0.4694	√ Σ(y[j]-ρ[j])²/n
平均 絶対偏差	0.2854	0.3861	Σ |y[j]-ρ[j]|/n
誤分類率	0.1915	0.3810	Σ (ρ[j]≠ρMax)/n
N	94	42	n

混同行列

学習

実測値 重症	予測値 High	予測値 Low
High	46	12
Low	6	30

検証

実測値 重症	予測値 High	予測値 Low
High	15	11
Low	5	11

決定行列

学習

実測値 重症	決定行列 High	決定行列 Low
High	51	7
Low	12	24

検証

実測値 重症	決定行列 High	決定行列 Low
High	20	6
Low	6	10

指定された利益行列

実測値	決定 High	決定 Low
High	1	-5
Low	-3	1

実測値 重症	割合 決定行列 High	割合 決定行列 Low
High	0.879	0.121
Low	0.333	0.667

実測値 重症	割合 決定行列 High	割合 決定行列 Low
High	0.769	0.231
Low	0.375	0.625

誤分類率	誤分類率
0.2021	0.2857

「あてはめの詳細」レポートの指標リストの下には「混同行列」レポートと「決定行列」レポートが表示されます。「混同行列」レポートと「決定行列」レポートの混同行列とでは、表示されている数値が異なることに注意してください。これは、利益行列に基づく分類が、何も重みづけないで行われた分類と異なるからです。

検証セットの混同行列には、予測確率のみに基づく分類が示されています。これによると、実際には重症度がHighの患者のうち11人はLowに分類され、実際には重症度がLowの患者のうち5人はHighに分類されることになります。

「決定行列」レポートでは、利益行列の重みが考慮されています。これらの重みを使用することにより、実際には重症度がHighの患者のうちLowに分類されるのは6人だけになります。その代り、実際には重要度がLowの患者のうち6人が誤ってHighの重症度グループに分類されてしまいます（1人多くなる）。

13.「重症のパーティション」の赤い三角ボタンをクリックし、**［列の保存］>［予測式の保存］**を選択します。

8個の列がデータテーブルに追加されます。

ヒント：データテーブルにすばやく戻るには、レポートウィンドウの右下隅にある［データの表示］ボタン をクリックします。

– 最初の3列は、予測確率に関するものです。このうち**「最尤　重症」**列は、最も予測確率の高い水準に分類しています。「混同行列」レポートの度数は、この結果に基づいています。予測確率は、**「確率(重症y==High)」**列と**「確率(重症==Low)」**列に含まれています。

– 最後の5列は、指定した利益行列に基づくものです。**「利益が最大となる重症の予測値」**列には利益行列に基づく決定が含まれています。そこでは、利益が最大となる水準に各患者が分類されています。利益は**「Highの利益」**列と**「Lowの利益」**列に含まれています。

統計的詳細

ここでは、統計量の詳細などを説明します。

目的変数と説明変数

応答変数（目的変数）は、連続尺度とカテゴリカル（名義／順序尺度）のどちらでもかまいません。

- 応答変数がカテゴリカルの場合は、それぞれの葉における各応答水準の割合が予測値となります。この場合、各水準の割合は、その葉のエントロピー（負の対数尤度）を最小にします。
- 応答変数が連続尺度の場合は、それぞれの葉における応答平均が予測値となります。この場合、平均は、その葉の誤差平方和を最小にします。

説明変数にも、連続尺度とカテゴリカル（名義／順序尺度）の両方を使用できます。

- 説明変数がカテゴリカルの場合は、特定の境界値の上下で分岐が行われます。
- 説明変数がカテゴリカルの場合、可能な限りのグループ分けが検討されて、Xの水準が2つのグループに振り分けられます。

分岐基準

ノードの分岐は、「候補」レポートに表示される対数価値の値に従って行われます。対数価値は次の式で計算されます。

$-\log_{10}(p\text{ 値})$

p 値は、考えられる分岐候補の組み合わせ数を考慮した複雑な方法で調整されています。未調整の *p* 値を用いると、水準数の多い説明変数に有利となります。また、それを修正しようとして、Bonferroni の *p* 値などを用いると、逆に、水準数の少ない説明変数を優先しぎます。JMP のパーティションで使われている調整 *p* 値は、それらに比べて公平な分析を可能にします。この方法の詳細については、Sall（2002）を参照してください。

応答変数が連続尺度の場合は、分岐によって誤差平方和がどれぐらい減少するかが考慮されます。そして、各ノードの統計量としては平方和（SS）が報告されます。

選択された候補の SS は次のように計算されます。

$SS_{test} = SS_{parent} - (SS_{right} + SS_{left})$、この式でノード内の SS は $s^2(n - 1)$

応答変数が連続尺度の場合は、統計量として差も報告されます。これは、親ノードから分岐した 2 つの子ノードの予測値の差です。

応答がカテゴリカルな場合は、G^2（尤度比カイ 2 乗）がレポートに表示されます。これは（自然対数）エントロピーに 2 を掛けたもの、またはエントロピーの変化量に 2 を掛けたものとして計算されます。各オブザベーションの応答が起こる確率を *p* とすると、エントロピーは $\Sigma\ -\log(p)$ です。

選択された候補の G^2 は次のように計算されます。

$G^2_{\text{ test}} = G^2_{\text{ parent}} - (G^2_{\text{ left}} + G^2_{\text{ right}})$

「パーティション」では、度数から計算される「割合」と、ゼロにならないように若干のバイアスを「割合」に加えた「確率」の 2 つが使用されます。「確率」の方はゼロにならないよう工夫されているため、検証セットや除外したデータセットに対しても、その対数を計算でき、エントロピー R2 乗も計算できます。

ディシジョンツリーとブートストラップ森の予測確率

この節では、ディシジョンツリーとブートストラップ森における予測確率の計算方法を説明します。

ディシジョンツリーにおいて、応答変数がカテゴリカルの場合、［割合を表示］コマンドにより次の統計量が表示されます。

割合　応答の各水準の、ノードに占める割合を示します。

p値　応答の各水準の予測確率を示します。あるノードにおける、応答の第i水準の「確率」は、次式によって計算されます。

$$Prob_i = \frac{n_i + prior_i}{\sum(n_i + prior_i)}$$

上の式で、合計は応答の全水準にわたって行われ、n_iはそのノードにおける第i水準のオブザベーション数、$prior_i$は第i水準の事前確率です。事前確率は次の式で計算されます。

$$prior_i = \lambda p_i + (1-\lambda)P_i$$

上の式で、p_iは親ノードの$prior_i$、P_iは親ノードの$Prob_i$、λは重み係数で現在0.9に設定されています。

このように定義された「確率」は、ゼロにならないという長所があります。

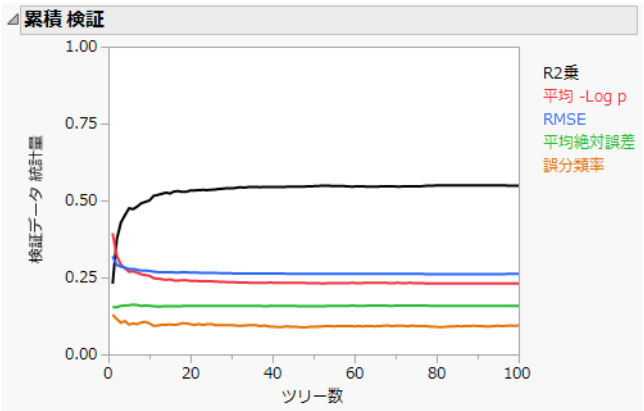
第6章

JMP PRO ブートストラップ森
多数のツリーを平均化して予測

「ブートストラップ森」プラットフォームは JMP Pro でのみ使用できます。

「ブートストラップ森」プラットフォームは、複数のディシジョンツリー（決定木）を組み合わせたモデルをあてはめます。この際、複数のツリーが、データから一部分を何度も復元無作為抽出（ブートストラップ抽出）することにより求められます。各ツリーの各分岐では、説明変数も無作為抽出して決められます。このようにして得られる、多くの「弱いモデル」を組み合わせることにより、予測精度の高いモデルを作成します。最終的な予測値は、すべてのディシジョンツリーから得られる予測値を平均したものです。

図6.1「累積 検証」レポートの例



JMP PRO「ブートストラップ森」プラットフォームの概要

「ブートストラップ森」プラットフォームは、多くのディシジョンツリーにおける応答の予測値を平均することによって応答を予測します。各ツリーはそれぞれ学習データの**ブートストラップ標本**に対して生成されます。ブートストラップ標本はデータから一部分を無作為に復元抽出したものです。何度も無作為抽出されたデータにツリーをあてはめていきます。さらに、各ツリーの各分岐において、説明変数も無作為抽出されます。各ツリーは、「パーティション」章で説明している手法によって求められます。

「ブートストラップ森」では、現在の学習セットに対して以下の手順によってモデルをあてはめます。

1. 学習セットからブートストラップ標本を抽出します。
2. その抽出された標本に対して、ディシジョンツリー（決定木）をあてはめます。
   - この際、各分岐において、説明変数を無作為に選択します。
   - 「ブートストラップ森の指定」ウィンドウで指定されている停止ルールの条件が満たされるまで、分岐を続けます。
3. 「ブートストラップ森の指定」ウィンドウで指定されているツリー数に達するまで、または早期打ち切りが発生するまで、手順1と手順2を繰り返します。

ブートストラップ標本の抽出には、復元抽出が使用されます。抽出されるオブザベーションの割合は指定できます。100%のオブザベーションが抽出されるように指定した場合は、各復元抽出において1度も抽出されないオブザベーションの割合は、およそ1/e（約36.8%）です。各抽出において、これらの抽出されなかったオブザベーションは**「バッグ外標本（out-of-bag）」**と呼ばれます。逆に、1度以上、抽出されたオブザベーションは**「バッグ内標本（in-bag）」**と呼ばれます。応答変数が連続尺度である場合、「ブートストラップ森」プラットフォームはバッグ外標本から計算される統計量（**「バッグ外誤差（out-of-bag error）」**という）も求めます。

応答変数が連続尺度である場合、あるオブザベーションにおける最終的な予測値は、個々のツリーにおける予測値をまとめて平均したものです。応答変数がカテゴリカルな場合、最終的な予測確率は、個々のツリーにおける予測確率をまとめて平均したものです。そして、各オブザベーションは、その最終的な予測確率が最も高い水準に分類されます。

ブートストラップの詳細については、Hastie et al.（2009）を参照してください。

JMP PRO カテゴリカルな応答を扱うブートストラップ森の例

この例では、顧客の信用リスクが悪いかどうかを予測するブートストラップ森モデルを構築します。また、データセットには欠測値が含まれていることがわかっているため、データがどれぐらい欠測しているかも調べてみます。

JMP PRO ブートストラップ森モデル

1. ［ヘルプ］>［サンプルデータライブラリ］を選択し、「Equity.jmp」を開きます。

2. ［分析］>［予測モデル］>［ブートストラップ森］を選択します。
3. 「BAD」を選択し、［Y, 目的変数］をクリックします。
4. 「LOAN」から「DEBTINC」までを選択し、［X, 説明変数］をクリックします。
5. 「Validation」列を選択し、［検証］ボタンをクリックします。
6. ［OK］をクリックします。
7. 「ツリーあたりの最大分岐数」の横に「30」と入力します。
8. ［項数に対する複数のあてはめ］を選択し、「項の最大数」の横に「5」と入力します。
9. （オプション）［マルチスレッドをオフにする］を選択し、「乱数シード値」の横に「123」と入力します。

   ブートストラップ森では無作為抽出が行われるため、上記の操作を行うことにより、下の図とまったく同じ結果を得ることができます。
10. ［OK］をクリックします。

図6.2「全体の統計量」レポート

ブートストラップ森（BAD）

検証セットでのモデル要約

これらのなかで最適なモデルの結果が表示されています。

項の数	ツリーの数	エントロピーR2乗	誤分類率	平均 -Log p	RMSE	平均絶対誤差
3	44	0.4649	0.1183	0.2890	0.2937	0.1924
4	16	0.4707	0.1208	0.2859	0.2930	0.1853
5	24	0.4828	0.1133	0.2793	0.2886	0.1779

設定

全体の統計量

指標	学習	検証	テスト	定義
エントロピーR2乗	0.5124	0.4828	0.4824	1-Loglike(model)/Loglike(0)
一般化R2乗	0.6302	0.6153	0.6068	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
平均 -Log p	0.2354	0.2793	0.2618	Σ -Log(p[j])/n
RMSE	0.2596	0.2886	0.2771	√ Σ(y[j]-p[j])²/n
平均 絶対偏差	0.1580	0.1779	0.1688	Σ |y[j]-p[j]|/n
誤分類率	0.0898	0.1133	0.0990	Σ (p[j]≠pMax)/n
N	3576	1192	1192	n

混同行列

学習

実測値	予測値 度数	
BAD	良いリスク	悪いリスク
良いリスク	2819	86
悪いリスク	235	436

検証

実測値	予測値 度数	
BAD	良いリスク	悪いリスク
良いリスク	884	33
悪いリスク	102	173

テスト

実測値	予測値 度数	
BAD	良いリスク	悪いリスク
良いリスク	921	28
悪いリスク	90	153

［項数に対する複数のあてはめ］オプションでの指定に従って、各分岐で説明変数が3～5個、使用されます。「検証セットでのモデル要約」レポートを見ると、検証セットにおいて「エントロピー R2 乗」が最大となるのは、各分岐で説明変数を5個、抽出したモデルです。このモデルは、誤分類率も最小になっています。このモデルが最良のモデルであり、「全体の統計量」レポートには、このモデルの結果が表示されています。

「全体の統計量」レポートを見ると、「学習」セットと「検証」セットの誤分類率が、それぞれ11.3%と9.9%になっています。混同行列を見ると、誤分類においては、「信用リスクが悪い顧客を信用リスクが良いとして分類している」ほうが多いことが分かります。

「テスト」セットの結果は、将来の独立したデータに対する予測精度を示します。「検証」セットは現在のブートストラップ森モデルを選択する際に使用されました。そのため、「検証」セットの結果は、独立した将来のデータに一般化するには、バイアスをもっています。

次に、このモデルに最も寄与している説明変数を見てみましょう。

11.「ブートストラップ森 (BAD)」の横にある赤い三角ボタンをクリックし、[**列の寄与**] を選択します。

図6.3「列の寄与」レポート

列の寄与

項	分岐数	G^2	割合
DEBTINC	96	567.182771	0.5456
VALUE	80	104.83095	0.1008
DELINQ	83	102.545902	0.0986
DEROG	77	73.3253058	0.0705
CLAGE	68	54.1092672	0.0520
CLNO	55	30.2200294	0.0291
JOB	50	26.1866129	0.0252
NINQ	51	24.5238489	0.0236
LOAN	41	19.7652555	0.0190
MORTDUE	42	17.0533251	0.0164
YOJ	40	16.3836362	0.0158
REASON	13	3.45680616	0.0033

「列の寄与」レポートを見ると、顧客の信用リスクに関する最も強い説明変数は「**DEBTINC**」であることがわかります。これは、収入に対する債務の比です。その次に大きく寄与しているのは、顧客の評価である「**VALUE**」や、滞納している支払いの回数である「**DELINQ**」です。

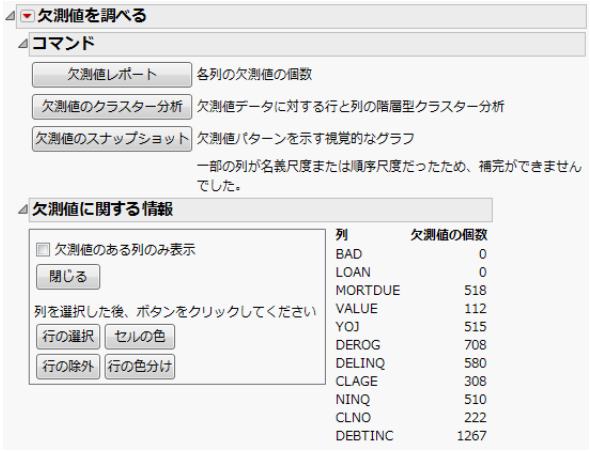
JMP PRO 欠測値

次に、説明変数がどれぐらい欠測値になっているかを調べてみましょう。

1. [**分析**] > [**スクリーニング**] > [**欠測値を調べる**] を選択します。
2. 「**BAD**」から「**DEBTINC**」までを選択し、[**Y, 列**] をクリックします。
3. 表示される警告で、[**OK**] をクリックします。

   「**REASON**」と「**JOB**」は、値のデータタイプが文字型であるため、これらの列は [Y, 列] リストに追加されません。これらの2つの列にどのくらいの欠測値があるかを見るには、「一変量の分布」プラットフォームを使用してください（この例では図示されていません）。
4. [**OK**] をクリックします。

図6.4 欠測値のレポート



「**DEBTINC**」列には1267個の欠測値があり、これはオブザベーション数の約21%に相当します。また、その他のほとんどの列にも欠測値があります。先ほどの例では起動ウィンドウにある［欠測値をカテゴリとして扱う］オプションを用いましたが、その場合、これらの欠測値を含んだデータも分析に使われます。詳細は、「パーティション」章の「欠測値をカテゴリとして扱う」（88ページ）を参照してください。

JMP PRO 連続尺度の変数を扱うブートストラップ森の例

この例では、男性の体脂肪率を予測するブートストラップ森モデルを構築します。

1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Body Fat.jmp」を開きます。
2. ［**分析**］>［**予測モデル**］>［**ブートストラップ森**］を選択します。
3. 「**体脂肪率**」を選択し、［**Y, 目的変数**］をクリックします。
4. 「**年齢**」から「**手首囲**」までを選択し、［**X, 説明変数**］をクリックします。
5. 「**検証**」を選択し、［**検証**］ボタンをクリックします。
6. ［**OK**］をクリックします。
7. （オプション）［**マルチスレッドをオフにする**］を選択し、「乱数シード値」の横に「123」と入力します。

   ブートストラップ森では無作為抽出が行われるため、上記の操作を行うことにより、下の図とまったく同じ結果を得ることができます。
8. ［**OK**］をクリックします。

図6.5 全体の統計量

全体の統計量

個々のツリー	RMSE
バッグ内	2.916888
バッグ外	6.751874

	R2乗	RMSE	N
学習	0.794	3.7179446	180
検証	0.673	4.9794361	72

「全体の統計量」レポートを見ると、「検証」のR2乗が0.673であることがわかります。

次に、モデルに最も寄与している説明変数を見てみましょう。

9. 「ブートストラップ森(体脂肪率)」の横にある赤い三角ボタンをクリックし、[**列の寄与**]をクリックします。

図6.6 列の寄与

列の寄与

項	分岐数	平方和	割合
胴囲(cm)	20	1581.52266	0.2805
体重(ポンド)	10	1308.72038	0.2321
胸囲(cm)	15	977.233672	0.1733
腰囲(cm)	11	505.633363	0.0897
身長(インチ)	13	229.750549	0.0408
手首囲(cm)	11	207.267987	0.0368
首囲	9	188.935633	0.0335
年齢	13	162.704606	0.0289
大腿囲(cm)	8	155.50402	0.0276
上腕囲(伸ばした状態)(cm)	11	122.315188	0.0217
足首囲(cm)	8	71.8606273	0.0127
膝囲(cm)	7	66.7490189	0.0118
前腕囲(cm)	8	59.7237924	0.0106

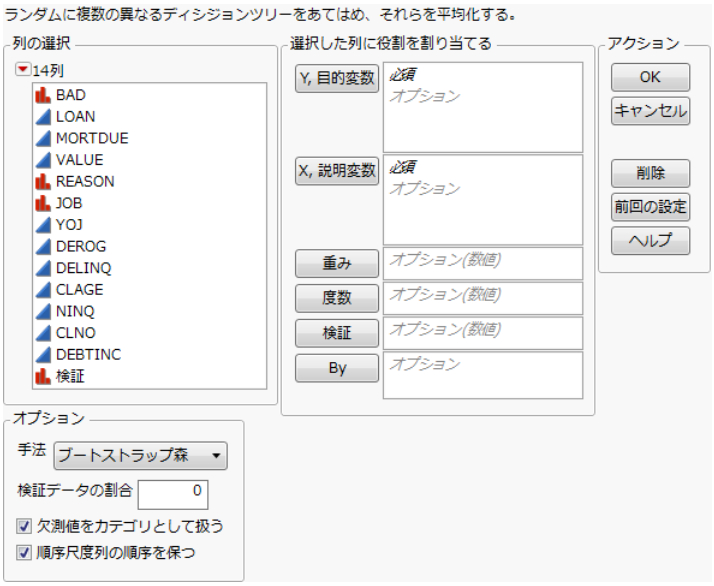
「列の寄与」レポートを見ると、「**胴囲(cm)**」、「**体重(ポンド)**」、および「**胸囲(cm)**」が、「**体脂肪率**」の予測に寄与している説明変数であることがわかります。

JMP PRO「ブートストラップ森」プラットフォームの起動

「ブートストラップ森」プラットフォームを起動するには、[**分析**] > [**予測モデル**] > [**ブートストラップ森**]を選択します。

JMP PRO 起動ウィンドウ

図6.7「ブートストラップ森」起動ウィンドウ



「ブートストラップ森」プラットフォームの起動ウィンドウには、以下のオプションがあります。

Y, 目的変数　分析する目的変数。

X, 説明変数　説明変数。

重み　分析において各行の重みとして使用される数値を含む列。

度数　分析において各行の度数として使用される数値を含む列。

検証　多くとも3つの数値を含む数値列。「パーティション」章の「検証」（93ページ）を参照してください。

By　別々に分析を行いたいときに、そのグループ分けをする変数を指定します。指定された列の各水準ごとに、別々に分析が行われます。各水準の結果は別々のレポートに表示されます。複数のBy変数を割り当てた場合、それらのBy変数の水準の組み合わせごとに別々のレポートが作成されます。

手法　パーティションの手法としてディシジョンツリー、ブートストラップ森、ブースティングツリー、K近傍法、単純Bayesを選択できます。ディシジョンツリー以外の手法は、JMP Proでのみ利用できます

「ブートストラップ森」以外の手法については、第5章「パーティション」、第7章「ブースティングツリー」、第8章「K近傍法」、および第9章「単純Bayes」を参照してください。

検証データの割合　データ全体のうち検証に用いるデータの割合。「パーティション」章の「検証」（93ページ）を参照してください。

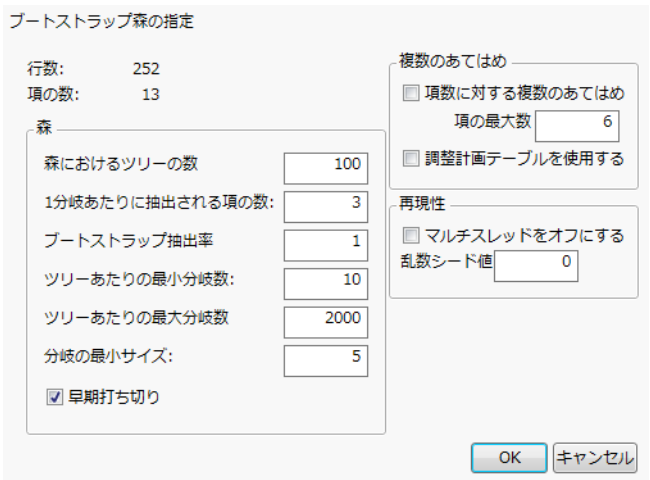
欠測値をカテゴリとして扱う このオプションを選択すると、カテゴリカルな説明変数の場合は、分析において、欠測値が1つのカテゴリとして扱われます。連続尺度の説明変数の場合は、欠測値が同一の数値をもつものとして扱われます。「パーティション」章の「欠測値をカテゴリとして扱う」（88ページ）を参照してください。

順序尺度列の順序を保つ このオプションを選択すると、順序尺度の列において、順序を保つ分岐だけが考慮されるようになります。

JMP PRO 指定ウィンドウ

起動ウィンドウで［**OK**］をクリックすると、「ブートストラップの森の指定」ウィンドウが表示されます。

図6.8「ブートストラップの森の指定」ウィンドウ



JMP PRO 指定パネル

行数 データテーブルの行数。

項の数 説明変数として指定された列の数。

JMP PRO「森」パネル

森におけるツリーの数 作成されるツリーの総数。

1分岐あたりに抽出される項の数 各分岐において、分岐の候補として検討される説明変数の個数。分岐ごとに、分岐の候補としてここに指定した個数だけ、無作為に説明変数が抽出されます。

ブートストラップ抽出率 各ツリーの作成時に抽出するオブザベーションの割合（抽出方法は復元抽出です）。ツリーごとに新しく無作為抽出が行われます。

ツリーあたりの最小分岐数　各ツリーで行う分岐の最小数。

ツリーあたりの最大分岐数　各ツリーで行う分岐の最大数。

分岐の最小サイズ　分岐候補を求めるのに必要とするオブザベーションの最低数。

早期打ち切り　（検証セットを使用している場合のみ。）このオプションを選択すると、ツリーの作成を続けても検証データの適合度統計量がこれ以上改善されない時点で、処理が打ち切られます。この適合度統計量には、応答変数がカテゴリカルな場合には検証セットの「エントロピー R2 乗」値が、応答変数が連続尺度の場合には検証セットの「R2 乗」値が使われます。このオプションを選択しなかった場合は、指定されたツリー数に達するまで処理が続行されます。

JMP PRO「複数のあてはめ」パネル

項数に対する複数のあてはめ　このオプションを選択すると、複数の「1 分岐あたりに抽出される項の数」に対してブートストラップ森が作成されます。レポートに表示される結果は、検証セットの「エントロピー R2 乗」値（応答変数がカテゴリカルな場合）や「R2 乗」値（応答変数が連続尺度の場合）が最大となっているモデルのものです。

下限は「1 分岐あたりに抽出される項の数」で指定されている値です。上限は次のオプションで指定されます。

項の最大数　1 つの分岐に対して考慮される項の最大数。

調整計画テーブルを使用する　ブートストラップ森の設定値を含むデータテーブルを選択するためのウィンドウが表示されます。このような設定が含まれたデータテーブルのことを、JMP では**「調整計画テーブル」**と呼んでいます。この調整計画テーブルには、指定したい各オプションにつき 1 つの列が含まれています。そして、1 つ 1 つのブートストラップ森モデルの設定値を含んだものが、1 行ずつで構成されています。調整計画テーブルでオプションが指定されていない設定は、デフォルト値が使用されます。

JMP は、指定された調整計画テーブルの 1 行につき 1 つのブートストラップ森モデルを作成します。調整計画テーブルで複数のモデルが指定されている場合、「検証セットでのモデル要約」レポートには各モデルの R2 乗値がリストされます。レポート全体には、R2 乗値が最も大きいモデルの結果が表示されます。

調整計画テーブルは実験計画の機能を使って作成してもよいでしょう。調整計画テーブルには、以下の列名をもつ列を含めてください（列の順番は任意です）。列名を英語で指定する場合には、大文字と小文字が区別されます。

- ツリー数（Number Trees）
- 項の数（Number Terms）
- ブートストラップの抽出割合（Portion Bootstrap）
- ツリーあたりの最小分岐数（Minimum Splits per Tree）
- ツリーあたりの最大分岐数（Maximum Splits per Tree）
- 分岐の最小サイズ（Minimum Size Split）

JMP PRO「再現性」パネル

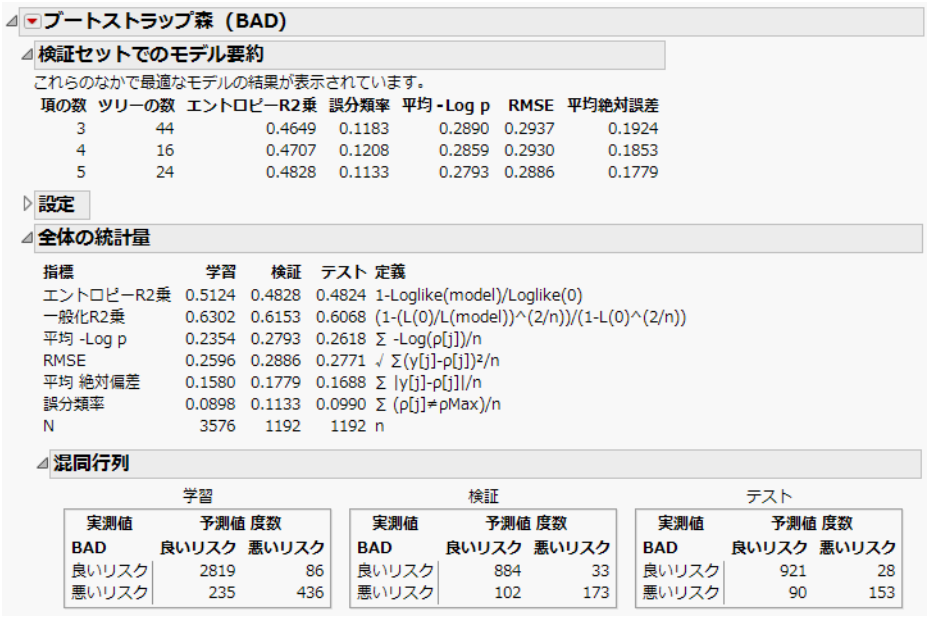
マルチスレッドをオフにする このオプションを選択すると、すべての計算が1つのスレッドで実行されます。

乱数シード値 分析を再実行したときに同じ結果を再現したい場合は、ここにゼロ以外の乱数シード値を指定してください。デフォルトでは、この乱数シード値はゼロとなっており、結果を再現しないように設定されています。分析をスクリプトに保存するとき、ここに入力した乱数シード値がスクリプトに保存されます。

JMP PRO「ブートストラップ森」レポート

「ブートストラップの森の指定」ウィンドウで［OK］をクリックすると、「ブートストラップ森」レポート表示されます。

図6.9 カテゴリカルな応答の「ブートストラップ森」レポート



ブートストラップ森（BAD）

検証セットでのモデル要約

これらのなかで最適なモデルの結果が表示されています。

項の数	ツリーの数	エントロピーR2乗	誤分類率	平均 -Log p	RMSE	平均絶対誤差
3	44	0.4649	0.1183	0.2890	0.2937	0.1924
4	16	0.4707	0.1208	0.2859	0.2930	0.1853
5	24	0.4828	0.1133	0.2793	0.2886	0.1779

設定

全体の統計量

指標	学習	検証	テスト	定義
エントロピーR2乗	0.5124	0.4828	0.4824	1-Loglike(model)/Loglike(0)
一般化R2乗	0.6302	0.6153	0.6068	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
平均 -Log p	0.2354	0.2793	0.2618	Σ -Log(ρ[j])/n
RMSE	0.2596	0.2886	0.2771	√ Σ(y[j]-ρ[j])²/n
平均 絶対偏差	0.1580	0.1779	0.1688	Σ |y[j]-ρ[j]|/n
誤分類率	0.0898	0.1133	0.0990	Σ (ρ[j]≠ρMax)/n
N	3576	1192	1192	n

混同行列

学習

実測値	予測値 度数	
BAD	良いリスク	悪いリスク
良いリスク	2819	86
悪いリスク	235	436

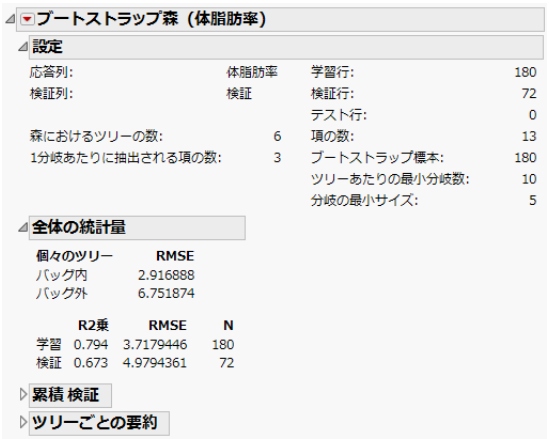
検証

実測値	予測値 度数	
BAD	良いリスク	悪いリスク
良いリスク	884	33
悪いリスク	102	173

テスト

実測値	予測値 度数	
BAD	良いリスク	悪いリスク
良いリスク	921	28
悪いリスク	90	153

図6.10 連続尺度の応答の「ブートストラップ森」レポート



以下のレポートが表示されます。応答変数がカテゴリカルか連続尺度であるかによって、内容は異なります。

- 「検証セットでのモデル要約」（117ページ）
- 「設定」（117ページ）
- 「全体の統計量」（117ページ）
- 「累積 検証」（119ページ）
- 「ツリーごとの要約」（119ページ）

JMP PRO 検証セットでのモデル要約

（「ブートストラップの森の指定」ウィンドウで［項数に対する複数のあてはめ］オプションを選択した場合に使用できます。）すべてのモデルの適合度統計量が表示されます。図6.9および「「複数のあてはめ」パネル」（115ページ）を参照してください。

JMP PRO 設定

モデルをあてはめたときに使用された設定が表示されます。

JMP PRO 全体の統計量

学習セットの適合度統計量が表示されます（検証セットやテストセットを指定した場合、それらの適合度等計量も表示されます）。表示される結果の内容は応答変数の尺度によって異なります。

たとえば、「ブートストラップの森の指定」ウィンドウの［項数に対する複数のあてはめ］オプションを使用して、複数のモデルをあてはめたとしましょう。その場合、「全体の統計量」に結果が表示されるのは、応答変数がカテゴリカルな場合には、検証セットの「エントロピー R2 乗」値が最大となったモデルです。一方、応答変数が連続尺度の場合には、「R2 乗」値が最大となったモデルです。

JMP PRO カテゴリカルな応答

「指標」レポート

学習セットに対して以下の統計量が表示されます（検証セットやテストセットが指定された場合には、それらに対しても同じ統計量が表示されます）。

メモ：「エントロピー R2 乗」と「一般化 R2 乗」は、1 に近いほど適合度が良いことを示します。「平均 -Log p」、「RMSE」、「平均 絶対偏差」、「誤分類率」は、値が小さいほど、適合度が良いことを示します。

エントロピー R2 乗　あてはめたモデルの対数尤度と、切片だけのモデルの対数尤度を比較している指標です。あてはめたモデルの対数尤度を、切片だけのモデルの対数尤度で割り、その値を 1 から引いたものです。この指標の範囲は 0 ～ 1 です。

一般化 R2 乗　この指標は、一般的な回帰モデルに適用できるものです。一般化 R2 乗は、尤度 L から算出され、最大が 1 となるように尺度化されています。完全にモデルがデータにあてはまっている場合は 1、切片だけのモデルと同等なあてはまりの場合には 0 になります。一般化 R2 乗は、通常の R2 乗（正規分布に従う連続尺度の応答変数に対する標準最小 2 乗法の R2 乗）を一般化したものです。この一般化 R2 乗は、「Nagelkerke の R^2」、または「Craig and Uhler の R^2」とも呼ばれており、Cox and Snell の疑似 R^2 を最大が 1 になるように尺度化したものです。詳細は、Nagelkerke（1991）を参照してください。

平均 -Log p　$-\log(p)$ の平均です。p は、実際に生じた応答水準に対する予測確率です。

RMSE　誤差の標準偏差（誤差平方和を自由度で割ったものの平方根）。誤差は $(1-p)$ で計算されます。ここで、p は、実際に生じた応答水準に対する予測確率です。

平均 絶対偏差　誤差の絶対値の平均。誤差は $(1-p)$ で計算されます。ここで、p は、実際に生じた応答水準に対する予測確率です。

誤分類率　予測確率が最も大きい応答の水準が、観測された水準と一致しない割合。

N　オブザベーションの数。

混同行列

（応答がカテゴリカルの場合のみ。）学習セットにおける分類結果の度数が表示されます。検証セットやテストセットが指定された場合には、それらに対する結果も表示されます。

決定行列

（応答がカテゴリカルであり、かつ応答に「利益行列」列プロパティがあるか、または［利益行列の指定］オプションを使用して損失を指定した場合にのみ表示されます。）学習セットに対する「決定行列 度数」と「決定行列 割合」が表示されます。検証セットやテストセットが指定された場合には、それらに対する結果も表示されます。「パーティション」章の「パーティションの別例」（95 ページ）を参照してください。

JMP PRO 連続尺度の応答

「個々のツリー」レポート

「バッグ内」と「バッグ外」のデータから求められたRMSEが表示されます。全体のRMSEは、すべてのツリーにおけるMSEの平均を求め、その平方根をとったものです。ツリーの作成に使われた学習セットは**「バッグ内」**、ツリーの作成に使われなかった学習セットは**「バッグ外」**（OOB; Out-Of-Bag）と呼ばれます。

各ツリーのバッグ外RMSE（OOB RMSE）は、誤差平方和をバッグ外のオブザベーション数で割って、その平方根をとったものです。「ツリーごとの要約」レポートの「OOB SSE/N」で表示されている値は、バッグ外RMSEを2乗した値になっています。

「R2乗」レポートと「RMSE」レポート

学習セットに対して、R2乗、誤差の標準偏差、およびオブザベーション数が表示されます。検証セットやテストセットが指定された場合には、それらに対する結果も表示されます。

JMP PRO 累積 検証

（ここで説明する結果は、検証セットを使用した場合にのみ表示されます。）ツリー数に対して、検証セットの適合度統計量をプロットしたグラフが表示されます。

応答変数が連続尺度の場合は、プロットされる適合度統計量はR2乗だけです。カテゴリカルな変数の場合は、以下に示す適合度統計量がプロットされます。これらの統計量については、「「指標」レポート」（118ページ）で説明しています。

- R2乗（エントロピーR2乗）
- 平均 -Log p
- RMSE
- 平均絶対誤差
- 誤分類率

「累積 検証」プロットの下にある「累積の詳細」レポートには、プロットに使用された統計量が表示されます。

JMP PRO ツリーごとの要約

「ツリーごとの要約」レポートでは、「バッグ内（IB; In-Bag）」および「バッグ外（OOB; Out-Of-Bag）」という考えが使われています。各ツリーについて、ツリーのあてはめに使用されるデータは、復元抽出されます。100%のデータが復元抽出されるように指定した場合は、1回、1回の抽出において1度も抽出されないデータの割合は約$1/e$となります。各ツリーでのあてはめにおいて、これらの1度も抽出されなかったデータは**「バッグ外標本」**と呼ばれます。一方、あてはめに使用されたデータは**「バッグ内標本」**と呼ばれます。

「ツリーごとの要約」レポートには、各ツリーにおける以下の要約統計量が表示されます。

分岐 ディシジョンツリー内の分岐の数。

順位 「OOB 損失」を昇順で並べた、ツリーの順位。「OOB 損失」が最も小さいツリーの順位が1位になっています。

OOB 損失 バッグ外標本から計算された、求められたツリーの予測の不確かさを示す指標。この指標が小さいほど、予測精度が良いことを示します。

OOB 損失/N 「OOB 損失」を、「OOB N」（バッグ外標本の標本サイズ）で割った値。

R2 乗 （応答が連続尺度の場合のみ。）ツリーのR2乗値。

IB SSE （応答が連続尺度の場合のみ。）バッグ内標本から計算された誤差平方和。

IB SSE/N （応答が連続尺度の場合のみ。）バッグ内の誤差平方和を、バッグ内の標本サイズで割ったもの。バッグ内の標本サイズは、学習セットの標本サイズに、「ブートストラップの森の指定」ウィンドウで指定したブートストラップ抽出率を掛けた値です。

OOB N （応答が連続尺度の場合のみ。）バッグ外標本の標本サイズ。

OOB SSE （応答が連続尺度の場合のみ。）バッグ外標本から計算された、求められたツリーの誤差平方和。

OOB SSE/N （応答が連続尺度の場合のみ。）「OOB SSE」を「OOB N」（バッグ外標本の標本サイズ）で割った値。

JMP PRO「ブートストラップ森」プラットフォームオプション

「ブートストラップ森」レポートにある赤い三角ボタンには、次のようなオプションが用意されています。

予測値と実測値のプロット （応答が連続尺度の場合のみ。）予測値と実測値のプロットを作成します。

列の寄与 各列があてはめにどれだけ寄与したかを示すレポートを作成します。このレポートには以下の情報も表示されます。

- 該当の列が分岐に使われた合計回数。
- 応答変数がカテゴリカルな場合には、該当の列による分岐のG^2を合計したもの。連続尺度の場合には、SS（平方和）を合計したもの。
- G^2またはSSの棒グラフ。
- 該当の列によるG^2またはSSの割合。

ツリーの表示 「ツリーの表示」レポートに表示されるツリーの形式に関するオプションが用意されています。オプションのいずれかを選択すると、ブースティングの各層におけるツリーが描かれます。[名前・カテゴリ・推定値の表示] オプションによって表示される確率については、「パーティション」章の「ディシジョンツリーとブートストラップ森の予測確率」（104ページ）を参照してください。

ROC曲線　（応答がカテゴリカルの場合のみ。）「パーティション」章の「ROC曲線」（90ページ）を参照してください。

リフトチャート　（応答がカテゴリカルの場合のみ。）「パーティション」章の「リフトチャート」（91ページ）を参照してください。

列の保存　モデルやツリーの結果を保存するオプション、およびSASコードを作成するオプションがあります。

予測値の保存　モデルの予測値をデータテーブルに保存します。

予測式の保存　予測式をデータテーブルの列に保存します。予測式は、条件節が枝分かれした形で、ツリーの構造を反映しています。応答変数が連続尺度の場合、予測式の列には、［予測対象］プロパティが割り当てられます。カテゴリカルな場合は、［応答確率］プロパティが割り当てられます。

欠測処理予測式の保存　（このオプションの代わりに［予測式の保存］オプションを使用してください。このオプションは［予測式の保存］を利用したくない場合にのみ使用してください。）データに欠測値があり、かつ［欠測値をカテゴリとして扱う］チェックボックスがオフの場合に、欠測値をランダムに処理する予測式を保存します。この予測式では、説明変数が欠測値の場合、どちらに分岐するかがランダムに決められます。応答変数が連続尺度の場合、予測式の列には、［予測対象］プロパティが割り当てられます。カテゴリカルな場合は、［応答確率］プロパティが割り当てられます。［欠測値をカテゴリとして扱う］チェックボックスをオンにした場合は、Shift キーを押しながらレポートの赤い三角ボタンをクリックすると、［欠測処理予測式の保存］を選択できます。

残差の保存　（応答が連続尺度の場合のみ。）残差をデータテーブルに保存します。

累積の詳細の保存　（検証セットを使用している場合のみ。）各ツリーの適合度統計量を含んだデータテーブルを作成します。

予測式を発行　予測式を作成し、それを「計算式デポ」プラットフォームの計算式列のスクリプトとして保存します。「計算式デポ」レポートが開いていない場合は、このオプションによって「計算式デポ」レポートが作成されます。「計算式デポ」章（167ページ）を参照してください。

欠測処理予測式を発行　（このオプションの代わりに［予測式を発行］オプションを使用してください。このオプションは［予測式を発行］を利用したくない場合にのみ使用してください。）欠測処理予測式を作成し、それを「計算式デポ」プラットフォームの計算式列のスクリプトとして保存します。「計算式デポ」レポートが開いていない場合は、このオプションによって「計算式デポ」レポートが作成されます。「計算式デポ」章（167ページ）を参照してください。［欠測値をカテゴリとして扱う］チェックボックスをオンにした場合は、Shift キーを押しながらレポートの赤い三角ボタンをクリックすると、このオプションを使用できます。

SAS DATA ステップの作成　データセットのスコア計算に使用できる SAS コードを作成します。

利益行列の指定　（応答がカテゴリカルの場合のみ。）分類の判定が正しいときや正しくないときの利益や損失を指定できます。「パーティション」章の「「あてはめの詳細」の表示」（84ページ）を参照してください。

プロファイル　「予測プロファイル」を表示します。詳細については、『プロファイル機能』の「プロファイル」章を参照してください。

以下のオプションについて詳しくは、『JMPの使用法』の「JMPのレポート」章を参照してください。

ローカルデータフィルタ 現在のレポートにおいてデータをフィルタリングするローカルデータフィルタを表示するか、非表示にします。

やり直し 分析を繰り返すオプションや、やり直すオプションを含みます。また、［自動再計算］オプションを選択すると、このオプションに対応しているプラットフォームにおいて、データテーブルに加えた変更が、該当するレポートに即座に反映されるようになります。

スクリプトの保存 レポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。

Byグループのスクリプトを保存 By変数の全水準に対するレポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。起動ウィンドウでBy変数を指定した場合のみ使用可能です。

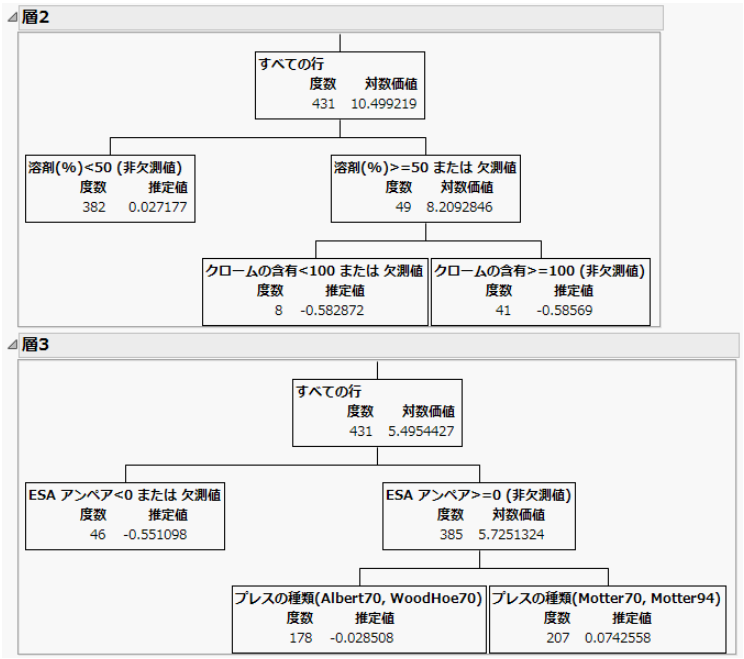
第7章

JMP PRO ブースティングツリー
ツリーを逐次的にあてはめていく

「ブースティングツリー」プラットフォームは JMP Pro でのみ利用できます。

ブースティングツリー（boosted tree）は、複数の小さいツリーを逐次的にあてはめていき、それらの結果を合わせて加法的な大きなモデルを構築する手法です。小さいツリーは、**「層」**と呼ばれています。各層のツリーは、少数の分岐で構成されます。前の層における残差に対してツリーがあてはめられていきます。そのため、各層は、それ以前までの層で生じている残差を小さくしていきます。最終的な予測値は、その残差に対する予測値を、すべての層で合計したものです。

図7.1 ブースティングツリーの層の例



JMP PRO「ブースティングツリー」プラットフォームの概要

「ブースティングツリー」プラットフォームは、多数の小さいツリーをあてはめて**層**を形成し、それに基づいて加法的な大きなモデルを構築します。通常、各層のツリーは5つ以下の分岐で構成されます。各層は、「パーティション」章で解説されている再帰的な手法であてはめられます。ただし、各層において、指定された分岐回数に達した時点で、分岐を止めます。なお、ツリーにおける葉の予測値は、その葉に含まれるすべてのデータの平均です。

ブースティングツリーは、次のようにして求められます。

1. 最初の層をあてはめます。
2. 残差を計算します。残差は、葉に含まれるデータの実測値から、予測値（それらデータの平均）を引いて求めます。
3. 求められた残差に、新たなツリーをあてはめます。
4. 加法的な大きなモデルを構築します。それには、1つ1つのデータ行において、すべての層の予測値を合計します。
5. 指定した層数に達するまで、あるいは、検証セットを使用している場合は層を追加しても検証セットの適合度統計量がこれ以上改善されなくなるまで、手順2～手順4を繰り返します。

最終的な予測値は、すべての層における予測値を合計したものです。

前の層までの残差に対してツリーをあてはめていくことで、モデルの適合度が改善していきます。

カテゴリカルな応答変数に対しては、2水準の応答変数のみがサポートされます。このとき、各層のあてはめでは、オフセット項の和をロジスティック変換したものが使われます。最終的な予測値も、すべての層に関してオフセット項を合計したものをロジスティック変換したものです。

ブースティングツリーの詳細については、Hastie et al.（2009）を参照してください。

JMP PRO カテゴリカルな応答変数に対するブースティングツリーの例

この例では、ブースティングツリーによって、どの印刷物に**「印刷縞」**と呼ばれる不良が生じるかを予測します。

1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Bands Data.jmp」を開きます。
2. ［**分析**］>［**予測モデル**］>［**ブースティングツリー**］を選択します。
3. **「印刷縞の有無」**を選択し、［**Y, 目的変数**］をクリックします。
4. **「Predictors」**列グループを選択し、［**X, 説明変数**］をクリックします。
5. **「検証データの割合」**に「0.2」と入力します。
6. ［**OK**］をクリックします。

   「ブースティングツリー」の設定ウィンドウが開きます。

7. （オプション）「再現性」パネルで、［**マルチスレッドをオフにする**］を選択し、「乱数シード値」として「123」を入力します。

   ここでの例の結果は乱数に依存しますが、この設定により数値結果が以下で紹介するものと同じになります。

8. ［**OK**］をクリックします。

図7.2 名義尺度の目的変数の全体の統計量

全体の統計量

指標	学習	検証	定義
エントロピーR2乗	0.5032	0.2871	1-Loglike(model)/Loglike(0)
一般化R2乗	0.6678	0.4291	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
平均 -Log p	0.3403	0.4663	Σ -Log(p[j])/n
RMSE	0.3202	0.3910	√ Σ(y[j]-p[j])²/n
平均 絶対偏差	0.2604	0.3248	Σ |y[j]-p[j]|/n
誤分類率	0.1230	0.2222	Σ (p[j]≠pMax)/n
N	431	108	n

混同行列

学習

実測値	予測値 度数	
印刷縞の有無	band	noband
band	143	45
noband	8	235

検証

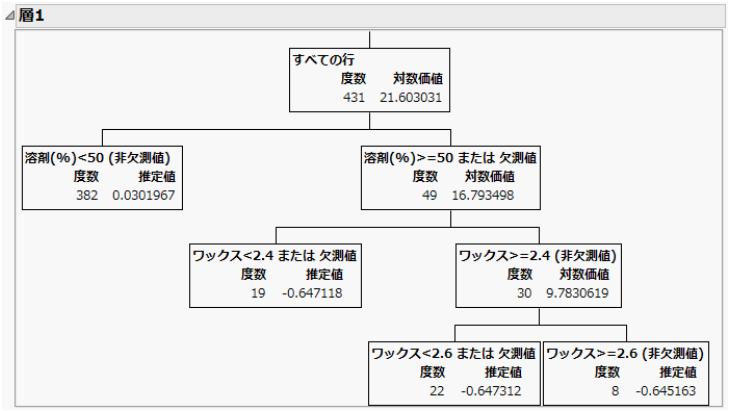
実測値	予測値 度数	
印刷縞の有無	band	noband
band	22	17
noband	7	62

目的変数とする**「印刷縞の有無」**がカテゴリカルであるため、「指標」に「誤分類率」が含まれ、「混同行列」レポートが作成されます。検証セットの誤分類率は0.2222、およそ22%です。

9. **「印刷縞の有無のブースティングツリー」**の赤い三角ボタンをクリックし、［**ツリーの表示**］>［**名前・カテゴリ・推定値の表示**］を選択します。

   「ツリーの表示」レポートが作成され、層ごとにアウトラインが作成されます。各層のアウトラインを開けば、その層であてはめられたツリーと予測値を確認できます。

図7.3 ブースティングツリーの層1



10. **「印刷縞の有無のブースティングツリー」**の赤い三角ボタンをクリックし、［**列の保存**］>［**予測式の保存**］を選択します。

「確率(印刷縞の有無==noband)」、**「確率(印刷縞の有無==band)」**、**「最尤 印刷縞の有無」**という3つの列がデータテーブルに追加されます。**「確率(印刷縞の有無==noband)」**の列を調べ、モデルの予測値が層からどのように計算されたかを確認してみてください。

JMP PRO 連続尺度の応答変数に対するブースティングツリーの例

この例では、ブースティングツリーを作成して、体脂肪率を予測します。説明変数には、名義尺度と連続尺度があります。

1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Body Fat.jmp」を開きます。
2. ［**分析**］>［**予測モデル**］>［**ブースティングツリー**］を選択します。
3. **「体脂肪率」**を選択し、［**Y, 目的変数**］をクリックします。
4. **「年齢」**から**「手首囲(cm)」**までを選択し、［**X, 説明変数**］をクリックします。
5. **「検証」**列を選択し、［**検証**］をクリックします。
6. ［**OK**］をクリックします。
7. ［**OK**］をクリックします。

図7.4 連続尺度の目的変数の全体の統計量

全体の統計量

	R2乗	RMSE	N
学習	0.818	3.5000537	180
検証	0.603	5.4804358	72

「全体の統計量」レポートには、ブースティングツリーモデルのR2乗とRMSEが表示されます。検証セットの「R2乗」は0.603です。検証セットの「RMSE」は約5.48です。

「体脂肪率」を予測するうえで重要な説明変数がどれを探してみましょう。

8. 「体脂肪率のブースティングツリー」の赤い三角ボタンをクリックし、［**プロファイル**］を選択します。
9. 「予測プロファイル」の赤い三角ボタンをクリックし、［**変数重要度の評価**］>［**独立な一様分布の入力**］を選択します。

メモ：［変数重要度の評価］オプションは計算で乱数を用いるために、結果は図7.5とはまったく同じにはなりません。

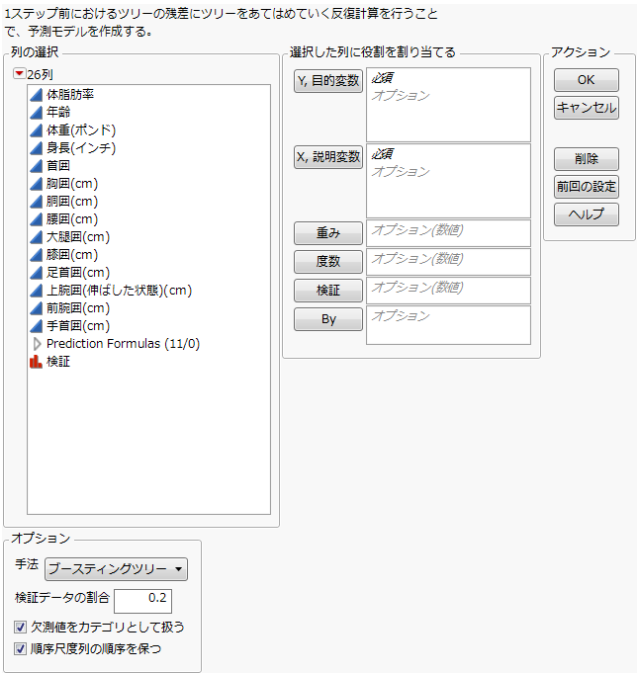
図7.5 変数重要度の要約レポート

「要約レポート」を見ると、**「体脂肪率」**の最も重要な説明変数は**「胴囲(cm)」**であることがわかります。

JMP PRO「ブースティングツリー」プラットフォームの起動

「ブースティングツリー」プラットフォームを起動するには、**[分析]** > **[予測モデル]** > **[ブースティングツリー]**を選択します。

「Body Fat.jmp」を使ったときの「ブースティングツリー」起動ウィンドウ



「ブースティングツリー」プラットフォームの起動ウィンドウには以下のオプションがあります。

Y, 目的変数 分析の対象とする応答変数。

X, 説明変数 説明変数。

重み この役割を割り当てた列の数値は、分析において各行の重みとして使用されます。

度数 この役割を割り当てた列の数値は、分析において各行の度数として使用されます。

検証 多くとも3つの数値を含む数値列。「パーティション」章の「検証」（93ページ）を参照してください。

By 別々に分析を行いたいときに、そのグループ分けをする変数を指定します。指定された列の各水準ごとに、別々に分析が行われます。各水準の結果は別々のレポートに表示されます。複数のBy変数を割り当てた場合、それらのBy変数の水準の組み合わせごとに別々のレポートが作成されます。

手法 パーティションの手法としてディシジョンツリー、ブートストラップ森、ブースティングツリー、K近傍法、単純Bayesを選択できます。ディシジョンツリー以外の手法はJMP Proでのみ利用できます。

「ブースティングツリー」以外の手法については、第5章「パーティション」、第6章「ブートストラップ森」、第8章「K近傍法」、および第9章「単純Bayes」を参照してください。

検証データの割合 データ全体のうち検証に用いるデータの割合。「パーティション」章の「検証」（93ページ）を参照してください。

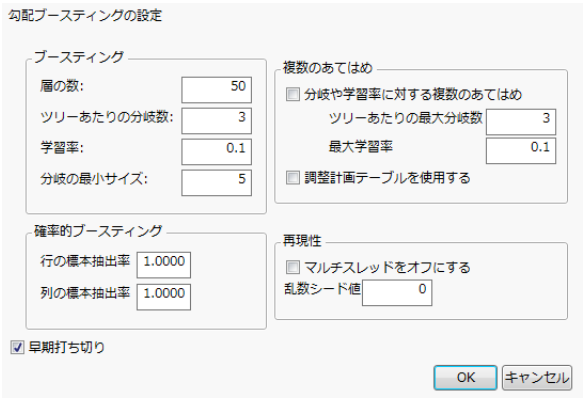
欠測値をカテゴリとして扱う このオプションを選択すると、カテゴリカルな説明変数の場合は、分析において欠測値が1つのカテゴリとして扱われます。連続尺度の説明変数の場合は、欠測値が同一の数値をもつものとして扱われます。「パーティション」章の「欠測値をカテゴリとして扱う」（88ページ）を参照してください。

順序尺度列の順序を保つ このオプションを選択すると、順序尺度の列において、順序を保つ分岐だけが考慮されるようになります。

JMP PRO 設定ウィンドウ

起動ウィンドウで［OK］を選択すると、「勾配ブースティングの設定」というウィンドウが開きます。

図7.6 ブースティングツリーの設定ウィンドウ



JMP PRO「ブースティング」パネル

層の数 最終的なツリーに含める層の最大数。

ツリーあたりの分岐数 各層の分岐の数。

学習率 $0 < r \leq 1$の範囲で設定します。学習率が1に近い値だと、最終モデルへの収束が速くなりますが、データにオーバーフィットしやすくなります。「層の数」に小さい数を指定した場合は、学習率を1に近い値に設定してください。通常、学習率は0.01～0.1の小さな値に設定し、モデルの収束を遅らせます。学習率に小さな値を指定したほうが、前の層とは異なる分岐を、後に続く層が探し出すようになります。

オーバーフィットペナルティ （カテゴリカルな目的変数にのみ使用可能。）予測確率が0になるのを防ぐバイアスパラメータ。「オーバーフィットペナルティ」（136ページ）を参照してください。

分岐の最小サイズ 分岐候補を求めるのに必要とするオブザベーションの最小数。

JMP PRO「複数のあてはめ」パネル

分岐および学習率に対する複数のあてはめ このオプションを選択すると、ツリーあたりの分岐数（増分は整数）と学習率（増分は0.1）のすべての組み合わせに対してブースティングツリーが作成されます。

この組み合わせの下限には、「ツリーあたりの分岐数」と「学習率」に指定された値が使われます。組み合わせの上限には、以下で指定された値が使われます。

ツリーあたりの最大分岐数 ツリーあたりの分岐数の上限。

最大学習率 学習率の上限。

調整計画テーブルを使用する ブースティングツリーの設定値を含むデータテーブルを選択するためのウィンドウが表示されます。このような設定が含まれたデータテーブルのことを、JMPでは**「調整計画テーブル」**と呼んでいます。この調整計画テーブルには、指定したい各オプションにつき1つの列が含まれています。そして、1つ1つのブースティングツリーモデルの設定値を含んだものが、1行ずつで構成されています。調整計画テーブルでオプションが指定されていない設定は、デフォルト値が使用されます。

JMPは、指定された調整計画テーブルの1行につき1つのブースティングツリーモデルを作成します。調整計画テーブルで複数のモデルが指定されている場合、「検証セットでのモデル要約」レポートには各モデルのR2乗値がリストされます。レポート全体には、R2乗値が最も大きいモデルの結果が表示されます。

調整計画テーブルは実験計画の機能を使って作成してもよいでしょう。調整計画テーブルには、以下の列名をもつ列を含めてください（列の順番は任意です）。列名を英語で指定する場合には、大文字と小文字が区別されます。

- 層の数（Number of Layers）
- ツリーあたりの分岐数（Splits per Tree）
- 学習率（Learning Rate）
- 分岐の最小サイズ（Minimum Size Split）
- 行の標本抽出率（Row Sampling Rate）
- 列の標本抽出率（Column Sampling Rate）

JMP PRO「確率的ブースティング」パネル

行の標本抽出率 各層で抽出する学習セットの行の割合。

メモ：応答変数がカテゴリカルな場合は、学習セットからの無作為抽出には層化抽出が使われます。

列の標本抽出率 各層で抽出する説明変数の列の割合。

JMP PRO「再現性」パネル

マルチスレッドをオフにする このオプションを選択すると、すべての計算が単一のスレッドで行われます。

乱数シード値 分析を再実行したときに同じ結果を再現したい場合は、ここにゼロ以外の乱数シード値を指定してください。デフォルトでは、この乱数シード値はゼロとなっており、結果を再現しないように設定されています。分析をスクリプトに保存すると、入力した乱数シード値もスクリプトに保存されます。

JMP PRO 早期打ち切り

早期打ち切り このオプションを選択すると、層の作成を続けても検証データの適合度統計量がこれ以上改善されない時点で、処理が打ち切られます。選択しなかった場合は、指定された層数に達するまで処理が続行されます。このオプションは、検証データを使用している場合のみ表示されます。

JMP PRO「ブースティングツリー」レポート

「勾配ブースティングの設定」ウィンドウで［OK］をクリックすると、「ブースティングツリー」レポートが作成されます。

図7.7 連続尺度の目的変数の「ブースティングツリー」レポート

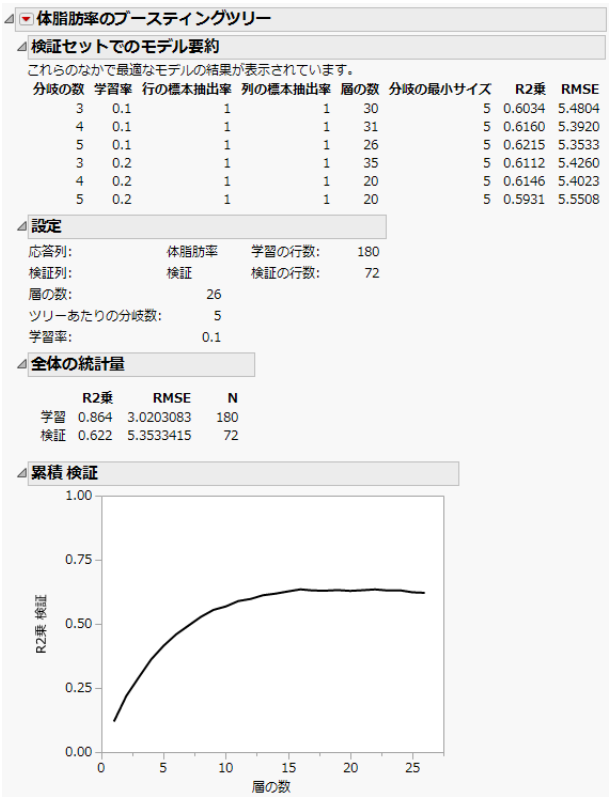
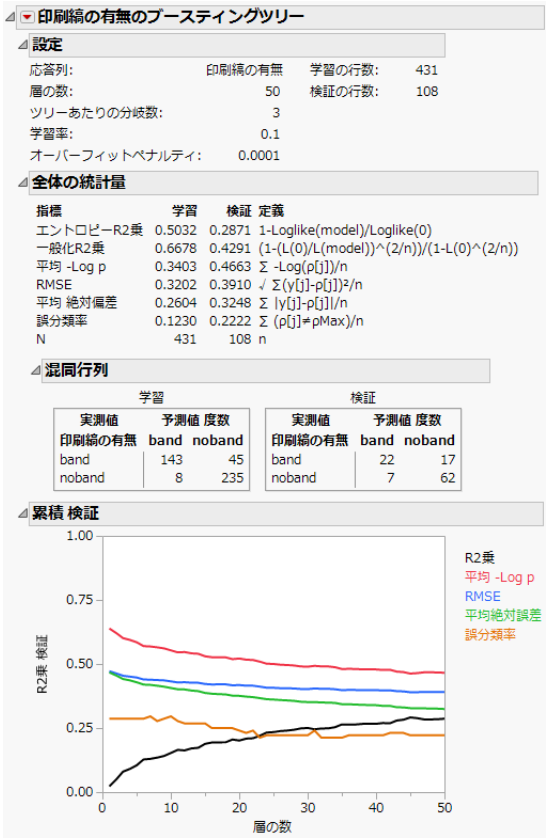


図7.8 カテゴリカルな目的変数の「ブースティングツリー」レポート



以下のレポートが表示されます。応答変数がカテゴリカルか連続尺度であるかによって、内容は異なります。

- 「検証セットでのモデル要約」（132ページ）
- 「設定」（133ページ）
- 「全体の統計量」（133ページ）
- 「累積 検証」（134ページ）

JMP PRO 検証セットでのモデル要約

設定ウィンドウで［分岐および学習率に対する複数のあてはめ］オプションを選択した場合には、それらのすべてのモデルに対して、適合度統計量が表示されます。図7.7および「「複数のあてはめ」パネル」（129ページ）を参照してください。

JMP PRO 設定

モデルをあてはめたときに使用された設定が表示されます。

JMP PRO 全体の統計量

学習セットの適合度統計量が表示されます（検証セットやテストセットを指定した場合、それらの適合度等計量も表示されます）。

たとえば、「ブートストラップの森の指定」ウィンドウの［項数に対する複数のあてはめ］オプションを使用して、複数のモデルをあてはめたとしましょう。その場合、「全体の統計量」に結果が表示されるのは、応答変数がカテゴリカルな場合には、検証セットの「エントロピー R2 乗」値が最大となったモデルです。一方、応答変数が連続尺度の場合には、「R2 乗」値が最大となったモデルです。

JMP PRO「指標」レポート

（応答変数がカテゴリカルな場合のみ。）学習セットに対して以下の統計量が表示されます（検証セットやテストセットが指定された場合には、それらに対しても同じ統計量が表示されます）。

メモ：「エントロピー 2 乗」と「一般化 R2 乗」は、値が 1 に近いほど、適合度が良いことを示します。「平均 -Log p」、「RMSE」、「平均 絶対偏差」、「誤分類率」は、値が小さいほど、適合度が良いことを示します。

エントロピー R2 乗　あてはめたモデルの対数尤度と、切片だけのモデルの対数尤度を比較する指標です。あてはめたモデルの対数尤度を、切片だけのモデルの対数尤度で割り、その値を 1 から引いたものです。エントロピー R2 乗は 0 ～ 1 の値を取ります。

一般化 R2 乗　この指標は、一般的な回帰モデルに適用できるものです。一般化 R2 乗は、尤度 L から算出され、最大が 1 となるように尺度化されています。完全にモデルがデータにあてはまっている場合は 1、切片だけのモデルと同等なあてはまりの場合には 0 になります。一般化 R2 乗は、通常の R2 乗（正規分布に従う連続尺度の応答変数に対する標準最小 2 乗法の R2 乗）を一般化したものです。この一般化 R2 乗は、「Nagelkerke の R^2」、または「Craig and Uhler の R^2」とも呼ばれており、Cox and Snell の疑似 R^2 を最大が 1 になるように尺度化したものです。詳細は、Nagelkerke（1991）を参照してください。

平均 -Log p　$-\log(p)$ の平均。p は実際に生じた応答水準に対する予測確率です。

RMSE　誤差の標準偏差（誤差平方和を自由度で割ったものの平方根）。誤差は $(1-p)$ で計算されます。ここで、p は実際に生じた応答水準に対する予測確率です。

平均 絶対偏差　誤差の絶対値の平均。誤差は $(1-p)$ で計算されます。ここで、p は実際に生じた応答水準に対する予測確率です。

誤分類率　予測確率が最も大きい応答の水準が、観測された水準と一致しない割合。

N　オブザベーションの数。

JMP PRO 混同行列

（応答がカテゴリカルの場合のみ。）学習セットにおいて分類した結果の度数が表示されます。検証セットやテストセットが指定された場合には、それらに対する結果も表示されます。

JMP PRO 決定行列

（応答変数がカテゴリカルで、利益行列の列プロパティを持つ場合、または［利益行列の指定］オプションを使って利益を指定した場合のみ。）学習セットに対する「決定行列 度数」と「決定行列 割合」が表示されます。検証セットやテストセットが指定された場合には、それらに対する結果も表示されます。「パーティション」章の「パーティションの別例」（95ページ）を参照してください。

JMP PRO 累積 検証

（ここで説明する結果は、検証セットを使用した場合にのみ表示されます。）層の数に対して、検証セットの適合度統計量をプロットしたグラフが表示されます。

応答変数が連続尺度の場合は、プロットされる適合度統計量はR2乗だけです。カテゴリカルな変数の場合は、以下に示す適合度統計量がプロットされます（詳細は「「指標」レポート」（133ページ）を参照）。

- R2乗（エントロピーR2乗）
- 平均 - Log p
- RMS誤差（RMSE）
- 平均絶対誤差（平均 絶対偏差）
- 誤分類率

「累積 検証」プロットの下にある「累積の詳細」レポートには、プロットに使用された統計量が表示されます。

JMP PRO「ブースティングツリー」プラットフォームのオプション

「ブースティングツリー」レポートにある赤い三角ボタンをクリックすると、次のようなオプションが表示されます。

ツリーの表示 「ツリーの表示」レポートに表示するツリーのオプション。ブースティングの各層におけるツリー構造の図が表示されます。

予測値と実測値のプロット （応答変数が連続尺度の場合のみ。）予測値と実測値のプロットを作成します。

列の寄与 各列があてはめにどれだけ寄与したかを示すレポートを作成します。このレポートには、以下の情報も表示されます。

- 該当の列が分岐に使われた合計回数。
- 応答変数がカテゴリカルな場合には、該当の列による分岐の G^2 を合計したもの。連続尺度の場合には、SS（平方和）を合計したもの。
- G^2 または SS の棒グラフ。
- 該当の列による G^2 または SS の割合。

ROC 曲線 （応答変数がカテゴリカルの場合のみ。）「パーティション」章の「ROC 曲線」（90 ページ）を参照してください。

リフトチャート （応答変数がカテゴリカルの場合のみ。）「パーティション」章の「リフトチャート」（91 ページ）を参照してください。

列の保存 モデルやツリーの結果を保存するオプション、SAS コードを作成するオプションがあります。

予測値の保存 モデルの予測値をデータテーブルに保存します。

予測式の保存 予測式をデータテーブルに保存します。予測式は、条件節が枝分かれした形で、ツリーの構造を反映しています。応答変数が連続尺度の場合、予測式の列には、［予測対象］プロパティが割り当てられます。カテゴリカルな場合は、［応答確率］プロパティが割り当てられます。

欠測処理予測式の保存 （このオプションの代わりに［予測式の保存］オプションを使用してください。このオプションは［予測式の保存］を利用したくない場合にのみ使用してください。）データに欠測値があり、かつ［欠測値をカテゴリとして扱う］チェックボックスがオフの場合に、欠測値をランダムに処理する予測式を保存します。この予測式では、説明変数が欠測値の場合、どちらに分岐するかがランダムに決められます。応答変数が連続尺度の場合、予測式の列には、［予測対象］プロパティが割り当てられます。カテゴリカルな場合は、［応答確率］プロパティが割り当てられます。［欠測値をカテゴリとして扱う］チェックボックスをオンにした場合は、Shift キーを押しながらレポートの赤い三角ボタンをクリックすると、［欠測処理予測式の保存］を選択できます。

残差の保存 （応答変数が連続尺度の場合のみ。）残差をデータテーブルに保存します。

オフセット推定値の保存 （応答変数がカテゴリカルの場合のみ。）線形成分の和を保存します。これらは、予測確率のロジットです。

ツリーの詳細の保存 各層における分岐の詳細と推定値を含んだデータテーブルを作成します。

累積の詳細の保存 （検証セットを使用している場合のみ。）各層の適合度統計量を含んだデータテーブルを作成します。

予測式を発行 予測式を作成し、それを「計算式デポ」プラットフォームの計算式列のスクリプトとして保存します。「計算式デポ」レポートが開いていない場合は、このオプションによって「計算式デポ」レポートが作成されます。「計算式デポ」章（167 ページ）を参照してください。

欠測処理予測式を発行 （このオプションの代わりに［予測式を発行］オプションを使用してください。このオプションは［予測式を発行］を利用したくない場合にのみ使用してください。）欠測処理予測式を作成し、それを「計算式デポ」プラットフォームの計算式列のスクリプトとして保存します。「計算式デポ」レポートが開いていない場合は、このオプションによって「計算式デポ」レポートが作成されます。「計算式デポ」章（167 ページ）を参照してください。［欠測値をカテゴリとして扱う］チェックボックスをオンにした場合は、Shift キーを押しながらレポートの赤い三角ボタンをクリックすると、このオプションを使用できます。

SAS データステップの作成 データセットのスコア計算に使用できる SAS コードを作成します。

利益行列の指定 （応答変数がカテゴリカルの場合のみ。）分類の判定が正しいときや正しくないときの利益や損失を指定できます。「パーティション」章の「「あてはめの詳細」の表示」（84 ページ）を参照してください。

プロファイル 予測プロファイルを表示します。詳細については、『プロファイル機能』の「プロファイル」章を参照してください。

以下のオプションについて詳しくは、『JMP の使用法』の「JMP のレポート」章を参照してください。

ローカルデータフィルタ 現在のレポートにおいてデータをフィルタリングするローカルデータフィルタを表示するか、非表示にします。

やり直し 分析を繰り返すオプションや、やり直すオプションを含みます。また、［自動再計算］オプションを選択すると、このオプションに対応しているプラットフォームにおいて、データテーブルに加えた変更が、該当するレポートに即座に反映されるようになります。

スクリプトの保存 レポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。

By グループのスクリプトを保存 By 変数の全水準に対するレポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。起動ウィンドウで By 変数を指定した場合のみ使用可能です。

JMP PRO 「ブースティングツリー」プラットフォームの統計的詳細

この節では、「ブースティングツリー」プラットフォームに特有の計算についてだけ説明します。再帰的なディシジョンツリー（決定木）の計算に関する詳細は、「パーティション」章の「統計的詳細」（103 ページ）を参照してください。

JMP PRO オーバーフィットペナルティ

応答変数がカテゴリカルな場合のパラメータ推定では、ペナルティ（罰則）を課しています。負の対数尤度にペナルティを加算したものを目的関数として、それを最小化しています。このペナルティは、各オブザベーションにおける推定値の平方和に、定数を掛けたものです。ペナルティを課すことにより、各層において、学習データへオーバーフィット（過学習）することを防いでいます。

第8章

JMP PRO K近傍法
近くにあるデータで応答を予測する

「K近傍法」プラットフォームはJMP Proでのみ利用できます。

「K近傍法」プラットフォームは、あるデータ行の応答値を、その近くに位置するデータの応答値で予測します。応答変数がカテゴリカルな場合には「分類」を、応答変数が連続尺度の場合には「予測」を行います。

K近傍法は、近くにあるデータとの距離に基づくノンパラメトリックな手法です。そのため、K近傍法では説明変数と応答変数との関係が複雑な場合でも使えます。しかし、K近傍法は関係のない説明変数にも予測が影響を受けやすいため、実行する前に説明変数を選択しておく必要があります。

K近傍法は、衛星写真や心電図の分類といったいろいろな分野で役立ちます。

図8.1「K近傍法」プラットフォームの例

K近傍法

BAD

学習セット

K	度数	誤分類率	誤分類	
1	3576	0.06432	230	*
2	3576	0.08809	315	
3	3576	0.10263	367	
4	3576	0.11885	425	
5	3576	0.12724	455	
6	3576	0.13647	488	
7	3576	0.13870	496	
8	3576	0.14234	509	
9	3576	0.14178	507	
10	3576	0.14541	520	

検証セット

K	度数	誤分類率	誤分類	
1	1192	0.07131	85	*
2	1192	0.11493	137	
3	1192	0.12248	146	
4	1192	0.13674	163	
5	1192	0.14597	174	
6	1192	0.16023	191	
7	1192	0.16527	197	
8	1192	0.16443	196	
9	1192	0.16946	202	
10	1192	0.17198	205	

テストセット

K	度数	誤分類率	誤分類	
1	1192	0.05789	69	*
2	1192	0.08557	102	
3	1192	0.10738	128	
4	1192	0.12332	147	
5	1192	0.13003	155	
6	1192	0.14094	168	
7	1192	0.14597	174	
8	1192	0.14849	177	
9	1192	0.15017	179	
10	1192	0.15688	187	

最良の混同行列 K=1

学習セット

実測値	予測値 度数	
BAD	Good Risk	Bad Risk
Good Risk	2894	11
Bad Risk	219	452

検証セット

実測値	予測値 度数	
BAD	Good Risk	Bad Risk
Good Risk	917	0
Bad Risk	85	190

テストセット

実測値	予測値 度数	
BAD	Good Risk	Bad Risk
Good Risk	949	0
Bad Risk	69	174

JMP PRO「K近傍法」プラットフォームの概要

「K近傍法」プラットフォームは、最も近くにあるk行の応答値に基づいて予測します。ある行に最も近いk行を特定するときには、まず、その行の説明変数と、その行以外の説明変数とのユークリッド距離を求めます。そして、そのユークリッド距離が最短となっているk行を特定します。応答変数が連続尺度の場合は、最も近いk行における応答変数の平均を予測値とします。また、応答変数がカテゴリカルな場合は、最も近いk行で最も頻繁に出現している応答水準を予測値とします。最も頻繁に出現している水準が複数あるときは、そこから無作為に選んだ水準を予測値とします。

メモ: カテゴリカルな変数では、最も頻繁に出現する水準が複数ある場合には、そこから1つが無作為に選択されるため、プラットフォームを何度か実行したときに、それらの結果が異なる可能性があります。スクリプトでは、「K近傍法」プラットフォームを呼び出すメッセージ内に、`Nonrandom`というキーワードを追加すれば、再現可能な結果が得られます。

連続尺度の説明変数は、各説明変数の標準偏差によって尺度化されます。この尺度化により、各説明変数が距離の計算に与える影響を等しくします。なお、連続尺度の説明変数にある欠測値は、その説明変数の平均によって補完されます。

カテゴリカルな各説明変数は、各水準の出現を表す指示変数に変換されます。なお、カテゴリカルな説明変数が欠測値である行は、すべての指示変数の値をゼロにします。

*K*近傍法には、次のような欠点があります。

- K近傍法は、巨大なデータに対して実用的な予測式を作成することができません。
- K近傍法は、カテゴリカルな応答に対して予測確率を求めることができません。

*K*近傍法の詳細については、Hastie et al.（2009）、Hand et al.（2001）、Shmueli et al.（2017）を参照してください。

JMP PRO カテゴリカルな応答に対するK近傍法の例

住宅担保ローンに応募した5,960人の顧客の資産状況に関する履歴データがあります。各顧客は「良いリスク」と「悪いリスク」に分類されています。ほとんどの説明変数に欠測値があります。未来の顧客のクレジットリスクを分類するためのモデルを作成してみましょう。

1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Equity.jmp」を開きます。
2. ［**分析**］>［**予測モデル**］>［**K近傍法**］を選択します。
3. 「**BAD**」を選択し、［**Y, 目的変数**］をクリックします。

   説明変数の1つである「**DEBTINC**」は、欠測値が多いです。何らかの情報が得られる可能性もありますが、ここではモデルに含めないことにします。
4. 「**LOAN**」から「**CLNO**」までを選択し、［**X, 説明変数**］をクリックします。

5. 「**Validation**」列を選択し、［**検証**］をクリックします。
6. ［**OK**］をクリックします。

図8.2「K近傍法」レポート

K近傍法

BAD

学習セット

K	度数	誤分類率	誤分類	
1	3576	0.06432	230	*
2	3576	0.08809	315	
3	3576	0.10263	367	
4	3576	0.11885	425	
5	3576	0.12724	455	
6	3576	0.13647	488	
7	3576	0.13870	496	
8	3576	0.14234	509	
9	3576	0.14178	507	
10	3576	0.14541	520	

検証セット

K	度数	誤分類率	誤分類	
1	1192	0.07131	85	*
2	1192	0.11493	137	
3	1192	0.12248	146	
4	1192	0.13674	163	
5	1192	0.14597	174	
6	1192	0.16023	191	
7	1192	0.16527	197	
8	1192	0.16443	196	
9	1192	0.16946	202	
10	1192	0.17198	205	

テストセット

K	度数	誤分類率	誤分類	
1	1192	0.05789	69	*
2	1192	0.08557	102	
3	1192	0.10738	128	
4	1192	0.12332	147	
5	1192	0.13003	155	
6	1192	0.14094	168	
7	1192	0.14597	174	
8	1192	0.14849	177	
9	1192	0.15017	179	
10	1192	0.15688	187	

最良の混同行列 K=1

学習セット

実測値	予測値 度数	
BAD	Good Risk	Bad Risk
Good Risk	2894	11
Bad Risk	219	452

検証セット

実測値	予測値 度数	
BAD	Good Risk	Bad Risk
Good Risk	917	0
Bad Risk	85	190

テストセット

実測値	予測値 度数	
BAD	Good Risk	Bad Risk
Good Risk	949	0
Bad Risk	69	174

Kの各値ごとに、学習セットで近傍のK行を用いたモデルが作成されます。そして、その作成された各モデルに対して、検証セットのデータが分類されます。この例において、「検証セット」の誤分類率が最も低いのは1つの近傍点（K = 1）に基づいたモデルです。また、「テストセット」でも、1つの近傍点（K = 1）に基づいたモデルが誤分類率が最も低いです。

7. 「K近傍法」の赤い三角ボタンをクリックし、［**予測式を発行**］を選択します。
8. ［**近傍点の個数 , K**］として「1」を入力します。

   この操作により、予測式が計算式デポに保存されます。計算式デポを使えば、さまざまなモデルもあてはめてみて、それらの予測精度を［モデルの比較］で比較できます。

JMP PRO 連続尺度の応答に対するK近傍法の例

この例では、13個の説明変数を使って男性の体脂肪率を予測します。「Body.jmp」データテーブルには、体脂肪率の推定値が含まれています（この体脂肪率の推定値は、水中で測定した体重と身体各部の周囲長に基づいて算出されています）。

1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Body Fat.jmp」を開きます。
2. ［**分析**］>［**予測モデル**］>［**K近傍法**］を選択します。
3. 「**体脂肪率**」を選択し、［**Y, 目的変数**］をクリックします。
4. 「**年齢**」から「**手首囲 (cm)**」までを選択し、［**X, 説明変数**］をクリックします。

5. 「**検証**」列を選択し、[**検証**] をクリックします。
6. [**OK**] をクリックします。

図8.3「K近傍法」レポート

K近傍法

体脂肪率

学習セット

K	度数	RMSE	SSE	
1	180	6.5710	7772.11	
2	180	5.5842	5612.94	
3	180	5.5228	5490.23	
4	180	5.3852	5220.16	
5	180	5.2984	5053.22	
6	180	5.1630	4798.2	
7	180	5.2055	4877.47	
8	180	5.1432	4761.53	*
9	180	5.2274	4918.7	
10	180	5.2199	4904.48	

検証セット

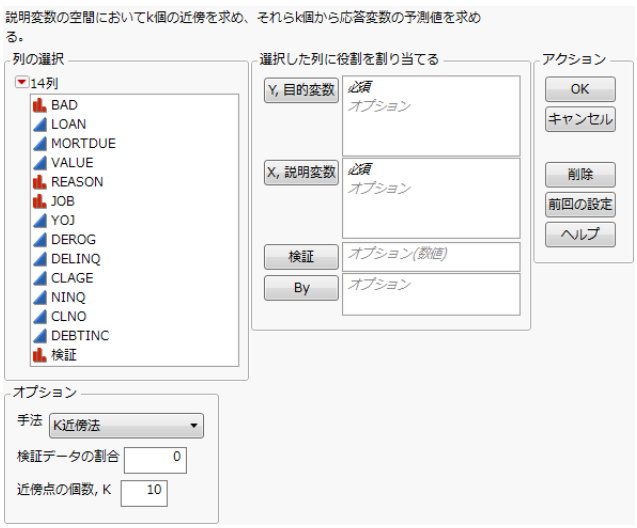
K	度数	RMSE	SSE	
1	72	7.8453	4431.56	
2	72	6.6167	3152.24	
3	72	5.9508	2549.69	
4	72	5.6600	2306.56	
5	72	5.6957	2335.73	
6	72	5.6412	2291.28	
7	72	5.6392	2289.68	
8	72	5.5460	2214.61	*
9	72	5.6555	2302.89	
10	72	5.6406	2290.81	

「検証セット」における「RMSE」の値が最も小さくなっているモデルは、K = 8のモデルです。K近傍法モデルのなかでは、8個の近傍点に基づくモデルが最も予測精度が高いようです。

JMP PRO「K近傍法」プラットフォームの起動

「K近傍法」プラットフォームを起動するには、[**分析**] > [**予測モデル**] > [**K近傍法**] を選択します。

図8.4「K近傍法」の起動ウィンドウ



「K近傍法」の起動ウィンドウには、以下のオプションがあります。

Y, 目的変数　分析の対象とする応答変数。

X, 説明変数　説明変数。

検証　多くとも3つの数値を含む数値列。「パーティション」章の「検証」（93ページ）を参照してください。

By　別々に分析を行いたいときに、そのグループ分けをする変数を指定します。指定された列の各水準ごとに、別々に分析が行われます。各水準の結果は別々のレポートに表示されます。複数のBy変数を割り当てた場合、それらのBy変数の水準の組み合わせごとに別々のレポートが作成されます。

方法　パーティションの手法としてディシジョンツリー、ブートストラップ森、ブースティングツリー、K近傍法、単純Bayesを選択できます。ディシジョンツリー以外の手法は、JMP Proでのみ利用できます。

「K近傍法」以外の手法については、第5章「パーティション」、第6章「ブートストラップ森」、第7章「ブースティングツリー」、第9章「単純Bayes」を参照してください。

検証データの割合　データ全体のうち検証に用いるデータの割合。「パーティション」章の「検証」（93ページ）を参照してください。

近傍点の個数, K　近傍点の最大数。1個から、指定した個数までのモデルがあてはめられます。

JMP PRO「K近傍法」レポート

「K近傍法」プラットフォームの結果は、応答ごとに、応答変数の列名をタイトルとしたレポートとなっており、そこにモデルの情報が表示されます。この結果は、あてはめた*K*個のモデルの要約になっています。結果は、学習セットと検証セットに分かれています（検証セットの表が表示されるのは、検証セットを使用している場合のみです）。

算出される統計量は、応答変数の尺度によって異なります。*K*の値として1から［近傍点の個数, *K*］で指定した値までモデルが作成され、結果における表の各行にその統計量が表示されます。

JMP PRO 連続尺度の応答

アスタリスクがついているモデルは、RMSE が最小となっているモデルです。応答変数が連続尺度の場合には、以下の統計量が算出されます。

K　モデルに使用された近傍点の個数。*K*の値として1から［近傍点の個数, K］で指定した値までモデルが作成されます。

度数　モデルのあてはめに使用されたオブザベーションの数。

RMSE　モデルにおける誤差の標準偏差。RMSEが最小となっているモデルにはアスタリスクがつきます。なお、RMSEが同じで最小となっているモデルが複数ある場合は、*K*が最小であるモデルにアスタリスクがつきます。

SSE　モデルの誤差平方和。

JMP PRO カテゴリカルな応答に対するオプション

要約表

アスタリスクがついているモデルは、誤分類率が最小となっているモデルです。カテゴリカルな応答の要約表には、以下の統計量が算出されています。

K モデルに使用された近傍点の個数。*K*の値として1から［近傍点の個数, K］で指定した値までモデルが作成されます。

度数 モデルのあてはめに使用されたオブザベーションの数。

誤分類率 モデルによって誤分類されたオブザベーションの割合。誤分類された度数を、全体の度数で割ったものです。誤分類率の最小となっているモデルにはアスタリスクがつきます。誤分類率が同じで最小となっているモデルが複数ある場合は、*K*が最小であるモデルにアスタリスクがつきます。

誤分類 モデルによって正しく予測されなかったオブザベーションの数。

混同行列

誤分類率が最小となっているモデル（複数ある場合はその中で*K*が最小であるモデル）の混同行列が表示されます。検証セットやテストセットを使用した場合、それらの混同行列も表示されます。混同行列は、応答の実測値と予測値を2元度数表にまとめたものです。モデルを選択するときは、混同行列や誤分類率を参考にしてください。

「K近傍法」プラットフォームのオプション

「K近傍法」レポートの赤い三角ボタンのメニューには、次のようなオプションがあります。

予測値の保存 予測値の列を*K*個、データテーブルに保存します。列の名前は**「予測値 <Y, 応答変数> <k>」**になります。*k*番目の列には、*k*個の近傍点から計算したモデルの予測値が保存されます。「<Y, 応答変数>」は応答列の名前を指します。

近傍行の保存 *K*個の列をデータテーブルに保存します。列の名前は**「近傍行 <k>」**になります。ある行の*k*番目の列にある値は、その行の*k*番目の近傍行の番号です。

注意：データテーブル内で行を並べ替えても、列**「近傍行 <k>」**の行番号は更新されません。行を並べ替えた場合、これらの列の値は意味がなくなります。

予測式の保存 指定の*k*近傍法モデルの予測式を含む列を保存します。*k*の値を入力してください。予測式にすべての学習データが含まれるため、データテーブルの大きさによってはこのオプションが実用的でない場合があります。

注意：［予測式の保存］で得た値と［予測値の保存］で得た値は、必ずしも一致しません。

予測式を発行 指定のkでK近傍法モデルの予測式を作成し、それを「計算デポ」プラットフォームに計算式列のスクリプトとして保存します。「計算式デポ」レポートが開いていない場合は、このオプションによって「計算式デポ」レポートが作成されます。『予測モデルおよび発展的なモデル』の「計算式デポ」章を参照してください。

以下のオプションについて詳しくは、『JMPの使用法』の「JMPのレポート」章を参照してください。

ローカルデータフィルタ 現在のレポートにおいてデータをフィルタリングするローカルデータフィルタを表示するか、非表示にします。

やり直し 分析を繰り返すオプションや、やり直すオプションを含みます。また、［自動再計算］オプションを選択すると、このオプションに対応しているプラットフォームにおいて、データテーブルに加えた変更が、該当するレポートに即座に反映されるようになります。

スクリプトの保存 レポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。

Byグループのスクリプトを保存 By変数の全水準に対するレポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。起動ウィンドウでBy変数を指定した場合のみ使用可能です。

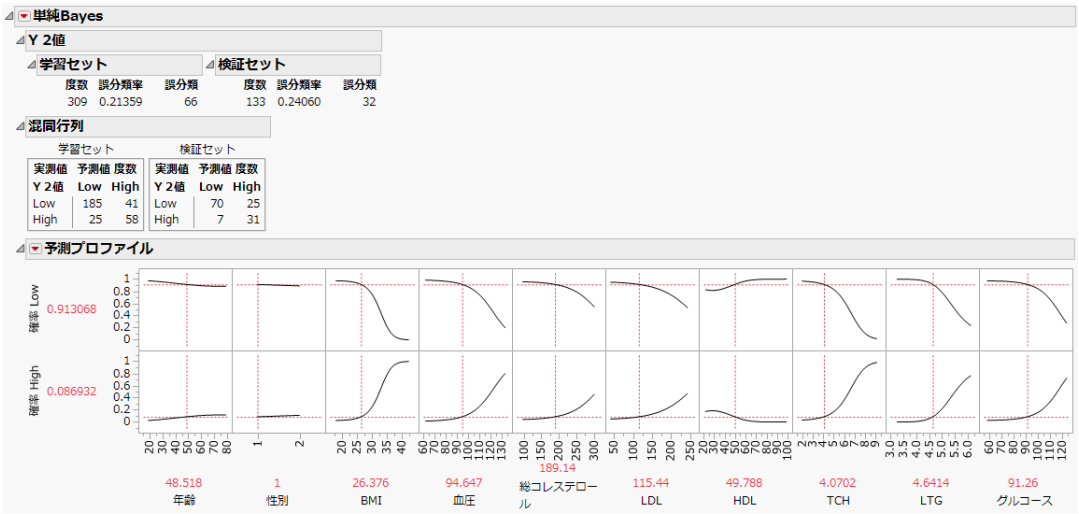
第9章

JMP PRO 単純Bayes
ベイズ定理にもとづく分類

「単純 Bayes」プラットフォームは JMP Pro でのみ利用できます。

「単純Bayes」プラットフォームは、カテゴリカルな変数を予測するモデルをあてはめます。単純Bayesモデル（naive Bayes model; ナイーブベイズモデル）は、計算式が単純で、計算するのに時間がかからない手法です。説明変数の個数が多い場合に特に適しています。

図9.1「単純Bayes」分析の例



JMP PRO「単純 Bayes」プラットフォームの概要

「単純 Bayes」プラットフォームは、カテゴリカルな応答変数の水準のいずれかに、各オブザベーションを分類します。この「カテゴリカルな応答変数の水準」は、**「クラス」**とも呼ばれています。また、分類に使用される説明変数は、データマイニングに関する文献では、**「特徴（feature）」**とも呼ばれています。

単純 Bayes 法は、特徴で条件付けたときの、各クラスに属する条件付き確率（事後確率）を計算します。なお、特徴が連続尺度の場合は、1 変量正規分布の密度関数が使われます。単純 Bayes 法では、「各クラスで条件付けたときに、特徴が互いに独立である」と仮定します（このような単純な仮定を置いているので、この手法は「単純（ナイーブ）」と呼ばれています）。各オブザベーションは、特徴で条件付けたときの条件付き確率（事後確率）が最大となっているクラスに分類されます。Hastie et al.（2001）を参照してください。

単純 Bayes 法は、1 変量の密度関数や確率しか計算に使わないため、計算時間がとても短くてすみます。そのため、大規模データや、特徴数の多いデータに向いています。説明変数が欠測値となっていない行に対して、事後確率が計算されます。

各データ行に対して、各クラスの**スコア**が計算されます。このスコアは、学習セットにおいて該当のクラスに属しているものの割合に、該当のクラスで条件付けたときの各特徴の条件付き確率の総積を掛けたものです。特徴で条件付けたときのあるクラスに属する**条件付き確率（事後確率）**は、該当のクラスのスコアを、全クラスのスコアの和で割ったものです。そして、各オブザベーションは、この事後確率が最大となっているクラスに分類されます。

注意：単純 Bayes 法では、クラスで条件付けたときの特徴の条件付き確率において、独立性を仮定しています。そのため、単純 Bayes 法による分類は精度が高くありません。

一般的には、学習データに含まれている特徴値やクラスに対して分類は行われます。しかし、JMP の「単純 Bayes」プラットフォームにおいては、検証セットには含まれているが、学習セットには含まれていない特徴値がある場合、それらの特徴値を含んだ予測式が作成されます。ただし、一度、保存された予測式は、新たに追加された特徴値を処理できません。

単純 Bayes 法の詳細については、Hand et al.（2016）と Shmueli et al.（2010）を参照してください。

JMP PRO「単純 Bayes」の例

糖尿病患者 442 人に関するデータがあります。このデータには、初診から 1 年後における症状の進行も収集されています。病気の進行度は、「Low」（低）と「High」（高）の 2 値で測定されました。ここでは、ある患者の病気の進行が「High」と「Low」のどちらになるかを予測する分類モデルを作成してみます。

1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Diabetes.jmp」を開きます。
2. ［**分析**］>［**予測モデル**］>［**単純 Bayes**］を選択します。
3. **「Y 2 値」**を選択し、［**Y, 目的変数**］をクリックします。
4. **「年齢」**から**「グルコース」**までを選択し、［**X, 説明変数**］をクリックします。

5. 「**検証**」列を選択し、[**検証**]をクリックします。
6. [**OK**]をクリックします。

図9.2「単純Bayes」レポート

単純Bayes

Y 2値

学習セット

度数	誤分類率	誤分類
309	0.21359	66

検証セット

度数	誤分類率	誤分類
133	0.24060	32

混同行列

学習セット

実測値 Y 2値	予測値 Low	予測値 High
Low	185	41
High	25	58

検証セット

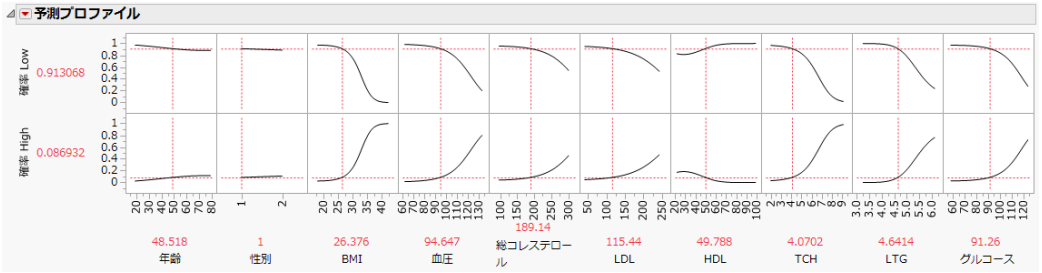
実測値 Y 2値	予測値 Low	予測値 High
Low	70	25
High	7	31

誤分類率は、学習セットで約21%、検証セットで約24%です。「混同行列」を見ると、学習セットと検証セットのどちらも、病気の進行が「High」である患者よりも、「Low」である患者のほうで誤分類が多く発生しています。検証セットの結果は、独立したデータでもある程度の分類が行えることを示唆しています。

次に、この単純Bayes法による分類において、どの特徴（説明変数）が重要になっているかを見てみましょう。

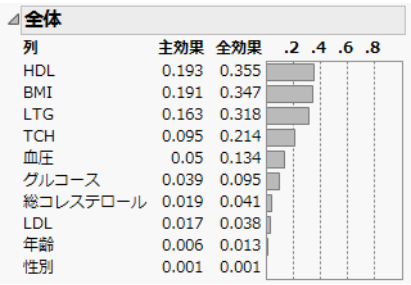
7. 「単純Bayes」の赤い三角ボタンをクリックし、[**プロファイル**]を選択します。

図9.3 病気の進行の予測プロファイル



8. 「予測プロファイル」の赤い三角ボタンをクリックし、[**変数重要度の評価**] > [**独立な一様分布の入力**]を選択します。

図 9.4 変数の重要度

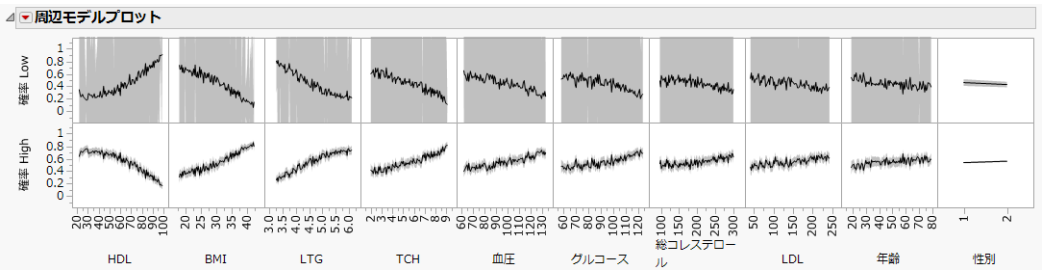


全体

列	主効果	全効果
HDL	0.193	0.355
BMI	0.191	0.347
LTG	0.163	0.318
TCH	0.095	0.214
血圧	0.05	0.134
グルコース	0.039	0.095
総コレステロール	0.019	0.041
LDL	0.017	0.038
年齢	0.006	0.013
性別	0.001	0.001

「要約レポート」は、**「HDL」**、**「BMI」**、**「LTG」**が確率の推定に最も大きな影響を与えていることを示しています。

図 9.5「周辺モデルプロット」レポート

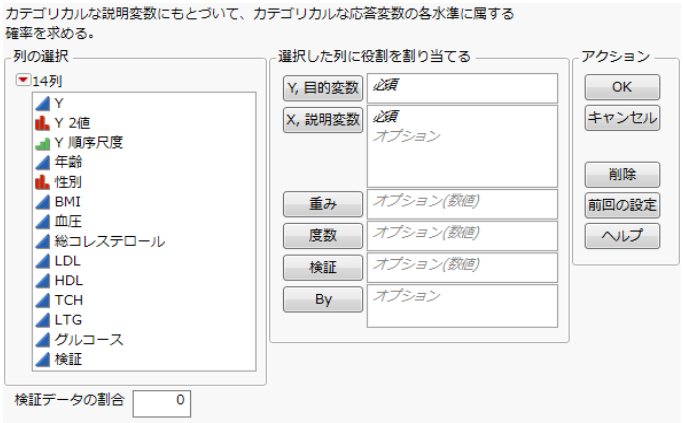


「周辺モデルプロット」レポートの 2 段目のプロットからは、**「HDL」**が高い患者は「High」に分類される確率が低いことがわかります。また、**「BMI」**と**「LTG」**が高い患者は、「High」に分類される確率が高くなっています。

JMP PRO「単純 Bayes」プラットフォームの起動

「単純 Bayes」プラットフォームを起動するには、**［分析］＞［予測モデル］＞［単純 Bayes］**を選択します。

図9.6「単純Bayes」起動ウィンドウ



「単純Bayes」起動ウィンドウには、以下のオプションがあります。

Y, 目的変数　カテゴリカルな応答の列。この列の値が、分析対象のクラスとなります。

X, 説明変数　カテゴリカルまたは連続尺度の説明変数の列。

重み　分析において各行の重みとして使用される数値を含む列。

度数　分析において各行の度数として使用される数値を含む列。

検証　多くとも3つの数値を含む数値列。「パーティション」章の「検証」（93ページ）を参照してください。

> **メモ：**起動ウィンドウで検証列も「検証データの割合」も指定せず、除外されている行がある場合は、それらの行が検証セットとして扱われます。

By　別々に分析を行いたいときに、そのグループ分けをする変数を指定します。指定された列の各水準ごとに、別々に分析が行われます。各水準の結果は別々のレポートに表示されます。複数のBy変数を割り当てた場合、それらのBy変数の水準の組み合わせごとに別々のレポートが作成されます。

検証データの割合　データ全体のうち検証に用いるデータの割合。「パーティション」章の「検証」（93ページ）を参照してください。

JMP PRO「単純Bayes」レポート

起動ウィンドウで［OK］をクリックすると、「単純Bayes」レポートが表示されます。デフォルトの「単純Bayes」レポートには、応答変数ごとに、分類に関する統計量と「混同行列」が表示されます。

JMP PRO 応答変数に対するレポート

応答変数の各列に対して、単純 Bayes モデルによる分類に関する統計量が、学習セット・検証セット・テストセットごとに表示されます。検証セットやテストセットに対する結果は、それらのデータが指定された場合にのみ計算されます。この表には以下の列があります。

度数　該当する各データ（学習セット・検証セット・テストセット）に含まれるオブザベーションの数。

誤分類率　該当する各データにおいて、モデルによって誤分類されたオブザベーションの割合。誤分類されている度数を、全度数で割ったものです。

誤分類　該当する各データにおいて、モデルによって誤分類されたオブザベーションの度数。

JMP PRO「混同行列」レポート

「混同行列」レポートには、学習セットの混同行列が表示されます。検証セットとテストセットを指定した場合は、その混同行列も表示されます。混同行列は、応答の実測値と予測値を 2 元度数表にまとめたものです。

JMP PRO「単純 Bayes」プラットフォームのオプション

「単純 Bayes」の赤い三角ボタンのメニューには、以下のオプションがあります。

予測値の保存　データテーブルに**「単純 予測値 <Y, 目的変数 >」**という列を作成し、そこにクラスの予測値を保存します。

予測式の保存　データテーブルに**「単純 予測式 <Y, 目的変数 >」**という列を作成し、そこにクラスの予測値を求める計算式を保存します。

確率の計算式の保存　各オブザベーションを分類するための計算式をデータテーブル内の新しい列に保存します。3 つの列グループが保存されます。

単純 スコア < クラス >、単純 スコア和　クラスごとに、そのクラスに属する事後確率に比例したスコアを求める計算式が保存されます。そして、「単純 スコア和」という列に、それらのスコアを全クラスで合計した値が保存されます。「確率の計算式」（153 ページ）を参照してください。

単純 確率 < クラス >　クラスごとに、特徴で条件付けたときの、そのクラスに属する条件付き確率（事後確率）を求める計算式が保存されます。「確率の計算式」（153 ページ）を参照してください。

単純 予測式 <Y, 目的変数 >　クラスの予測式が保存されます。

確率の計算式を発行　確率の計算式を作成し、それらを「計算式デポ」プラットフォームの計算式列のスクリプトとして保存します。「計算式デポ」レポートが開いていない場合は、このオプションによって「計算式デポ」ウィンドウが開きます。『予測モデルおよび発展的なモデル』の「計算式デポ」章を参照してください。

プロファイル　対話式のプロファイルを表示します。因子の値を変更すると、それに応じた応答の予測値の曲線が描かれます。詳細については、『プロファイル機能』の「プロファイル」章を参照してください。

以下のオプションについて詳しくは、『JMPの使用法』の「JMPのレポート」章を参照してください。

ローカルデータフィルタ　現在のレポートにおいてデータをフィルタリングするローカルデータフィルタを表示するか、非表示にします。

やり直し　分析を繰り返すオプションや、やり直すオプションを含みます。また、［自動再計算］オプションを選択すると、このオプションに対応しているプラットフォームにおいて、データテーブルに加えた変更が、該当するレポートに即座に反映されるようになります。

スクリプトの保存　レポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。

Byグループのスクリプトを保存　By変数の全水準に対するレポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。起動ウィンドウでBy変数を指定した場合のみ使用可能です。

JMP PRO「単純Bayes」の別例

住宅担保ローンに応募した5,960人の顧客の資産状況に関する履歴データがあります。各顧客は、「良いリスク」と「悪いリスク」に分類されています。ほとんどの説明変数に欠測値があります。未来の顧客のクレジットリスクを分類するためのモデルを作成してみましょう。

1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Eqquity.jmp」を開きます。
2. ［**分析**］>［**予測モデル**］>［**単純Bayes**］を選択します。
3. 「**BAD**」を選択し、［**Y, 目的変数**］をクリックします。

   潜在的な説明変数の1つである「**DEBTINC**」には、欠測値が多数あります。欠測値から何らかの情報が得られる場合もありますが、単純Bayes法は多数の欠測値をうまく処理することができないため、ここでは「**DEBTINC**」をモデルに含めないことにします。
4. 「**LOAN**」から「**CLNO**」までを選択し、［**X, 説明変数**］をクリックします。
5. 「**Validation**」列を選択し、［**検証**］をクリックします。
6. ［**OK**］をクリックします。

図 9.7「BAD」の「単純 Bayes」レポート

BAD

学習セット

度数	誤分類率	誤分類
3576	0.18009	644

検証セット

度数	誤分類率	誤分類
1192	0.19128	228

テストセット

度数	誤分類率	誤分類
1192	0.18960	226

混同行列

学習セット

実測値	予測値 度数	
BAD	Good Risk	Bad Risk
Good Risk	2691	214
Bad Risk	430	241

検証セット

実測値	予測値 度数	
BAD	Good Risk	Bad Risk
Good Risk	859	58
Bad Risk	170	105

テストセット

実測値	予測値 度数	
BAD	Good Risk	Bad Risk
Good Risk	885	64
Bad Risk	162	81

学習セット、検証セット、テストセットの誤分類率は 18～19% です。各セットの混同行列を見ると、誤分類の多くは「Bad Risk」（リスクが高い）の顧客を誤って「Good Risk」（リスクが低い）に分類したことに起因しています。

顧客の資産状況に基づいて、その顧客がリスクが高いかどうかを求めてみましょう。

7. 「単純 Bayes」の赤い三角ボタンをクリックし、**[確率の計算式の保存]** を選択します。

   データテーブルには、3 種類の列が追加されます。なお、説明変数に欠測値があるデータ行は、新しく作成された列の値も欠測値になります。

   – 「単純 スコア」の 3 列には、「Good Risk」と「Bad Risk」のスコア、および、それらの和が含まれます。
   – 「単純 確率」の 2 つの列には、「Good Risk」と「Bad Risk」の事後確率を求める計算式が含まれます。
   – **「単純 予測式 BAD」** 列には、事後確率が最大となっているクラスに割り当てる計算式が保存されています。

   これらの計算式は、新しい顧客のスコアを計算するときにも使用できます。計算式列の詳細については、「確率の計算式」（153 ページ）を参照してください。

JMP PRO「単純 Bayes」プラットフォームの統計的詳細

JMP PRO アルゴリズム

単純 Bayes 法は、特徴値に基づいて、所属する事後確率が最大となっているクラスに各データ行を分類します。単純 Bayes 法は、「ある 1 つのクラスで条件付けたときに、各特徴が独立である」と仮定しています。

分類（クラス）を C_1、…、C_k とします。また、特徴（説明変数）を X_1、X_2、…、X_p とします。

また、C_r で条件付けたときの、$X_j = x_j$ である条件付き確率を次のように定義します。

- X_j がカテゴリカルの場合: $P(C_r | x_j)$

- X_jが連続尺度の場合:

$$P(C_r|x_j) = \frac{1}{s}\phi((x_j - m)/s)$$

ここで、ϕは標準正規分布の密度関数です。また、mとsは、それぞれ、クラスがC_rであるデータ行の説明変数から計算された、平均と標準偏差です。

説明変数の値がx_1、x_2、...、x_pであるオブザベーションがC_rクラスに属する条件付き確率は、次のように求められます。

$$P(C_r|(x_1, \ldots, x_p)) = \left(P(C_r)\prod_{j=1}^{p}[P(x_j|C_r)]\right) \Big/ \left(\sum_{i=1}^{k} P(C_i)\left(\prod_{j=1}^{p}[P(x_j|C_i)]\right)\right)$$

クラスに分類するときには、この条件付き確率（事後確率）が最も大きいクラスに分類されます。

JMP PRO 確率の計算式

この節では、[確率の計算式の保存] オプションで保存される計算式について説明します。説明変数がx_1、x_2、...、x_pであるデータ行が、クラスC_rに属する条件付き確率（事後確率）は、「アルゴリズム」（152ページ）で説明した$P(C_r|(x_1, \ldots, x_p))$の計算式で求められます。

単純スコアの計算式

C_rというクラスの「単純 スコア」の計算式は、$P(C_r|(x_1, \ldots, x_p))$の式の分子です。

「単純 スコア和」の計算式は、全クラスの条件付き確率$P(C_r|(x_1, \ldots, x_p))$を合計します。
これは、$P(C_r|(x_1, \ldots, x_p))$の式の分母です。

単純確率の計算式

C_rクラスの「単純 確率」の計算式は、$P(C_r|(x_1, \ldots, x_p))$に等しくなります。

単純 予測式

「単純 予測式」は、各データ行において、$P(C_r|(x_1, \ldots, x_p))$が最大となっているクラスに分類します。これは、「単純 スコア」が最大となっているクラスに相当します。

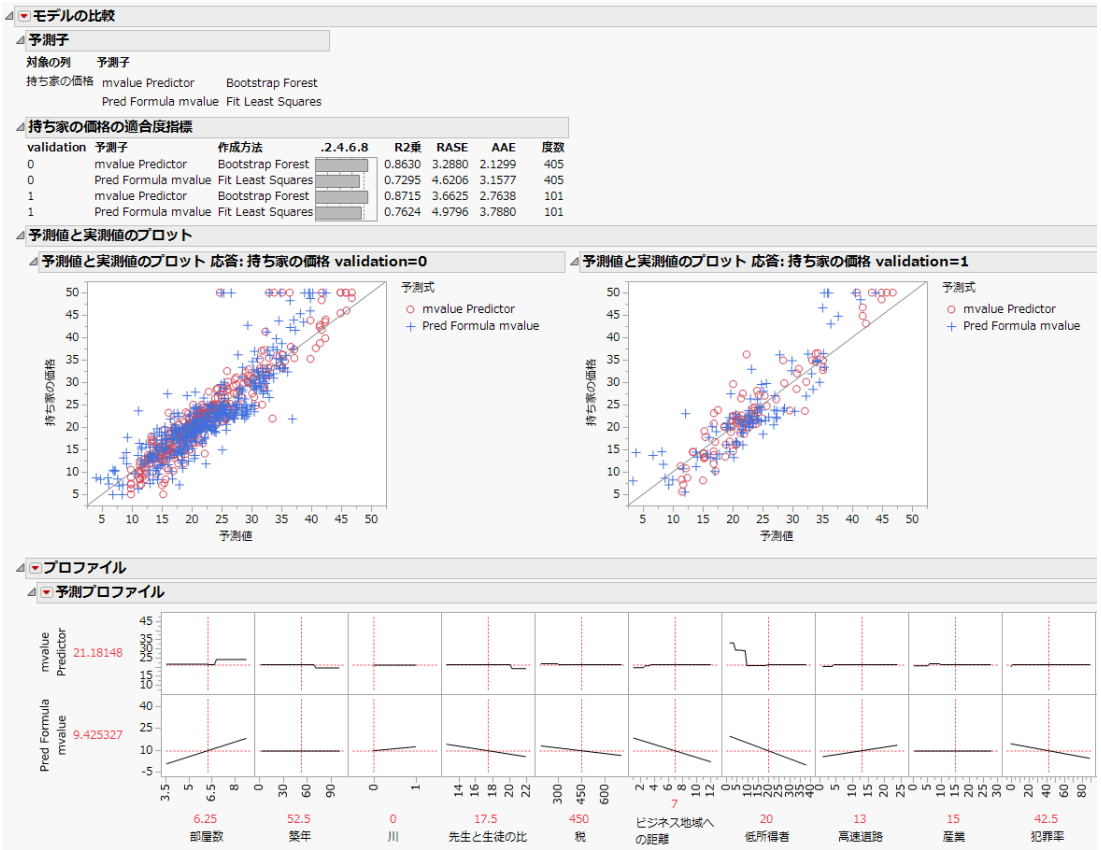
第10章

JMP PRO モデルの比較
予測モデルの精度比較

「モデルの比較」プラットフォームは JMP Pro でのみ利用できます。

JMP Proの「モデルの比較」プラットフォームでは、さまざまなモデルの予測精度を比較できます。各モデルの適合度を求めたり、診断プロットを描いたりできます。

図10.1「モデルの比較」の例



JMP PRO「モデルの比較」の例

この節では、「モデルの比較」プラットフォームの使用例を紹介します。使用するデータは、住宅価格のデータです。各地域における住宅価格の中央値を、その地域の属性から予測するモデルを作成します。回帰モデルとブートストラップ森を比較します。

まず、［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Boston Housing.jmp」を開いてください。

検証列の作成

1. **「検証」**という名前の列を作成します。
2. 列情報ウィンドウの「データの初期化」リストから［**乱数**］を選択します。
3. ［**指示乱数**］を選択します。
4. ［**OK**］をクリックします。

   0が割り当てられた行は学習セットとして使われます。一方、1が割り当てられた行は検証セットとして使われます。

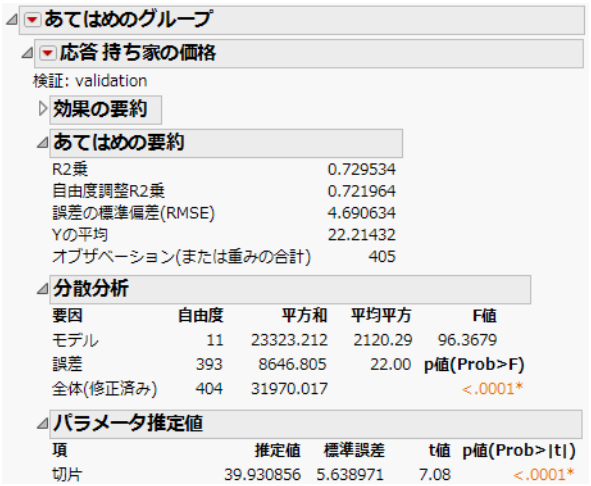
回帰モデルを作成し、予測式を列に保存

1. ［**分析**］>［**モデルのあてはめ**］を選択します。
2. **「持ち家の価格」**を選択し、［**Y**］をクリックします。
3. （**「検証」**を除く）その他の列をすべて選択し、［**追加**］をクリックします。
4. 「手法」リストから［**ステップワイズ法**］を選択します。
5. **「検証」**列を選択し、［**検証**］ボタンをクリックします。
6. ［**実行**］ボタンをクリックします。
7. 「停止ルール」リストから［**閾値ｐ値**］を選択します。
8. ［**実行**］ボタンをクリックします。
9. ［**モデルの実行**］ボタンをクリックします。

   「モデルのあてはめ」レポートが表示されます。図10.2はその一部です。

10. 「応答」の赤い三角ボタンのメニューから［**列の保存**］>［**予測式**］を選択し、予測式を列に保存します。

図10.2「モデルのあてはめ」レポート

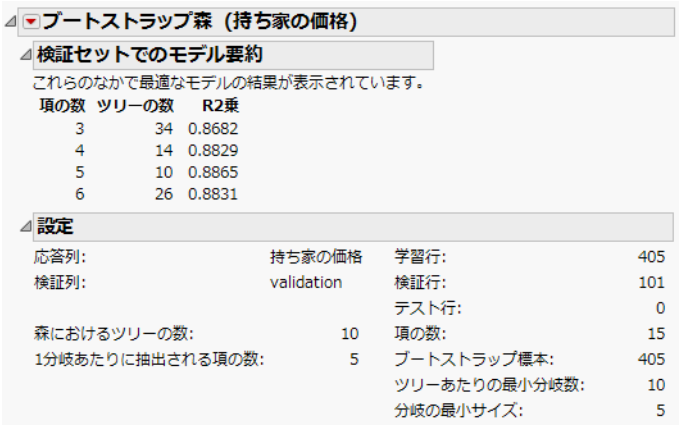


ブートストラップ森モデルを作成し、予測式を列に保存

1. ［**分析**］>［**予測モデル**］>［**パーティション**］を選択します。
2. 「**持ち家の価格**」を選択し、［**Y, 目的変数**］をクリックします。
3. （「**検証**」を除く）その他の列をすべて選択し、［**X, 説明変数**］をクリックします。
4. 「**検証**」列を選択し、［**検証**］ボタンをクリックします。
5. 「手法」リストから［**ブートストラップ森**］を選択します。
6. ［**OK**］をクリックします。
7. ［**早期打ち切り**］チェックボックスをオンにします。
8. ［**項数に対する複数のあてはめ**］チェックボックスをオンにします。
9. ［**OK**］をクリックします。

   「ブートストラップ森」レポートが表示されます。図10.3はその一部です。
10. 「ブートストラップ森」の赤い三角ボタンのメニューから［**列の保存**］>［**予測式の保存**］を選択し、予測式を列に保存します。

図10.3 ブートストラップ森のモデル



ブートストラップ森（持ち家の価格）

検証セットでのモデル要約

これらのなかで最適なモデルの結果が表示されています。

項の数	ツリーの数	R2乗
3	34	0.8682
4	14	0.8829
5	10	0.8865
6	26	0.8831

設定

応答列:	持ち家の価格	学習行:	405
検証列:	validation	検証行:	101
		テスト行:	0
森におけるツリーの数:	10	項の数:	15
1分岐あたりに抽出される項の数:	5	ブートストラップ標本:	405
		ツリーあたりの最小分岐数:	10
		分岐の最小サイズ:	5

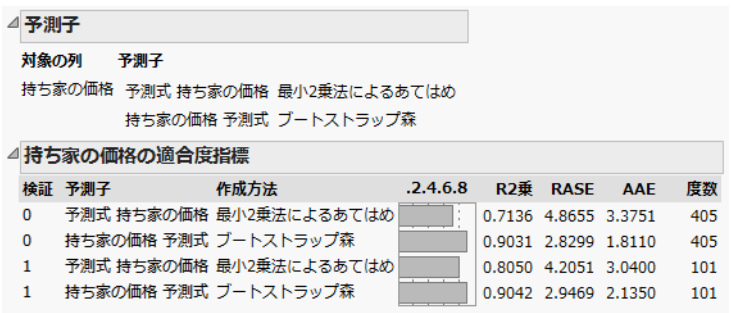
モデルの比較

1. ［**分析**］>［**予測モデル**］>［**モデルの比較**］を選択します。
2. 2つの予測式列を選択し、［**Y, 予測式**］をクリックします。
3. 「**検証**」を選択し、［**グループ化**］をクリックします。
4. ［**OK**］をクリックします。

「モデルの比較」レポートが表示されます（図10.4）。

メモ：学習セットと検証セットが無作為に決められているため、実際の結果は図と同じにはなりません。

図10.4「モデルの比較」レポート



予測子

対象の列	予測子
持ち家の価格	予測式 持ち家の価格 最小2乗法によるあてはめ
	持ち家の価格 予測式 ブートストラップ森

持ち家の価格の適合度指標

検証	予測子	作成方法	.2.4.6.8	R2乗	RASE	AAE	度数
0	予測式	持ち家の価格 最小2乗法によるあてはめ		0.7136	4.8655	3.3751	405
0	持ち家の価格	予測式 ブートストラップ森		0.9031	2.8299	1.8110	405
1	予測式	持ち家の価格 最小2乗法によるあてはめ		0.8050	4.2051	3.0400	101
1	持ち家の価格	予測式 ブートストラップ森		0.9042	2.9469	2.1350	101

レポートにおいて、「検証」の値が「0」になっているデータは、学習セットです。モデルの推定には学習セットが使われているため、学習セットのR2乗は過度に大きくなっている可能性があります。学習セットのR2乗は、将来のデータに対する予測精度を正しく反映していません。この傾向は、特にブートストラップ森に対して言えます。

「検証」の値が「1」であるデータにおけるR2乗を見てみましょう。ブートストラップ森の方が回帰モデルより、R2乗値が大きくなっています。

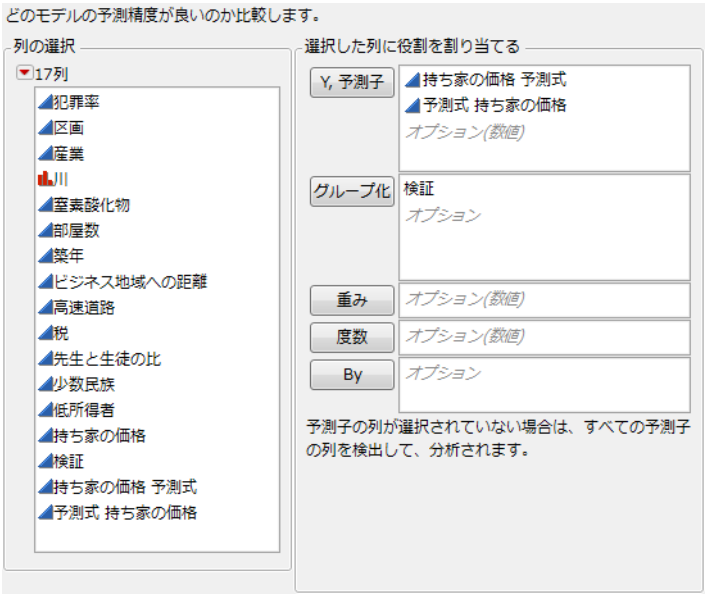
関連情報

- 『基本的な回帰モデル』の「モデルの指定」章
- 「パーティション」章（71ページ）

「モデルの比較」プラットフォームの起動

「モデルの比較」プラットフォームを起動するには、[**分析**] > [**予測モデル**] > [**モデルの比較**] を選択します。

図10.5「モデルの比較」起動ウィンドウ



Y, 予測子　モデルの予測式や予測値を含む列です。計算式として予測式を含んでいる列、および、データ値として予測値を含んでいる列のいずれも使用できます。JMPのプラットフォームで作成される予測式や予測値の列には、「予測対象」または「応答確率」の列プロパティが割り当てられます。どちらの列プロパティも持たない列を指定した場合は、Yによってどの列を予測したいのかを指定する必要があります。

カテゴリカルな応答に対して予測を行うプラットフォームのほとんどで、k個の水準を持つ応答に対し、各水準の確率を予測するk個の列がデータテーブルに保存されます。それらk個の列すべてを [**Y, 予測子**] に指定する必要があります。k個の列を保存しないプラットフォームでは、応答水準の予測値を含む列を [**Y, 予測子**] に指定してください。

[**Y, 予測子**] に何も指定しなかった場合は、現在のデータテーブルの中で「予測対象」または「応答確率」の列プロパティを持つ予測式や予測値の列が使用されます。

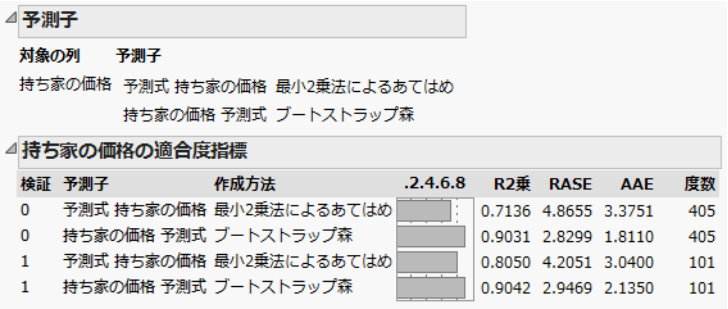
グループ化　データをグループ分けする列。各グループごとに、結果が計算されます。

その他の役割ボタンはJMPプラットフォームに共通のものです。詳細については、『JMPの使用法』の「はじめに」章を参照してください。

「モデルの比較」レポート

図10.6は、連続尺度の応答変数に対する「モデルの比較」レポートの例です。

図10.6「モデルの比較」レポート



予測子

対象の列	予測子
持ち家の価格	予測式 持ち家の価格 最小2乗法によるあてはめ
	持ち家の価格 予測式 ブートストラップ森

持ち家の価格の適合度指標

検証	予測子	作成方法	.2.4.6.8	R2乗	RASE	AAE	度数
0	予測式 持ち家の価格	最小2乗法によるあてはめ		0.7136	4.8655	3.3751	405
0	持ち家の価格 予測式	ブートストラップ森		0.9031	2.8299	1.8110	405
1	予測式 持ち家の価格	最小2乗法によるあてはめ		0.8050	4.2051	3.0400	101
1	持ち家の価格 予測式	ブートストラップ森		0.9042	2.9469	2.1350	101

「予測子」レポートには、すべての応答変数の列、および、すべてのモデルが表示されます。また、ここには、予測式や予測値の列を作成したプラットフォームの名前も表示されます。

「適合度指標」レポートには、各モデルの適合度統計量が表示されます。なお、表示される適合度統計量の種類は、応答変数が連続尺度かカテゴリカルかによって異なります。

連続尺度の応答に対する適合度指標

R2乗　決定係数（寄与率）。データに欠測値がない場合は、「モデルの比較」レポートにおけるこの値は、元のモデルのR2乗と一致します。欠測値がある場合は、異なった値になります。

RASE　誤差の平均平方根（Root of Average Squared Error）。

- 以下の式において、予測誤差（応答の実測値と予測値の差）の平方和をSSEとします。
- また、オブザベーションの数をnとします。
- RASEは次のように定義されます。

$$\text{RASE} = \sqrt{\frac{SSE}{n}}$$

AAE　平均絶対誤差（Average Absolute Error）。

度数　各行の度数を含む列。

カテゴリカルな応答に対する適合度指標

エントロピー R2 乗　あてはめたモデルの対数尤度と、切片だけのモデルの対数尤度を比較している指標です。あてはめたモデルの対数尤度を、切片だけのモデルの対数尤度で割り、その値を1から引いたものです。この指標の範囲は0～1です。

一般化 R2 乗　この指標は、一般的な回帰モデルに適用できるものです。一般化R2乗は、尤度Lから算出され、最大が1となるように尺度化されています。完全にモデルがデータにあてはまっている場合は1、切片だけのモデルと同等なあてはまりの場合には0になります。一般化R2乗は、通常のR2乗（正規分布に従う連続尺度の応答変数に対する標準最小2乗法のR2乗）を一般化したものです。この一般化R2乗は、「Nagelkerke の R^2」、または「Craig and Uhler の R^2」とも呼ばれており、Cox and Snell の疑似 R^2 を最大が1になるように尺度化したものです。詳細は、Nagelkerke（1991）を参照してください。

平均 -Log p　-log(p)の平均です。pは、実際に生じた応答水準に対する予測確率です。

RMSE　誤差の標準偏差（誤差平方和を自由度で割ったものの平方根）。応答がカテゴリカルの場合は、誤差は(1-p)で計算されます。ここで、pは、実際に生じた応答水準に対する予測確率です。

平均 絶対偏差　誤差の絶対値の平均。応答がカテゴリカルの場合は、誤差は(1-p)で計算されます。ここで、pは、実際に生じた応答水準に対する予測確率です。

誤分類率　予測確率が最も大きい応答の水準が、観測された水準と一致しない割合。

N　オブザベーションの数。

関連情報

「ニューラルネットワーク」章の「学習と検証における適合度」（65ページ）では、カテゴリカルな応答の適合度指標について具体的に説明しています。

「モデルの比較」プラットフォームのオプション

「モデルの比較」の赤い三角ボタンのメニューには、使用するデータによって若干異なるオプションが表示されます。

JMP PRO 連続尺度とカテゴリカルに共通のオプション

モデル平均化　予測値（連続尺度の場合）、もしくは、予測確率（カテゴリカルの場合）の、複数のモデルにおける平均を含んだ列を作成します。

連続尺度の応答のオプション

予測値と実測値のプロット　予測値と実測値の散布図を表示します。複数のモデルの結果が、重ね合わされてプロットされます。

行番号と残差のプロット　行番号と残差のプロットを表示します。複数のモデルの結果が、重ね合わされてプロットされます。

プロファイル　予測確率に対するプロファイルを表示します。プロファイルはモデルごとに表示されます。

カテゴリカルな応答のオプション

ROC曲線　応答変数の各水準に対するROC曲線を表示します。複数のモデルの曲線が、重ね合わされてプロットされます。

AUCの比較　各モデルのAUCを比較します。ここでのAUC（Area Under Curver）は、ROC曲線より下側の領域の面積を指します。各モデルのAUCは、そのモデルの適合度を表します。AUCが1なら、あてはまりが完全なことを意味します

レポートには次の情報も表示されます。

- 各AUCの標準誤差と信頼区間
- AUCのペアごとの比較。差、および、その標準誤差、信頼区間、仮説検定
- 「すべてのAUCが等しい」という帰無仮説に対する全体的な仮説検定

リフトチャート　応答変数の水準ごとに、リフト曲線を表示します。複数のモデルの曲線が、重ね合わされてプロットされます。

累積ゲイン曲線　応答変数の水準ごとに、累積ゲイン曲線を表示します。累積ゲイン曲線は、予測確率でデータを降順に並び替えて、すべてのデータでの割合に対して、該当の応答水準での割合をプロットしたものです。予測が完璧なモデルでは、累積ゲイン曲線は、横軸がデータ全体における応答水準の割合になった時点で、縦軸が1.0に達します。複数のモデルにおける曲線が、重ね合わされてプロットされます。

混同行列　各モデルの混同行列を表示します。混同行列は、応答の実測値と予測値を2元度数表にまとめたものです。度数と割合の混同行列が表示されます。グループ変数の水準ごとに個別に混同行列が作成されます。

応答に「利益行列」の列プロパティが割り当てられている場合は、混同行列の右側に「決定行列 度数」と「決定行列 割合」の行列が表示されます。これらの行列の詳細については、「パーティション」章の「パーティションの別例」（95ページ）を参照してください。

プロファイル　予測確率に対するプロファイルを表示します。プロファイルはモデルごとに表示されます。

関連情報

- 「パーティション」章の「ROC曲線」（90ページ）
- 「パーティション」章の「リフトチャート」（91ページ）

「モデルの比較」の別例

この例では、自動車のサイズを予測するモデルを扱います。ロジスティック回帰とディシジョンツリーを比較します。

まず、［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Car Physical Data.jmp」を開いてください。

ロジスティック回帰モデルの作成

1. ［**分析**］>［**モデルのあてはめ**］を選択します。
2. 「**タイプ**」を選択し、［**Y**］をクリックします。
3. 「**生産国**」、「**車両重量**」、「**最小回転半径**」、「**排気量**」、「**馬力**」の列を選択し、［**追加**］をクリックします。
4. ［**実行**］をクリックします。

   「名義ロジステイックのあてはめ」レポートが表示されます。
5. 「名義ロジステイックのあてはめ」の赤い三角ボタンのメニューから［**確率の計算式の保存**］を選択し、予測式を列に保存します。

ディシジョンツリーモデルの作成

1. ［**分析**］>［**予測モデル**］>［**パーティション**］を選択します。
2. 「**タイプ**」を選択し、［**Y, 目的変数**］をクリックします。
3. 「**生産国**」、「**車両重量**」、「**最小回転半径**」、「**排気量**」、「**馬力**」の列を選択し、［**X, 説明変数**］をクリックします。
4. 「手法」リストで［**ディシジョンツリー**］が選択されていることを確認します。
5. ［**OK**］をクリックします。

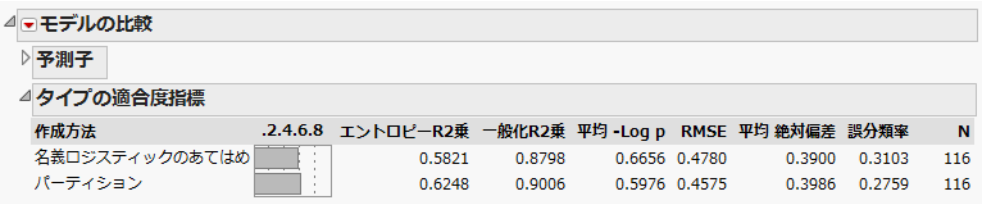
   「パーティション」レポートが表示されます。
6. ［**分岐**］を10回クリックします。
7. 「タイプのパーティション」の赤い三角ボタンのメニューから［**列の保存**］>［**予測式の保存**］を選択し、予測式を列に保存します。

モデルの比較

1. ［**分析**］>［**予測モデル**］>［**モデルの比較**］を選択します。
2. 「**確率**」で始まるすべての列を選択し、［**Y, 予測式**］をクリックします。
3. ［**OK**］をクリックします。

   「モデルの比較」レポートが表示されます（図10.7）。

図10.7「モデルの比較」レポート



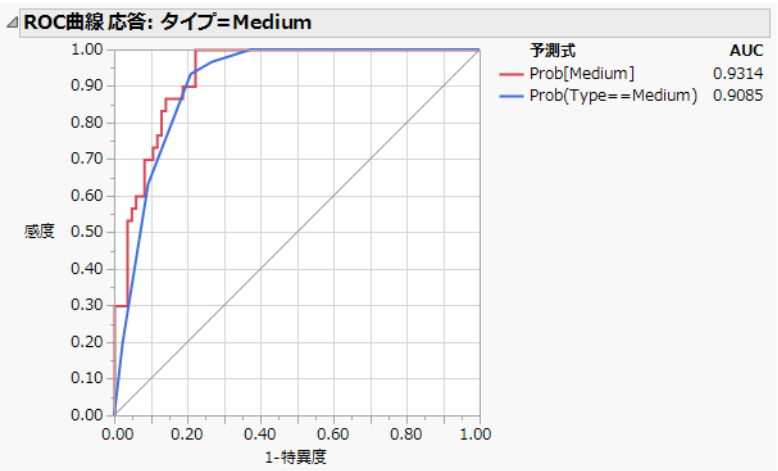
作成方法	.2.4.6.8	エントロピーR2乗	一般化R2乗	平均 -Log p	RMSE	平均 絶対偏差	誤分類率	N
名義ロジスティックのあてはめ		0.5821	0.8798	0.6656	0.4780	0.3900	0.3103	116
パーティション		0.6248	0.9006	0.5976	0.4575	0.3986	0.2759	116

レポートを見ると、パーティションモデルの「エントロピー R2 乗」と「一般化 R2 乗」は高め、「誤分類率」は低めになっています。

4. ［モデルの比較］の赤い三角ボタンのメニューから［**ROC 曲線**］を選択します。

「タイプ」別に ROC 曲線が表示されます。そのうちの 1 つを図 10.8 に示します。

図10.8「Medium」の ROC 曲線



ROC 曲線を見ると、この 2 つのモデルの予測能力は、ほぼ同じであることがわかります。

5. ［モデルの比較］の赤い三角ボタンのメニューから［**AUC の比較**］を選択します。

「タイプ」別に AUC の比較レポートが表示されます。そのうちの 1 つを図 10.9 に示します。

図10.9「Medium」のAUCの比較

AUCの比較 タイプ=Medium

予測子	AUC	標準誤差	下側95%	上側95%
確率[Medium]	0.9314	0.0218	0.8742	0.9637
確率(タイプ==Medium)	0.9085	0.0255	0.8448	0.9477

予測子	vs. 予測子	AUCの差	標準誤差	下側95%	上側95%	カイ2乗	p値(Prob>ChiSq)
確率[Medium]	確率(タイプ==Medium)	0.0229	0.0234	-0.023	0.0687	0.9553	0.3284

検定	カイ2乗	自由度	p値(Prob>ChiSq)
すべてのAUCが等しい	0.95534	1	0.3284

このレポートは、AUC（ROC曲線より下の領域の面積）の差を検定しています。結果から、**「タイプ」**のどの水準においても、AUCには統計的な有意差は見られません。

次の理由により、2つのモデルの予測能力の間には、大きな差はないだろうと結論付けることができます。

- R2乗値とROC曲線がほぼ同じである。
- AUCに、統計的な有意差が見られない。

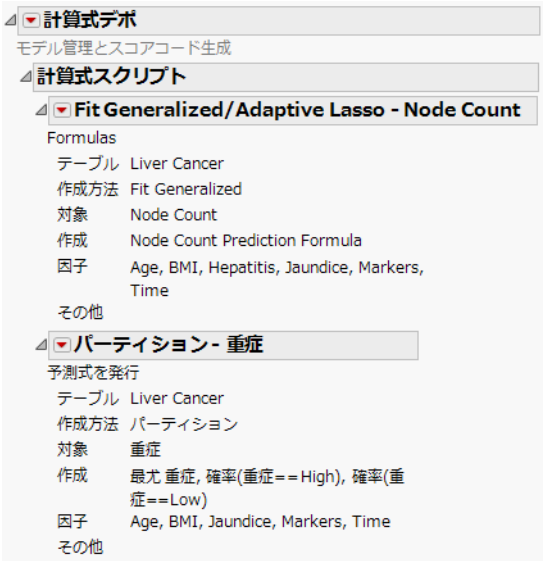
第 11 章

JMP PRO 計算式デポ
モデル管理とスコアコード生成

「計算式デポ」プラットフォームは、JMP Pro でのみ利用できます。

計算式デポは、モデルを整理・比較・保管するためのツールです。計算式デポにモデルを発行すると、スコアを計算するためのコードを生成できます（JMP 以外で実行できるコードも生成できます）。計算式デポには、候補となるモデルを元の JMP データテーブルとは切り離して保管できます。計算式デポでは保管されたモデルを比較でき、プロファイルやモデル比較のプラットフォームを開くことができます。選択されたモデルを、元の JMP データテーブルや新しいデータテーブルに保存できます。また、C、Python、JavaScript、SAS、SQL といった言語のコードも生成できます。

図 11.1 計算式デポの例



JMP PRO「計算式デポ」プラットフォームの概要

計算式デポは、モデルの整理・比較・保管を行うためのツールです。計算式デポでは、「モデル」は予測式を指します。計算式デポでは、各モデルの予測式が、列スクリプトとして保存されます。計算式デポは、JMP テーブルに予測式を追加したり、スコアを計算したりできます。また、スコアを計算するコードとして、JMP 以外での言語のものも生成できます。

計算式デポでは、以下の処理を実行できます。

- 予測式をデータテーブルと切り離して保存する
- 複数のデータテーブルの予測式を1箇所に保存する
- モデルを比較する
- モデルのプロファイルを作成する
- JMP 内で新しいデータのスコア計算を行うため、予測式をデータテーブルに追加する
- スコアを計算するための、C、Python、Java Script、SAS、SQL といった JMP 以外の言語のコードを生成する

JMP PRO 計算式デポの例

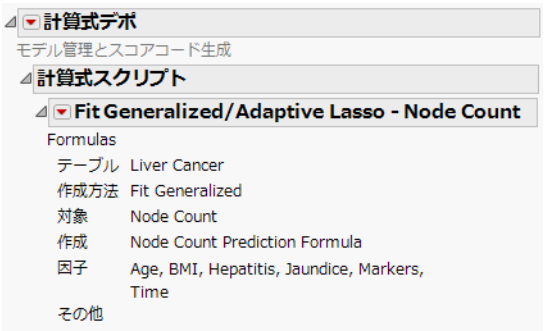
「Liver Cancer.jmp」データテーブルは、調査開始時における肝臓がん患者の重症度に関するデータです。このデータテーブルには、モデルのスクリプトも多数含まれています。この例では、計算式デポの使い方を学ぶために、これらの予め保存されているスクリプトを使ってモデルをあてはめます。

1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Liver Cancer」を開きます。
2. 「Lasso Poisson 分布 検証列」というスクリプトの緑色の三角ボタンをクリックします。
3. 「適応型 Lasso (検証法: 検証列)」の赤い三角ボタンをクリックし、［**列の保存**］>［**予測式を発行**］を選択します。

   「一般化モデルのあてはめ」の予測式を含む「計算式デポ」が開きます。
4. 予測式をデータテーブルに追加するには、「一般化モデルのあてはめ / 適応型 Lasso - 結節数」の赤い三角ボタンのメニューから［**スクリプトの実行**］を選択します。
5. JMP 以外の言語のコード、たとえば Python で書かれたスコアを求めるコードを生成するには、「一般化モデルのあてはめ / 適応型 Lasso - 結節数」の赤い三角ボタンのメニューから［**Python コードの作成**］を選択します。スクリプトウィンドウが開き、Python のコードが表示されます。「「計算式デポ」でスコア計算のコードを生成」（171ページ）を参照してください。
6. 「計算式デポ」を保存するには、［**ファイル**］>［**保存**］を選択します。

メモ：手順3で［**列の保存**］>［**予測式の保存**］を選択すると、手順3と手順4の結果が得られます。ただしその場合、予測式は計算式デポに含まれません。

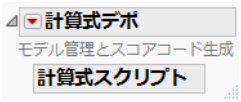
図11.2 計算式デポと一般化モデル



JMP PRO「計算式デポ」プラットフォームの起動

「計算式デポ」を起動するには、［**分析**］>［**予測モデル**］>［**計算式デポ**］を選択します。

図11.3 起動時の空白の計算式デポ



または、「計算式デポ」が開いていない状態で［発行］コマンドを選択すると、「計算式デポ」が開きます。

JMP PRO 計算式デポに予測式を発行するプラットフォーム

以下のプラットフォームは、計算式デポに予測式を発行し、スコア計算に使用するコードを生成することができます。

- 判別分析
- 最小2乗法
- ロジスティック回帰
- パーティション
- アップリフト
- K近傍法
- 単純Bayes
- ニューラル
- 潜在クラス分析
- 主成分分析

- 一般化回帰
- PLS
- Gauss過程

予測式を計算式デポに発行できないプラットフォームでは、まず、予測式をデータテーブルに保存します。そして、計算式デポにて［**列の計算式から追加**］を選択します。この操作により、保存された予測式を計算式デポに追加できます。ただし、このようなモデルについては、スコアを計算するコードが完全には動作しない場合があります。

スコアコードの詳細については、「「計算式デポ」でスコア計算のコードを生成」（171ページ）を参照してください。

JMP PRO「計算式デポ」プラットフォームのオプション

列の計算式から追加　現在の計算式デポに、予測式の列が追加されます。

スクリプトの表示　新しい計算式ウィンドウ（または開いている計算式ウィンドウ）に、現在の計算式デポにあるすべての計算式のスクリプトが表示されます。

スクリプトのコピー　現在の計算式デポにあるすべてのスクリプトがクリップボードにコピーされます。

変換として計算式をコピー　現在の計算式デポから選択したモデルがコピーされます。このコピーされたモデルは、`Transform Columns()` ステートメントに含まれている形式でクリップボードにコピーされます。

スクリプトの実行　現在の計算式デポから選択したモデルが、JMPデータテーブル内の新しい列に保存されます。

Cコードの生成、Pythonコードの生成、JavaScriptコードの生成、SASコードの生成、SQLコードの生成

現在の計算式デポから選択したモデルに基づいた、スコアを計算するためのコードが生成されます。新しいスクリプトウィンドウが開き、選択したモデルのスコアを求めるコード（C、Python、JavaScript、SAS DS2、SQLで書かれたコード）が表示されます。このコードを使うと、いろいろな言語や環境でスコアを計算することができます。「「計算式デポ」でスコア計算のコードを生成」（171ページ）を参照してください。

モデルの比較　現在の計算式デポに保存されている、複数のモデルを比較できます。計算式デポに複数のデータテーブルのモデルが保存されている場合は、まず、データテーブルやモデルを選択します。『予測モデルおよび発展的なモデル』の「モデルの比較」章を参照してください。

モデルの比較を削除　現在の計算式デポから「モデルの比較」レポートを削除します。

プロファイル　現在の計算式デポから選択されたモデルに対して、プロファイルを描きます。計算式デポに複数のデータテーブルのモデルが保存されている場合は、まず、データテーブルやモデルを選択します。『プロファイル機能』の「プロファイル」章を参照してください。

プロファイルを削除　現在の計算式デポからすべてのプロファイルを削除します。

以下のオプションについて詳しくは、『JMPの使用法』の「JMPのレポート」章を参照してください。

やり直し 分析を繰り返すオプションや、やり直すオプションを含みます。また、［自動再計算］オプションを選択すると、このオプションに対応しているプラットフォームにおいて、データテーブルに加えた変更が、該当するレポートに即座に反映されるようになります。

スクリプトの保存 レポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。

Byグループのスクリプトを保存 By変数の全水準に対するレポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。起動ウィンドウでBy変数を指定した場合のみ使用可能です。

JMP PRO 計算式デポのモデルのオプション

計算式デポに保存された予測式には、それぞれ個別のオプションメニューがあります。メインメニューと同じ表示、コピー、実行、コード生成の各オプションのほかに、次のようなオプションがあります。

新しい列の名前を変更 モデルの名前を変更できます。指定された名前は、スクリプトを実行して予測式の列をJMPデータテーブルに追加した際に、その列名に使用されます。また、生成されるコードでも、ここで指定された名前が使用されます。

削除 計算式デポからモデルを削除します。このコマンドは、一度実行すると元に戻せません。

JMP PRO「計算式デポ」でスコア計算のコードを生成

スコア計算のためのコードを生成すると、JMPで構築した予測モデルをさまざまな環境で使えるようになります。JMPの多くのプラットフォームにおいて、予測式を計算式デポに発行できます。なお、すべての場合において、完全なコードが生成されるわけではありません。完全なコードが生成される場合もあれば、コードの一部分だけが作成される場合もあります。また、該当のコードにサポートされていない関数が含まれていて、プログラミングを加えないと利用できない場合もあります。

たとえば、以下のような処理を実行できます。

- SASコードを生成して、そのコードをSAS Model Managerで用いる。
- SQLコードを生成して、そのコードをデータベースのETLプロセスに加える。
- Cコードを生成して、そのコードを使ってアプリケーションを構築し、データを変換する。
- Pythonコードを生成して、そのコードを使ってJupyter Notebookを作成し、スコアを動的に表示する。
- JavaScriptコードを生成して、そのコードをWebアプリケーションに含め、Web閲覧者が自分のデータのスコアを計算できるようにする。

C、Python、JavaScriptの各言語では、生成されたコードを展開またはコンパイルする際に、.hファイルのような補助的なコードとユーティリティライブラリを含める必要があります。これらのファイルは、JMPのインストール時に「Scoring」フォルダに保存されます。

Cコード 生成されたCスコアコードを用いるには、ライブラリの形式でコンパイルし、アプリケーションにリンクする必要があります。リンクは、静的リンクと動的リンクのどちらでもかまいません。コンパイルとリンクに必要なヘッダーファイル（jmp_lib.h、jmp_parms.h、jmp_score.h）が、JMPのインストール時に作成される「Scoring」フォルダ内の「C」フォルダに用意されています。

Pythonコード 生成されたPythonコードの実行に必要なファイル（jmp_score.py）が、JMPのインストール時に作成される「Scorin」フォルダ内の「Python」フォルダに用意されています。

JavaScriptコード 生成されたJavaScriptコードの実行に必要なファイル（jmp_score.js）が、「Scoring」フォルダ内の「JavaScript」フォルダに用意されています。

SASコード 生成されたSASコードを、PROC DS2ステートメントでラップすれば、SAS In-Database Code Acceleratorをはじめとする SASアプリケーションで使用できます。

ヒント：カテゴリカルな応答のロジスティックモデルやニューラルモデルなど、IfMaxを用いているコードでは、一時変数の宣言を **method run()** ステートメントの前に移動してください。また、変数名を、SASの命名規則に従って変更してください。

SQLコード 生成されたSQLコードを、SELECTステートメントでラップすれば、ほとんどのデータベースにおけるSQLクエリで使用できます。

メモ：「placeholder」または「ERROR」を含むコードは、サポートされていない関数呼び出しがあることを意味します。

第 12 章

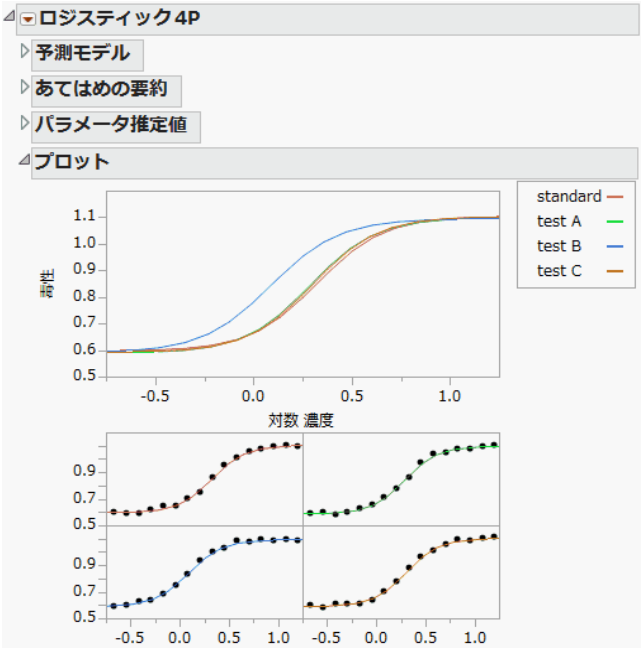
曲線のあてはめ

あらかじめ用意された非線形モデルをあてはめる

特に自然科学分野では、変数間の関係を表すいろいろな非線形モデルが知らています。たとえば、薬理学では、薬の濃度（用量）を変えたときに、反応がどのように変化するかを試験します。薬の濃度と反応の強さとの関係は、シグモイド曲線によってたびたび表されます。非線形モデルのほかの例として、指数成長曲線が挙げられます。指数成長曲線は、時間が経過する従って、集団の大きさがどのように変化していくかを表します。

「曲線のあてはめ」機能を使う場合、パラメータ推定値の開始値を指定する必要がなく、モデル計算式を作成する必要もありません。自分自身で開始値を指定したり、モデル式を指定したりしたいときには、「曲線のあてはめ」ではなく、「非線形回帰」プラットフォームを使ってください。「非線形回帰」プラットフォームでは、どのような非線形モデルの式も指定することができます。詳細については、「非線形回帰」章（195ページ）を参照してください。

図12.1「非線形回帰」プラットフォームの「曲線のあてはめ」機能



「曲線のあてはめ」プラットフォームについて

モデルの中には、パラメータに関して**線形**であるもの（たとえば2次式などの多項式）や、線形に変換できるもの（たとえばxを対数変換するなど）があります。そのようなモデルには、「モデルのあてはめ」プラットフォームや「二変量の関係」プラットフォームがより適しています。『基本的な回帰モデル』の「モデルの指定」章にある例は、酸素摂取量と走行時間の間に有意な線形関係があることを示しています。「モデルのあてはめ」の詳細については、『基本的な回帰モデル』の「モデルの指定」章を参照してください。「二変量の関係」の詳細については、『基本的な統計分析』の「「二変量の関係」プラットフォームの概要」章を参照してください。

「曲線のあてはめ」プラットフォームでは、パラメータに関して**非線形**であるモデルをあてはめることができます。この章の最初の例では、非線形な関係の例として、薬剤の毒性を濃度の関数としてモデル化します。毒性に対する濃度の効果は、低用量から高用量へと変化するため、関係は非線形です。

以下は、パラメータに関して線形なモデルと、パラメータに関して非線形なモデルの例です。

- パラメータに関して線形: $Y=\beta_0+\beta_1 e^{x}$
- パラメータに関して非線形: $Y=\beta_0+\beta_1 e^{\beta_2 x}$

「曲線のあてはめ」プラットフォームには、多項式・ロジスティック曲線・プロビット曲線・Gompertz曲線、指数モデル・ピークモデル・薬物動態モデルなど、いくつかのモデルが用意されています。また、グループ変数を指定すると、そのグループ変数の水準ごとに、モデルのパラメータが推定されます。このとき、パラメータ推定値を、グループ変数の水準間で比較することもできます。

「曲線のあてはめ」では、データテーブルに予測式を保存することもできます。保存した予測式を「非線形回帰」プラットフォームで利用すれば、パラメータ値に対して上限や下限を設定することもできます。詳細は、「非線形回帰」章の「パラメータの範囲を設定する例」（213ページ）を参照してください。

「曲線のあてはめ」機能の使用例

この例では、薬剤の毒性を濃度の関数としてモデル化します。標準的な製剤を、3つの新しい製剤と比較したいとしましょう。

各薬剤で、特定の濃度における生存細胞と非生存細胞の比を調べます。調査では、製剤ごとに16点における濃度の毒性を調べます。生存細胞の非生存細胞に対する比は、その値が低いほど毒性が高いことを意味します。毒性が高いと、その薬剤は好ましくありません。なお、この例では、濃度の範囲を狭くし、曲線の差を検出しやすくするため、濃度そのものではなく、濃度の対数をモデルで用います。

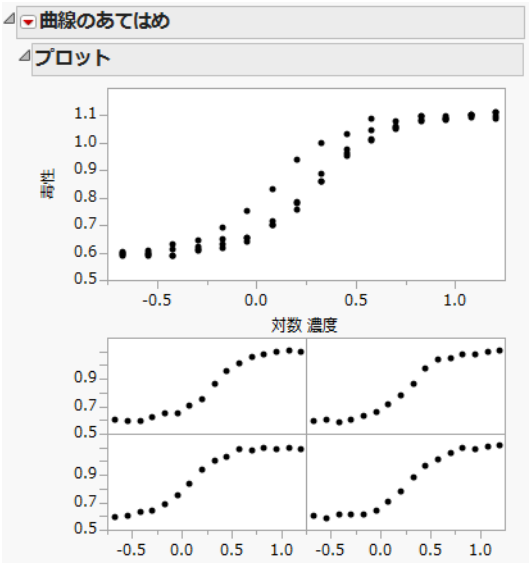
以下の手順に従ってモデルを作成します。

1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Nonlinear Examples¥Bioassay.jmp」を開きます。
2. ［**分析**］>［**発展的なモデル**］>［**曲線のあてはめ**］を選択します。
3. 「**毒性**」を［**Y, 応答変数**］に指定します。

4. 「**対数 濃度**」を［**X, 説明変数**］に指定します。
5. 「**製剤**」を［**グループ化**］に指定します。
6. ［**OK**］をクリックします。

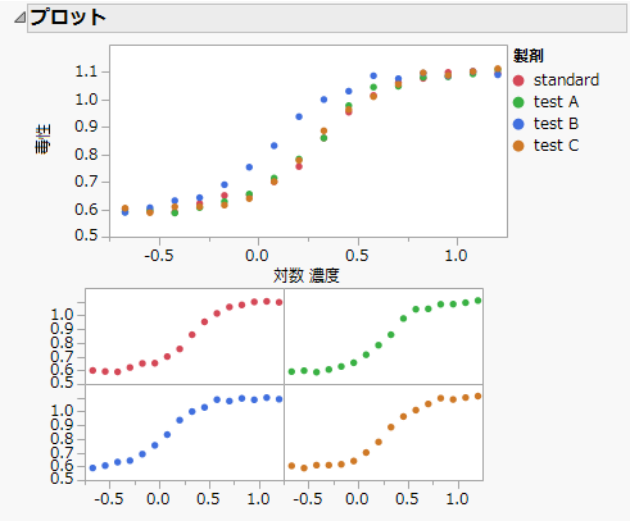
「曲線のあてはめ」レポートが表示されます（図12.2）。「プロット」レポートには、製剤ごとにあてはめられたモデルが重ねて表示されます。

図12.2 最初の「曲線のあてはめ」レポート



7. 薬剤の製剤を示す凡例を表示するため、いずれかのグラフを右クリックし、［**行の凡例**］を選択します。「**製剤**」の列を選択し、［**OK**］をクリックします。プロットが図12.3のように表示されます。

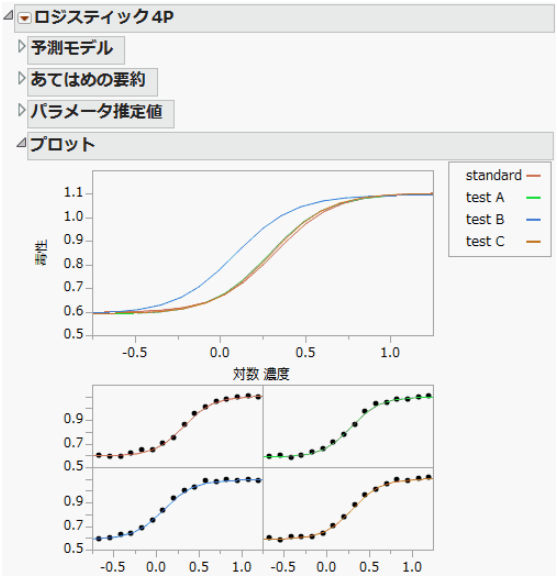
図12.3「曲線のあてはめ」レポートとプロットの凡例



曲線がS字型であることから、シグモイド曲線が適切であると判断できます。表12.1に、「曲線のあてはめ」機能で使用できる各種モデルのプロット例と計算式がまとめられています。

8. 「曲線のあてはめ」の赤い三角ボタンをクリックし、**[シグモイド曲線] > [ロジスティック曲線] > [ロジスティック 4Pのあてはめ]**を選択します。

図12.4「ロジステイック 4P」レポート

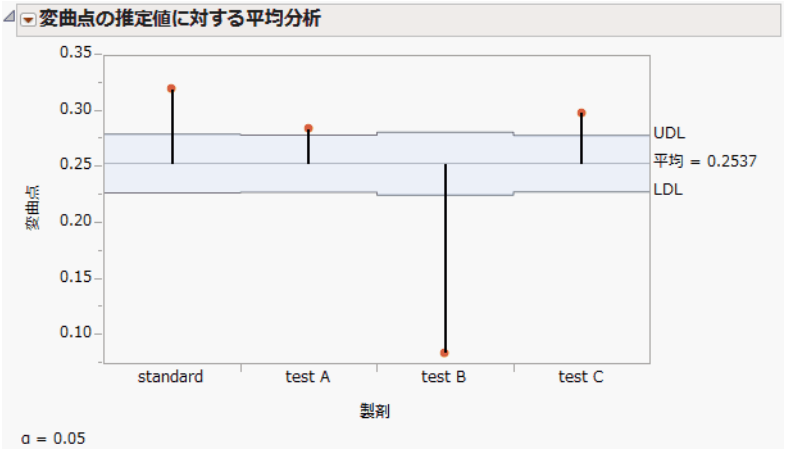


「ロジスティック 4P」レポートが表示されます（図12.4）。また、製剤ごとの、個別のプロットも表示されます。プロットの曲線を見ると、製剤Bは、他の薬剤と大きく異なっていることがわかります。「test B」の曲線は、他の曲線よりも早くに上昇し始めています。曲線の上昇がどれぐらいから始まるかは、変曲点のパラメータによって決まります。

9. 「ロジスティック 4P」の赤い三角ボタンをクリックし、［**パラメータ推定値の比較**］を選択します。

「パラメータの比較」レポートの一部を図12.5に示します。

図12.5「パラメータの比較」レポート

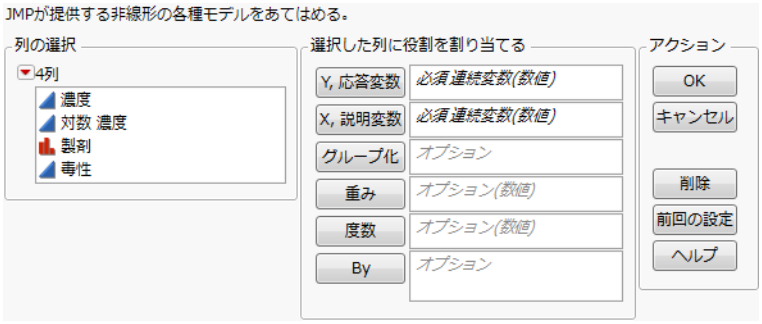


「test B」の製剤の変曲点が、全体の平均よりも、かなり低いことがわかります。これは、図12.4のプロットの様子と一致しています。製剤Bは、他の製剤より毒性が弱くなっています。

「曲線のあてはめ」プラットフォームの起動

「曲線のあてはめ」プラットフォームを起動するには、［**分析**］>［**発展的なモデル**］>［**曲線のあてはめ**］を選択します。図12.6に起動ウィンドウを示します。

図12.6「曲線のあてはめ」プラットフォームの起動ウィンドウ



「曲線のあてはめ」プラットフォームの起動ウィンドウには以下の機能があります。

Y, 目的変数　Y変数を選択します。

X, 説明変数　X変数を選択します。

グループ化　グループ変数を指定します。グループ変数の水準ごとに、個別にモデルがあてはめられます。グループ変数を指定すると、あてはめたモデルとパラメータ推定値を、グループ変数の水準間で比較できます。

重み　オブザベーションの重みを含む変数を指定します。

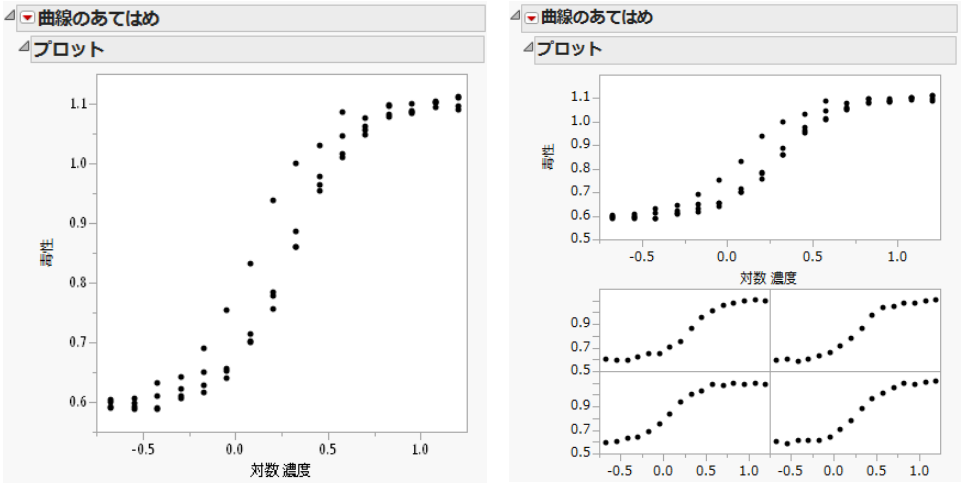
度数　オブザベーションの度数を含む変数を指定します。

By　ここで指定した変数の水準ごとに個別に分析が行われます。

「曲線のあてはめ」レポート

「曲線のあてはめ」レポートのプロットには、最初はXとYのデータ値しかプロットされていません（図12.7）。グループ変数を指定すると、レポートには、すべてのモデルを重ねて表示したプロットと、グループ別のプロットが表示されます（図12.7の右側）。

図12.7「曲線のあてはめ」レポート : グループ変数未指定（左）とグループ変数指定（右）



「曲線のあてはめ」の赤い三角ボタンのメニューから、次のいずれかのモデルを選択できます。

多項式　直線から5次までの多項式をあてはめます。

シグモイド曲線　ロジスティック、プロビット、Gompertz といったモデルをあてはめます。これらのモデルはS字型で、上側と下側に漸近線があります。ロジスティック2P、3P、4P、およびプロビット2P、4Pの各モデルは対称です。ロジスティック5Pモデルと両方のGompertzモデルは非対称です。ロジスティック2Pは、応答のデータ値が、0以上1以下の場合にだけ使用できます。シグモイド曲線の例としては、学習曲線や、腫瘍の成長を示すモデルが挙げられます。どちらも、最初は上昇し、やがて停滞します。

指数 成長と減衰　指数、双指数、単分子成長のモデルをあてはめます。「指数 2P」は漸近線が0に固定されていますが、「指数 3P」ではその漸近線が推定されます。双指数モデルは、成長過程（もしくは、減衰過程）を表す指数関数が2つあります。単分子成長モデルは、指数 3Pモデルを、別のパラメータ表現で表したモデルです。指数成長の例としてはウイルスの拡散、減衰関数の例としては薬物の半減期が挙げられます。

ピークモデル　Gauss型ピークとLorentz型ピークのモデルをあてはめます。これらのモデルは、ピークまで増加し、その後減少します。Gauss型ピークモデルは、正規分布の確率密度関数のスケールを変更したものです。Lorentz型ピークモデルは、連続確率分布のCauchy分布のスケールを変更したものです。これらの曲線は、一部の化学濃度における測定や人工ニューラルネットワークで利用されています。

薬物動態モデル　1コンパートメント経口投与、2コンパートメント急速静注モデル、および双指数 4Pのモデルをあてはめます。このオプションは、体内の薬剤濃度をモデル化するときに使用します。

Michaelis-Mentenのあてはめ　Michaelis-Menten生化学動態モデルをあてはめ、基質濃度に対する酵素反応速度を特定します。

以下のオプションについて詳しくは、『JMPの使用法』の「JMPのレポート」章を参照してください。

ローカルデータフィルタ　現在のレポートにおいてデータをフィルタリングするローカルデータフィルタを表示するか、非表示にします。

やり直し　分析を繰り返すオプションや、やり直すオプションを含みます。また、［自動再計算］オプションを選択すると、このオプションに対応しているプラットフォームにおいて、データテーブルに加えた変更が、該当するレポートに即座に反映されるようになります。

スクリプトの保存　レポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。

Byグループのスクリプトを保存　By変数の全水準に対するレポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。起動ウィンドウでBy変数を指定した場合のみ使用可能です。

最初の「曲線のあてはめ」レポート

モデルをあてはめる前、「曲線のあてはめ」レポートのプロットには、XとYのデータ値しかプロットされていません。モデルをあてはめると、そのモデルの曲線がプロットに追加されます（プラットフォームの起動ウィンドウでグループ変数を指定していない場合）。レポートには、以下の結果が表示されます。

「モデルの比較」レポート

図12.8に示すレポートを作成するには、「曲線のあてはめ」の赤い三角ボタンのメニューから、**［シグモイド曲線］＞［ロジスティック曲線］＞［ロジスティック4Pのあてはめ］**および**［シグモイド曲線］＞［ロジスティック曲線］＞［Gompertz 4Pのあてはめ］**を選択します。

図12.8「モデルの比較」レポート

曲線のあてはめ

モデルの比較

モデル	AICc	AICc 重み	.2 .4 .6 .8	BIC	SSE	MSE	RMSE	R2乗
ロジスティック 4P	-372.3021	1		-348.9054	0.005325	0.0001109	0.0105327	0.9980584
Gompertz 4P	-327.5372	1.903e-10		-304.1405	0.0107173	0.0002233	0.0149425	0.9960922

「モデルの比較」レポートには、複数のモデルを比較するための適合度統計量が表示されます。AICc、AICc重み、BIC、SSE、MSE、RMSE、R2乗が求められます。それぞれの定義を以下に紹介します。

AICc　推定した統計モデルの適合度を示す統計量で、2つ以上のモデルを比較したいときに使用できます。AICcは、標本サイズが少ない場面を考慮して、通常のAICを調整しています。標本サイズがパラメータ数より2以上大きくないと、AICcを求めることはできません。AICcの値が最も小さいモデルが最良であるため、この例ではロジスティック4Pが最良のモデルです。『基本的な回帰モデル』の付録にある「統計的詳細」を参照してください。

AICc 重み 合計が1になるように、AICcの値を正規化したものです。AICc重みは、あてはめたモデルのいずれかが真である場合に特定のモデルが真である確率と解釈できます。そのため、AICc重みが1に最も近いモデルが良いモデルを意味します。この例において良いモデルは、明らかにロジスティック4Pです。AICc重みは、複数のモデルのAICcから、次のように算出されます。

AICcの重み = exp[-0.5(AICc-min(AICc))] / sum(exp[-0.5(AICc-min(AICc))])

上の式で、min(AICc)は、あてはめたモデルの中で最も小さいAICc値です。「モデルの比較」表は、AICc重みの降順に並べ替えられます。

BIC 推定した統計モデルの適合度を示す統計量で、2つ以上のモデルを比較したいときに使用できます。BICの値が小さいほど、良いモデルです。『基本的な回帰モデル』の付録にある「統計的詳細」を参照してください。

SSE 観測値と予測値との差の平方和。

MSE 誤差の平方平均。

RMSE MSEの平方根で、誤差の標準偏差に対する推定値です。

R2乗 応答の変動のうち、偶然誤差ではなく、モデルによって説明される変動の割合を示します。R2乗値が1に近いほど、あてはまりが良いモデルです。

「モデルの比較」プラットフォームには、残差や実測値のプロットなど、追加のオプションがあります。詳細については、「モデルの比較」章（155ページ）を参照してください。

モデルのレポート

あてはめた各モデルのレポートが作成されます。各モデルレポートには、次のセクションがあります。

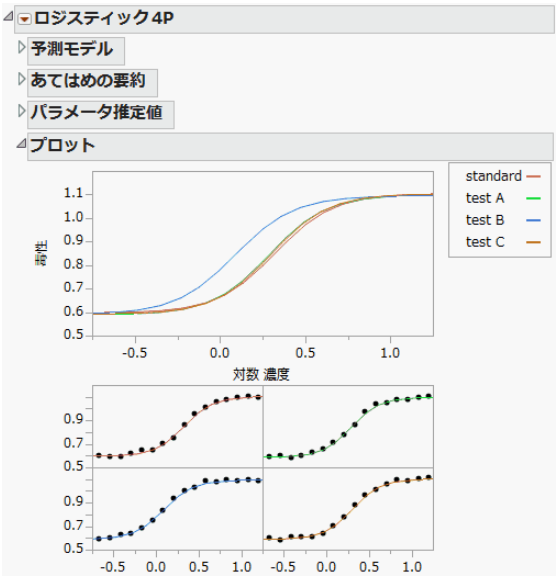
予測モデル 予測式とパラメータが、代数的な数式の形式で示されます。

あてはめの要約 「モデルの比較」レポートと同じ適合度統計量が表示されます。

パラメータ推定値 パラメータ推定値と、その標準誤差および信頼限界が表示されます。また、推定値の相関係数や共分散も表示されます。

プロット データとあてはめたモデルのプロットが表示されます。図12.9を参照してください。このプロットは、起動ウィンドウでグループ変数を指定した場合のみ表示されます。

図12.9 ロジスティック4Pモデルのレポート



各モデルのレポートにある赤い三角ボタンをクリックすると、次のようなオプションを含んだメニューが開きます。

平行性の検定　各曲線が、形状はまったく同じでX軸に沿ってずれているだけかどうかを判断するのに役立ちます。状況によっては、グループ間で詳細な比較を行う前に、平行性が成り立っているかどうかを調べることが重要な場合があります。このオプションは、起動ウィンドウでグループ変数を指定した場合にのみ使用できます。このオプションは、シグモイドモデル（ロジスティック、Gompertz）と、線形回帰モデルで使用できます（高次の多項式を除く）。詳細は、「平行性の検定」（189ページ）を参照してください。

AUC 曲線下面積　あてはめたモデルの曲線から、AUC（Area Under Curve;曲線下の面積）を計算します。このオプションは、1コンパートメント、2コンパートメント、Gauss型ピーク、Lorentz型ピークだけで使用できます。積分範囲は、モデルによって異なりますが、レポートに表示されます。

プラットフォームの起動ウィンドウでグループ変数を指定した場合は、AUCを群間比較する平均分析が実行されます。あるグループの結果が決定限界を超える場合、そのグループのAUCは、AUCの全体平均とは異なるとみなせます。

パラメータ推定値の比較　各グループのパラメータが、パラメータの全体平均と異なるかどうかを検定します。このオプションは、起動ウィンドウでグループ変数を指定した場合のみ使用できます。詳細は、「パラメータ推定値の比較」（191ページ）を参照してください。

同等性の検定　各グループ間において、パラメータの同等性検定を行います。このオプションは、起動ウィンドウでグループ変数を指定した場合のみ使用できます。詳細は、「同等性の検定」（192ページ）を参照してください。

パラメータテーブルの作成 パラメータ推定値、標準誤差、t値をデータテーブルに保存します。このオプションは、起動ウィンドウでグループ変数を指定した場合のみ使用できます。

予測値と実測値のプロット 縦軸にYの実測値、横軸にYの予測値をプロットします。

予測値と残差のプロット 縦軸に残差、横軸にYの予測値をプロットします。

プロファイル 予測式のプロファイルの表示／非表示が切り替わります。微分のプロファイルも含まれますが、それはX変数で予測式を微分したものです。プロファイルの詳細については、『プロファイル機能』の「プロファイル」章を参照してください。

計算式の保存 データテーブルに各種計算式の列を保存するオプションがあります。

予測式の保存 予測式を保存します。

予測値の標準誤差の保存 予測値の標準誤差を保存します。

パラメトリックな予測式の保存 パラメトリックな予測式を保存します。このオプションで保存された計算式は、「非線形回帰」プラットフォームで自分自身でモデルを作成する際に利用できます。

残差計算式の保存 残差を保存します。

スチューデント化残差計算式の保存 スチューデント化残差の計算式を保存します。スチューデント化残差は、残差をその標準偏差の推定値で割ったものです。

1 次微分の保存 X 変数に関する予測式の微分を、計算式として保存します。

1 次微分の標準誤差の保存 1 次微分の標準誤差の計算式を保存します。

逆推定計算式の保存 Y から X を予測するための計算式を保存します。

カスタム逆推定 特定のY値に対してX値を予測します。逆推定の詳細については、『基本的な回帰モデル』の標準最小2乗に関する章を参照してください。

あてはめの削除 モデルのレポート、「モデルの比較」レポートの該当部分、プロット上のあてはめ線を削除します。

「曲線のあてはめ」のオプション

モデル式

表12.1に、「曲線のあてはめ」の赤い三角ボタンのメニューで用意されているモデルの式を示します。

表12.1「曲線のあてはめ」のモデル計算式

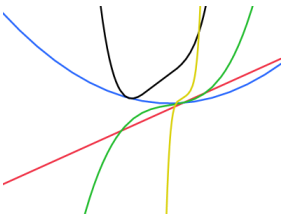
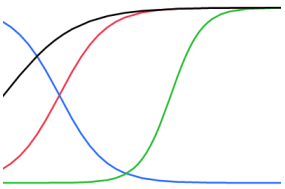
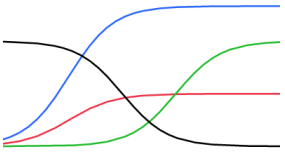
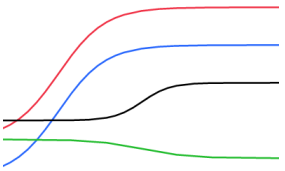
モデル	計算式
多項式 	$\beta_0 + \sum_{i=1}^{k} \beta_i x^i$ kは多項式の次数です。これらのモデルは、「モデルのあてはめ」プラットフォームや「二変量の関係」プラットフォームでもあてはめることができます。
ロジスティック 2P 	$\frac{1}{1 + \mathrm{Exp}(-a(x-b))}$ a = 増加率 b = 変曲点 応答変数のすべての値が0～1である場合にのみ使用可能です。
ロジスティック 3P 	$\frac{c}{1 + \mathrm{Exp}(-a(x-b))}$ a = 増加率 b = 変曲点 c = 漸近線
ロジスティック 4P 	$c + \frac{d-c}{1 + \mathrm{Exp}(-a(x-b))}$ a = 増加率 b = 変曲点 c = 下側漸近線 d = 上側漸近線

表12.1「曲線のあてはめ」のモデル計算式（**続き**）

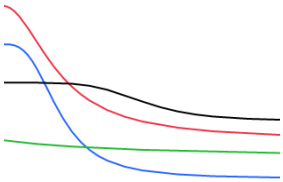
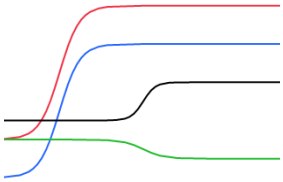
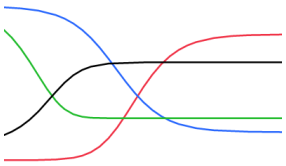
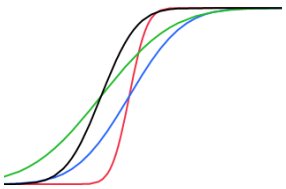
モデル	計算式
ロジスティック 4P Rodbard 	$c + \frac{d-c}{1+(x/b)^a}$ a = 増加率 b = 変曲点 c = 下側漸近線 d = 上側漸近線 説明変数の値が正の場合にのみ使用可能です。
ロジスティック 4P Hill 	$c + \frac{d-c}{1+10^{(-a(x-b))}}$ a = 増加率 b = 変曲点 c = 下側漸近線 d = 上側漸近線
ロジスティック 5P 	$c + \frac{d-c}{(1+\mathrm{Exp}(-a(x-b)))^f}$ a = 増加率 b = 変曲点 c = 漸近線1 d = 漸近線2 f = べき乗
プロビット 2P 	$\Phi\left(\frac{x-b}{a}\right)$ a = 増加率 b = 変曲点 Φ = 正規分布累積確率プロット 応答変数のすべての値が0～1である場合にのみ使用可能です。

表12.1「曲線のあてはめ」のモデル計算式（**続き**）

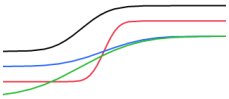
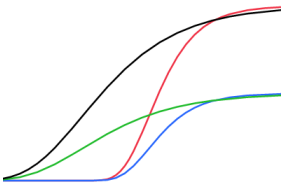
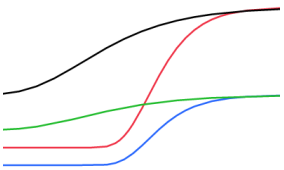
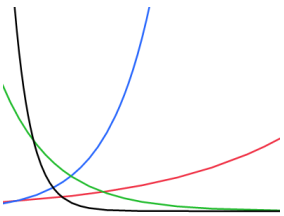
モデル	計算式
プロビット 4P 	$c + (d - c) \cdot \Phi\left(\frac{x - b}{a}\right)$ a = 増加率 b = 変曲点 c = 漸近線1 d = 漸近線2 Φ = 正規分布累積確率プロット
Gompertz 3P 	$a\text{Exp}(-\text{Exp}(-b(x - c)))$ a = 漸近線 b = 増加率 c = 変曲点
Gompertz 4P 	$a + (b - a)\text{Exp}(-\text{Exp}(-c(x - d)))$ a = 下側漸近線 b = 上側漸近線 c = 増加率 d = 変曲点
指数 2P 	$a\text{Exp}(bx)$ a = スケール b = 増加率

表12.1「曲線のあてはめ」のモデル計算式（**続き**）

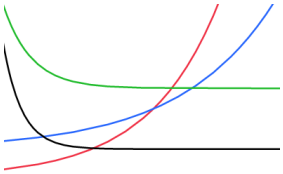
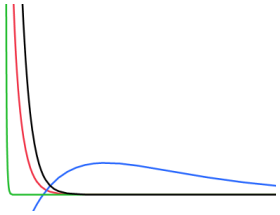
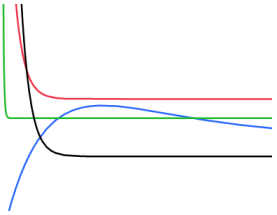
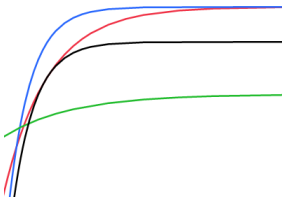
モデル	計算式
指数 3P 	$a + b\text{Exp}(cx)$ a = 漸近線 b = スケール c = 増加率
双指数 4P 	$a\text{Exp}(-bx) + c\text{Exp}(-dx)$ a = スケール1 b = 減衰率1 c = スケール2 d = 減衰率2 応答変数の値が正の場合にのみ使用可能です。
双指数 5P 	$a + b\text{Exp}(-cx) + d\text{Exp}(-fx)$ a = 漸近線 b = スケール1 c = 減衰率1 d = スケール2 f = 減衰率2
単分子成長 	$a(1 - b\text{Exp}(-cx))$ a = 漸近線 b = スケール c = 増加率

表12.1「曲線のあてはめ」のモデル計算式（続き）

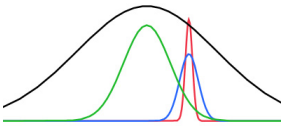
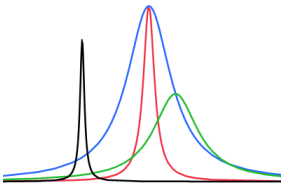
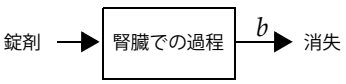
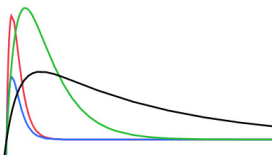
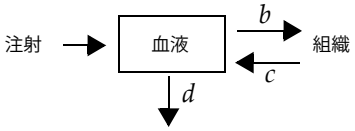
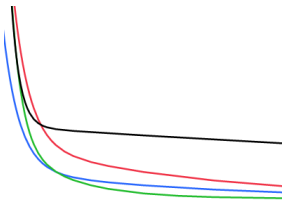
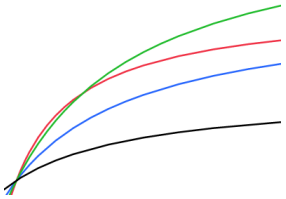
モデル	計算式
Gauss型ピーク 	$a\mathrm{Exp}\left(-\frac{1}{2}\left(\frac{x-b}{c}\right)^2\right)$ a = ピーク値 b = 臨界点 c = 増加率
Lorentz型ピーク 	$\frac{ab^2}{(x-c)^2+b^2}$ a = ピーク値 b = 増加率 c = 臨界点
1コンパートメント 経口投与  	$\frac{abc}{c-b}(\mathrm{Exp}(-bx)-\mathrm{Exp}(-cx))$ a = AUC（曲線下面積） b = 消失速度 c = 吸収速度 応答変数の値と説明変数の値がすべて正の場合にのみ使用可能です。

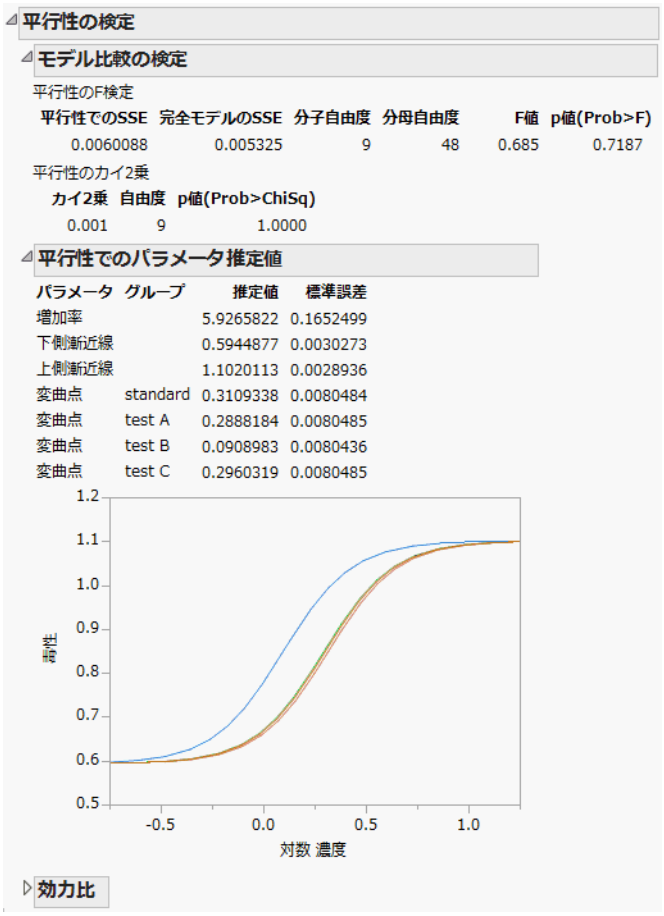
表12.1「曲線のあてはめ」のモデル計算式（**続き**）

モデル	計算式
2コンパートメント 急速静注  	$\frac{a}{\alpha-\beta}((\alpha-b)\mathrm{Exp}(-\alpha x)-(\beta-b)\mathrm{Exp}(-\beta x))$ $\alpha=\frac{1}{2}(b+c+d+\sqrt{(b+c+d)^2-4bd})$ $\beta=\frac{1}{2}(b+c+d-\sqrt{(b+c+d)^2-4bd})$ a = 初期濃度 b = 移行速度 流入 c = 移行速度 流出 d = 消失速度 応答変数の値と説明変数の値がすべて正の場合にのみ使用可能です。
Michaelis-Menten 	$\frac{ax}{b+x}$ a = 最大反応速度 b = 解離定数 応答変数の値と説明変数の値がすべて正の場合にのみ使用可能です。

平行性の検定

［平行性の検定］オプションを選択すると、あてはめたモデルがグループ間で同じ形状を示していて、X軸に沿ってずれているだけかどうかが検定されます（図12.10）。「Bioassay.jmp」の例では、製剤Bの曲線が、他の3つの左にずれています。しかし、すべての曲線が同じ形状なのか（つまり、平行にずれているだけなのか）、製剤Bの形状が異なるのかは、わかりません。平行性の検定を行うと、異なる製剤の曲線について、形状が似ているかどうか、横軸に沿ってずれているかどうかを判断できます。あてはめたモデルの赤い三角ボタンのメニューから［**平行性の検定**］を選択すると、レポートが追加されます。

図12.10 平行性の検定



このレポートには、以下の結果が表示されます。

モデル比較の検定 平行性に対するF検定およびカイ2乗検定の結果が表示されます。F検定は、完全モデルと縮小モデルの誤差平方和を比較したものです。この検定での完全モデルは、グループごとのパラメータがすべて異なっているモデルです。一方、縮小モデルは、変曲点以外のすべてのパラメータが、すべてのグループで同じモデルです。この例では、p値が0.05より大きいため、曲線の間に差があるという十分な証拠はありません。

平行性でのパラメータ推定値 縮小モデルにおけるパラメータ推定値が表示されます。また、縮小モデルの曲線もプロットされます。ここでの縮小モデルは、すべてのグループにおいて、変曲点以外のすべてのパラメータが同じモデルです。製剤Bの変曲点は、他の3つの製剤の変曲点よりずっと低い位置にあります。

効力比 グループ変数の水準ごとに、効力比（relative potency）を計算します。効力は10^（EC_{50}）です。ここで、EC_{50}は応答の基準値と最大値の中間値が得られる用量です。ロジスティック2P、3P、4Pの場合、効力は、変曲点パラメータを指数変換した値です。効力比は、これらの効力の比として計算されます。

図12.11 グループ別の効力比

効力比

standard に対する効力比

グループ	効力	効力比	標準誤差
standard	2.0461329	1	0
test A	1.9445466	1.0522416	0.0246535
test B	1.2328161	1.6597227	0.0388878
test C	1.977115	1.0349084	0.0242474

test A に対する効力比

グループ	効力	効力比	標準誤差
standard	2.0461329	0.950352	0.0222663
test A	1.9445466	1	0
test B	1.2328161	1.5773209	0.036957
test C	1.977115	0.9835273	0.0230435

test B に対する効力比

グループ	効力	効力比	標準誤差
standard	2.0461329	0.6025103	0.014117
test A	1.9445466	0.6339864	0.0148545
test B	1.2328161	1	0
test C	1.977115	0.6235429	0.0146098

test C に対する効力比

グループ	効力	効力比	標準誤差
standard	2.0461329	0.9662691	0.0226392
test A	1.9445466	1.0167486	0.0238219
test B	1.2328161	1.6037388	0.037576
test C	1.977115	1	0

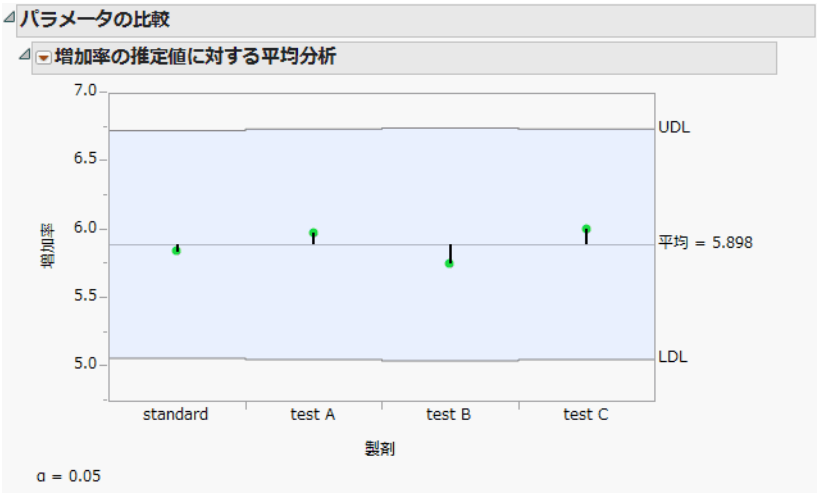
図12.11の［**standard に対する効力比**］パネルを見ると、製剤Aと製剤Cの効力比はほぼ1です。これは、2つの製剤の毒性が標準の製剤に近いことを意味します。製剤Bの効力はstandardより低い値です。つまり、製剤Bの毒性は、濃度に対してstandardよりも速く増加します。

平行性の検定からは、曲線が平行であることが示唆されており、もし平行性が成立していれば効力比が意味を持ちます。効力比を見る限り、製剤Bの毒性は他の製剤より強いことが示唆されています。これらの結果と過去の試験結果と総合して、製剤Bの毒性が最も高いかどうかを考えることができます。

パラメータ推定値の比較

「パラメータ推定値の比較」レポートは、各グループのパラメータが、パラメータの全体平均と異なるかどうかを検定した結果です。各パラメータが全体平均に等しいかどうかを検定するために、平均分析（ANOM; ANalysis Of Means）が行われます。あるパラメータ推定値が決定限界を超えている場合、そのパラメータは全体平均に等しくないと言えます。図12.12は、「増加率の推定値に対する平均分析」レポートです。あてはめたモデルの赤い三角ボタンのメニューから［**パラメータ推定値の比較**］を選択すると、レポートが追加されます。

図12.12 増加率の推定値に対するパラメータの比較



平均分析レポートのタイトルバーにある赤い三角ボタンをクリックすると、次のようなオプションが表示されます。

有意水準の設定　この検定のα水準を設定します。

要約レポートの表示　パラメータ推定値、決定限界、およびパラメータが決定限界を超えているかどうかを示すレポートの表示／非表示が切り替わります。

表示オプション　決定限界、陰影、および中心線の表示／非表示を切り替えるためのオプションがあります。また、点の表示方法を変更するためのオプションもあります。

「平均分析」レポートの詳細については、『基本的な統計分析』の「一元配置分析」章を参照してください。

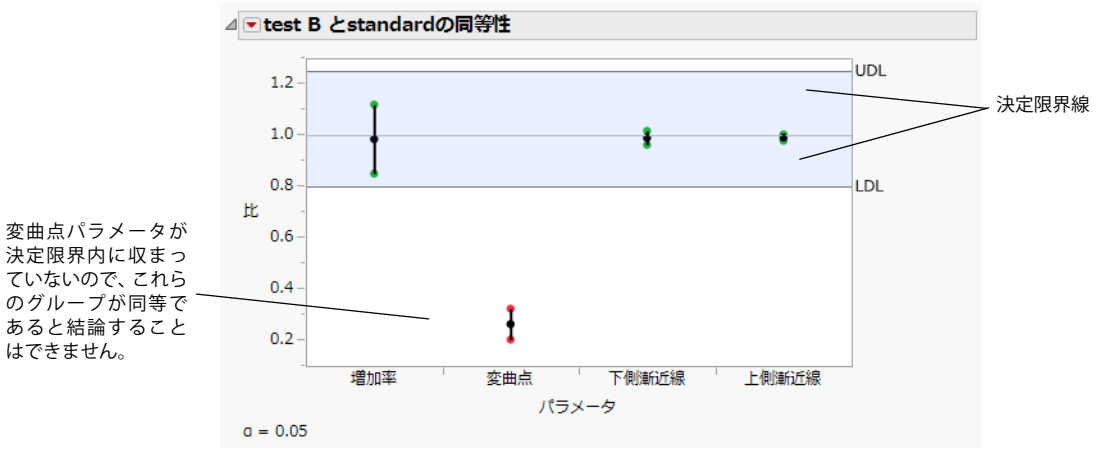
同等性の検定

「同等性の検定」レポートには、グループ間でパラメータの同等性検定を行った結果が表示されます(図12.13)。オプションを選択した後、基準もしくは標準とするグループ変数の水準を指定してください。そこで指定した水準と、他の各水準が比較されます。あてはめたモデルの赤い三角ボタンのメニューから［**同等性の検定**］を選択すると、このレポートが追加されます。

パラメータの同等性は、パラメータの比を分析することで検定されます。デフォルトでは、25%の差異を表す比（0.8と1.25）の位置に決定限界線が置かれます。

すべての信頼区間が決定限界線の内側にある場合は、2つのグループのモデルが実質的に等しいことを意味します。一方で、いずれかの信頼区間が決定限界内に収まっていない場合（図12.13を参照）は、「2つのグループのモデルは等しい」と結論することができません。製剤Bの変曲点は、standardより低く、過去の分析結果が示す傾向と一致しています。

図12.13 同等性の検定



同等性レポートの赤い三角ボタンのメニューには、次のようなオプションがあります。

有意水準の設定　この検定のα水準を設定します。デフォルトの値は0.05です。

決定限界線の設定　比の決定限界線を変更します。デフォルトでは、25%の差異を表す0.8および1.25の位置に設定されています。

要約レポートの表示　パラメータ推定値、決定限界、およびパラメータが決定限界を超えているかどうかを示すレポートの表示／非表示が切り替わります。

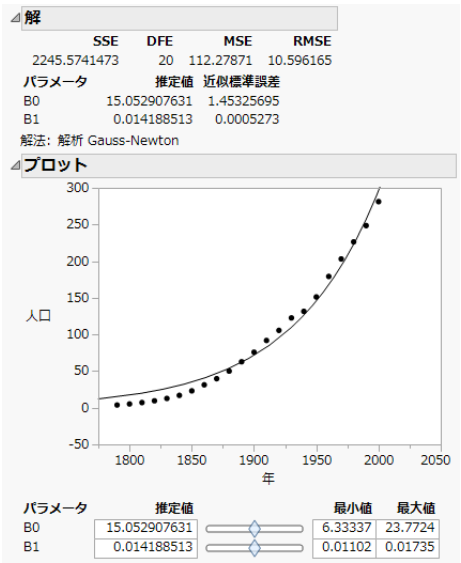
表示オプション　決定限界、陰影、および中心線の表示／非表示を切り替えるためのオプションがあります。また、点の表示方法を変更するためのオプションもあります。詳細については、グラフを右クリックし、[**カスタマイズ**]を選択してください。

第13章

非線形回帰
独自に定義した非線形モデルをあてはめる

「非線形回帰」プラットフォームは、パラメータに関して**非線形**であるモデルに適しています。この章では、非線形モデルのモデル式を分析者自身で設定する方法について説明します。この場合、まず、モデル式として、推定対象のパラメータを含む計算式を作成する必要があります。推定方法としては最小2乗法がデフォルトで使われますが、それだけでなく、分析者自身が独自の損失関数を定義することもできます。この場合、損失関数の合計を最小にするパラメータが求められます。

図13.1 独自に作成した非線形モデルの例



なお、「非線形回帰」や「曲線のあてはめ」プラットフォームには、多項式・ロジスティック曲線・Gompertz曲線・指数モデル・ピークモデル・薬物動態モデルなど、いくつかのモデルがあらかじめ用意されており、それらの非線形回帰モデルに関してはユーザがモデル式を設定する必要はありません。詳細については、「曲線のあてはめ」章（173ページ）を参照してください。

メモ: モデルの中には、パラメータに関して**線形**であるもの（たとえば2次式などの多項式）や、線形に変換できるもの（たとえばxを対数変換するなど）があります。そのようなモデルには、「モデルのあてはめ」プラットフォームや「二変量の関係」プラットフォームがより適しています。これらのプラットフォームの詳細については、『基本的な回帰モデル』の「モデルの指定」章と、『基本的な統計分析』の「「二変量の関係」プラットフォームの概要」章を参照してください。

独自の非線形回帰モデルを設定する例

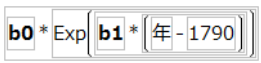
分析者が独自に考えた非線形モデルを推定するには、まず、計算式を含んだ列を作成する必要があります。この計算式で、パラメータとその初期値も設定します。この方法は、操作手順が少し複雑になりますが、どのような非線形モデルでも指定できます。また、損失関数をユーザ自身で定義したり、反復計算に関する詳細なオプションを選択したりすることもできます。

この節では、モデルの計算式を含む列を作成し、「非線形回帰」プラットフォームを実行する例を述べます。使用するデータは「US Population.jmp」データテーブルです。応答変数は、米国の人口（単位は百万人）で、説明変数は年です。

1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Nonlinear Examples¥US Population.jmp」を開きます。
2. 新しい列を作成し、その列名を「**モデル**」とします。
3. 「**モデル**」列を右クリックして、［**列プロパティ**］>［**計算式**］を選択します。

   計算式エディタが表示されます。
4. 左側の列のリストの上にあるドロップダウンリストから［**パラメータ**］を選択します。
5. ［**パラメータの新規作成**］を選択します。
6. デフォルト名のb0を使用します。
7. 「**値**」に「4」と入力します。これがパラメータの初期推定値です。
8. ［**OK**］をクリックします。
9. ［**パラメータの新規作成**］を選択します。
10. 名前はデフォルトのままにし、「**値**」に「0.02」と入力します。
11. ［**OK**］をクリックします。
12. 計算式エディタの関数、「**年**」列、およびパラメータを使用してモデル計算式を入力します。入力が完了したモデル計算式は図13.2のようになります。

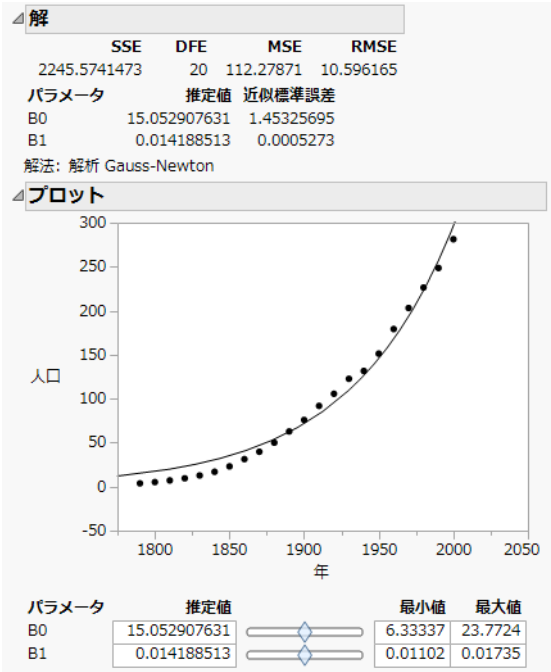
図13.2 入力が完了したモデル計算式



13. ［**OK**］をクリックします。
14. ［**分析**］>［**発展的なモデル**］>［**非線形回帰**］を選択します。
15. 「**モデル**」を［**X, 予測式列**］に指定します。
16. 「**人口**」を［**Y, 応答変数**］に指定します。
17. ［**OK**］をクリックします。
18. 設定パネルの［**実行**］ボタンをクリックします。

レポートの一部を図13.3に示します。

図13.3「プロット」と「解」レポート



「**解**」レポートには、最終的なパラメータ推定値とその他の適合度統計量が表示されます。プロットには、あてはめたモデルが表示されます。

グループ変数を含むモデルのパラメータ

計算式エディタでパラメータを追加するときに、[**選択された列をカテゴリに展開する**]というチェックボックスを使用できます。このオプションは、カテゴリカル変数の各水準ごとのパラメータを一度に作成するものです（複数のパラメータが一度に追加されます）。このオプションを選択すると、列を選択するためのダイアログボックスが開きます。列の選択が完了すると、パラメータのリストに「**D_***列名*」という名前の新しいパラメータが表示されます。「**D**」の部分は、パラメータの名前です。計算式にこのパラメータを使用すると、グループ変数の水準ごとに個別のパラメータを含んだMatch式が挿入されます。

「非線形回帰」プラットフォームの起動

「非線形回帰」プラットフォームを起動するには、［分析］>［発展的なモデル］>［非線形回帰］を選択します。図13.4に起動ウィンドウを示します。

図13.4「非線形回帰」起動ウィンドウ

「非線形回帰」プラットフォームの起動ウィンドウには次のような機能があります。

Y, 目的変数　Y変数を選択します。

X, 予測式列　X変数のデータ値を含んでいる列、もしくは、モデルの計算式を含んでいる列を選択します。

グループ化　グループ変数を指定します。グループ変数の水準ごとに、個別にモデルがあてはめられます。グループ変数を指定すると、あてはめたモデルとパラメータ推定値を、グループ変数の水準間で比較できます。

重み　オブザベーションの重みを含む変数を指定します。

度数　オブザベーションの度数を示す変数を指定します。

損失　損失関数を含む計算式列を指定します。

By　ここで指定した変数の水準ごとに個別に分析が行われます。

モデルライブラリ　モデルライブラリツールを起動します。初期値を選択して計算式列を作成できます。「モデルライブラリを使用した列の作成」（205ページ）を参照してください。

数値微分のみ 反復計算の微分において、数値微分だけを使用します。このオプションは、モデルが複雑で解析的な微分を求めるのが困難であるような場合に効果的です。解析的微分に基づく方法が収束しないようなケースに対しても役に立ちます。このオプションは、[X, 予測式列] に計算式列が指定されている場合のみ使用します。

中間計算式の展開 モデルに使われている列の中に、さらに計算式が含まれている場合に、(他の列を参照している) 内側の式が代入されます。このオプションを使うときに、特定の列だけ展開されないようにするには、列プロパティで**「その他」**を選択し、「計算式の展開」という名前の列プロパティを作成して、その値を0とします。このオプションは、[X, 予測式列] に計算式列が指定されている場合のみ使用します。

「非線形回帰のあてはめ」レポート

最初に表示される「非線形回帰のあてはめ」レポートには、次のような項目が含まれます (図13.5)。

設定パネル 反復計算の処理を制御します。

実行 反復計算を開始します。

停止 反復計算を途中で停止します。

ステップ 反復計算を1ステップだけ進めます。

リセット 反復計算の値がリセットされます。フィールドに入力したパラメータ値がモデル式に代入され、それに対する SSE が再計算されます。

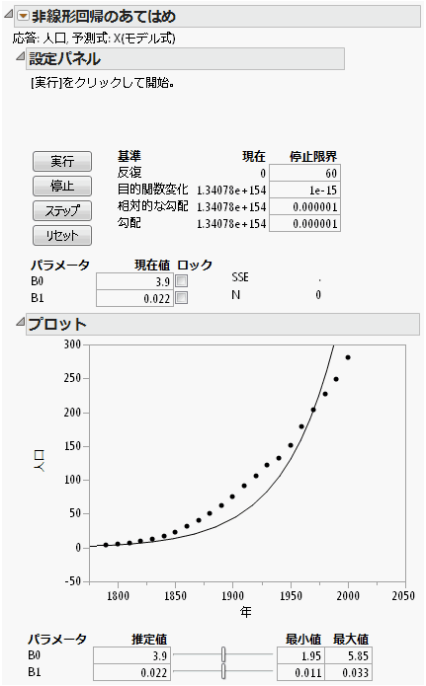
基準 反復計算が収束したとみなす基準です。

現在 各基準の現在の値を表示します。

停止限界 「基準」にリストされる指標が、この値以下になると反復計算を終了します。

プロット X変数が1つしかない場合、X変数とY変数のプロットが作成されます。このプロットには、現在のパラメータ値に基づくモデルが表示されます。現在のパラメータ値を変更するには、スライダを使用するか、またはプロットの下にあるボックスを編集します。

図13.5 最初の「非線形回帰のあてはめ」レポート



［**実行**］をクリックしてモデルをあてはめると、レポートに次の項目が追加されます（図13.6）。

推定値の保存　現在のパラメータ値を、データテーブル列の計算式に含まれているパラメータに保存します。

信頼限界　すべてのパラメータの信頼区間を計算します。これらの区間はプロファイル尤度信頼限界であり、「解」レポートに表示されます。信頼限界の計算では、各パラメータの各限界ごとに反復計算が行われます。この反復計算においては、信頼限界がうまく見つからないことがよくあります。**「α の編集」**と**「収束基準」**は信頼区間を計算するためのオプションです。**「信頼限界のための目標SSE」**の詳細については、「プロファイル尤度信頼限界」（216ページ）を参照してください。

解　パラメータ推定値とその他の統計量が表示されます。

SSE　残差平方和（誤差平方和）が表示されます。非線形回帰では、SSE を最小にするようなパラメータを求めることが目的です。損失関数を指定した場合、SSE は損失関数の合計となります。

DFE　誤差の自由度で、オブザベーションの数から、あてはめたパラメータの数を引いたものです。

MSE　誤差の平均平方が表示されます。残差誤差の分散の推定値で、SSE を DFE で割って求めます。

RMSE　残差誤差の標準偏差の推定値で、MSE の平方根です。

パラメータ　あてはめた計算式内にあるパラメータ名がリストされます。

推定値　計算されたパラメータ推定値がリストされます。非線形回帰では、処理が無事に完了した場合でも推定が完璧でない可能性があるので注意してください。

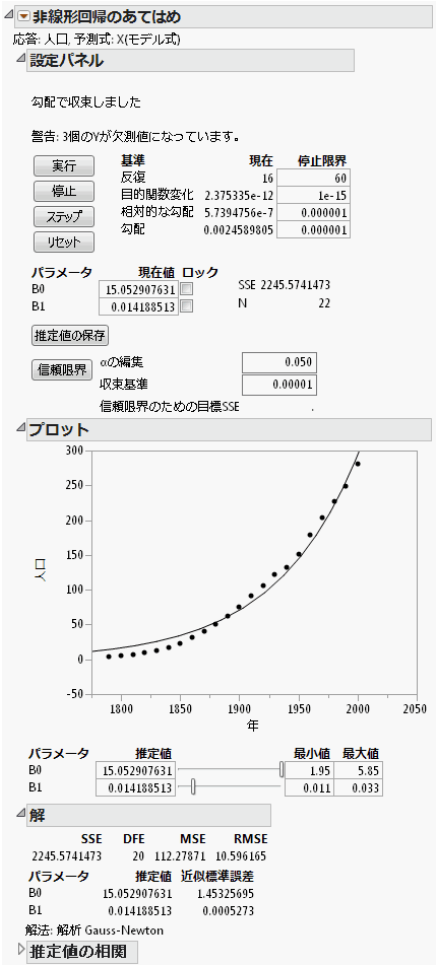
近似標準誤差　線形回帰で行われる方法に似たような枠組みで計算されます。微分したものの交差積和行列の逆行列における対角要素の平方根に、RMSE を掛けたものとして算出されます。

下側信頼限界と上側信頼限界　パラメータの信頼限界が表示されます。設定パネルで［**信頼限界**］をクリックすると表示されます。信頼区間の詳細は、「プロファイル尤度信頼限界」（216 ページ）を参照してください。

除外データ　除外された行の適合度統計量のレポートが表示されます。これは、モデルのあてはめに使用しなかったオブザベーションに関してモデルを検証する場合に便利です。この機能と［解を記録］オプションを併用すれば、除外するデータを変更し、その相違を反映させたレポートを作成することができます。

推定値の相関　パラメータ推定値間の相関を示します。

図13.6 あてはめたモデルのレポート



「非線形回帰」プラットフォームのオプション

「非線形回帰のあてはめ」レポートのタイトルバーにある赤い三角ボタンをクリックすると、次のようなオプションが表示されます。

パラメータの範囲　パラメータの範囲を設定します。このオプションを選択すると、設定パネルに編集ボックスが表示されます。制限しないパラメータについては、フィールドを空白にしておきます。

プロット　X変数が1つしかない場合、X変数とY変数のプロットが作成されます。プロットには、現在のパラメータ値に基づくモデルが表示されます。現在のパラメータ値を変更するには、スライダを使用するか、またはプロットの下にあるボックスを編集します。起動ダイアログボックスで［グループ化］変数を指定した場合、グループごとに曲線が作成されます。

反復オプション　このメニューには、反復計算のアルゴリズムに関するオプションが用意されています。

反復計算のログ　新しいウィンドウが開き、そこに反復計算の各ステップが記録されます。

数値微分のみ . モデルが複雑で解析的な微分を計算するのが困難であるような場合に効果的な機能です。解析的微分に基づく方法が収束しないようなケースに対しても役に立ちます。

中間計算式の展開 . 計算式に使われている列の中に、さらに計算式が含まれている場合に、（他の列を参照している）内側の式が代入されます。このオプションを使うときに、特定の列だけ展開されないようにするには、展開したくない列の列プロパティにおいて**「その他」**を選択し、「計算式の展開」（英語名は「Expand Formula」）という名前の列プロパティを作成し、その値を 0 とします。

Newton　最適化手法として Gauss-Newton 法（通常の最小 2 乗法の場合）または Newton-Raphson（損失関数のあるモデルの場合）が使われます。

準 Newton SR1　最適化方法として準 Newton SR1 が使われます。

準 Newton BFGS　最適化方法として準 Newton BFGS が使われます。

現在の推定値を採用　推定値が収束しなかった場合でも、現在の推定値に基づく「解」レポートが作成されます。

微分した式の表示　非線形回帰式を微分した式（導関数）が JMP ログに表示されます。微分した式の内容については、「微分した式について」（218 ページ）を参照してください。

スレッドを使用しない　反復がメインの計算スレッドで実行されます。JMP では、ほとんどの計算が個別の計算スレッドで行われます。そのため、非線形回帰の計算中に他の処理を行っても、JMP の応答性は高いままです。しかし、中にはメインスレッドで実行すべきケース（たとえば、表示ルーチンを呼び出すような副作用があるモデル）もあるので、その場合にはこのオプションをオンにします。

プロファイル　応答曲面を表示するさまざまなプロファイルが用意されています。

プロファイル　予測プロファイルが表示されます。［プロファイル］では、曲面を各 X 変数でスライスした断面が表示され、因子の最適設定を探索することができます。

等高線プロファイル　等高線プロファイルが表示されます。2 次元の等高線と 3 次元のメッシュプロットが表示されます。

曲面プロファイル 3次元の曲面プロットが作成されます。このオプションは、モデルに2つ以上のX変数がある場合のみ使用できます。

パラメータプロファイル 予測プロファイルが起動し、SSEまたは損失をパラメータの関数としてプロファイルが作成されます。

パラメータ等高線プロファイル 等高線プロファイルが起動し、SSEまたは損失をパラメータの関数として等高線プロファイルが作成されます。

パラメータ曲面プロファイル 3次元の曲面プロットが作成され、SSEまたは損失をパラメータの関数としてプロファイルが作成されます。このオプションは、モデルに2つ以上のパラメータがある場合のみ使用できます。

グリッド上のSSE 解推定値の周囲に値のグリッドを作成し、各値の誤差平方和（SSE）を計算します。解推定値のSSEは最小になるのが理想です。このオプションを選択すると、**「出力のグリッドを指定」**が表示され、次のようなオプションが使用できます。

パラメータ モデルのパラメータがリストされています。

最小値 グリッド計算で使用するパラメータの最小値を指定します。デフォルトの**「最小値」**は、解推定値から**「近似標準誤差」**の2.5倍を引いたものです。

最大値 グリッド計算で使用するパラメータの最大値を指定します。デフォルトの**「最大値」**は、解推定値に**「近似標準誤差」**の2.5倍を足したものです。

ポイントの数 各パラメータごとに、グリッドを作成するためのポイントの数を指定します。グリッドテーブル内に作成されるポイントの合計数を計算するには、すべての［ポイントの数］値を掛け合わせます。デフォルトでは、最初の2つのパラメータの［ポイントの数］が11で、残りのパラメータは3になっています。別の値を指定するときは、グリッドテーブルに解推定値が含まれるようにするため、奇数値を選んでください。［ポイントの数］が0のパラメータに対しては、推定値だけがグリッドテーブルに設定されます。

［実行］をクリックすると、指定したポイント数から構成されるグリッドが、新しいテーブルに作成されます。推定値がテーブルに含まれている場合は、その推定値の行が強調表示されます。

元のパラメータに戻す パラメータの値を、最初に設定したパラメータ値（計算式列パラメータで指定されている値）に戻します。

解を記録 現在のパラメータ推定値と要約統計量を含む「記録したモデル」というレポートが作成されます。複数のモデルの結果を記録し、比較することができます。これは、パラメータ制限の異なるモデルや、異なるオプションであてはめたモデルを比較する場合に便利です。特定のモデルのラジオボタンをクリックすると、そのモデルがプロットに表示され、パラメータ推定値が設定パネルに表示されます。

カスタム推定値 パラメータの関数に対する推定値が表示されます。パラメータだけから成る式を指定してください。その指定された式に現在のパラメータ推定値が代入されて計算されます。また、1次のTaylor展開による近似に基づいて式の標準誤差も計算されます。

カスタム逆推定 指定の Y 値から X 値を予測します。推定された X に対する標準誤差も計算されます。現在のモデル式に対する逆関数を、JMPが求められることが前提となります。標準誤差は、逆関数の1次Taylor展開による近似で求められます。また、信頼区間が、t 分位点と標準誤差を使って、Wald法により求められます。

予測信頼限界の保存 モデルに基づく予測の漸近信頼限界を保存します。これは、指定した X 値における、Y の平均に対する信頼区間です。

個別信頼限界の保存 個々の予測の漸近信頼限界を保存します。これは、指定した X 値における、個々の Y 値に対する信頼区間です。

計算式の保存 モデルの分析結果をデータテーブル列に保存するためのオプションを含みます。

予測式の保存 予測式と現在のパラメータ推定値が保存されます。

予測値の標準誤差の保存 モデルに基づく予測の標準誤差が保存されます。これは、指定した X 値における、Y の平均に対する標準誤差です。計算式は `Sqrt(VecQuadratic(行列1, ベクトル1))` という形を取ります。「行列1」はパラメータ推定値の共分散行列、「ベクトル1」はモデル式を各パラメータについて偏微分した式を要素とするベクトルです。

個々の標準誤差の保存 個々の予測の標準誤差が保存されます。これは、指定した X 値における、個々の Y 値に対する標準誤差です。計算式は `Sqrt(VecQuadratic(行列1, ベクトル1)+mse)` という形を取ります。「行列1」はパラメータ推定値の共分散行列、「ベクトル1」はモデル式を各パラメータについて偏微分した式を要素とするベクトル、「`mse`」は誤差分散の推定値です。

残差計算式の保存 残差の計算式が保存されます。

予測信頼限界の計算式の保存 モデルに基づく予測の信頼区間を計算する式が保存されます。これは、指定した X 値における、Y の平均に対する信頼区間です。

個別信頼限界の計算式の保存 個々の予測の信頼区間を計算する式が保存されます。これは、指定した X 値における、個々の Y 値に対する信頼区間です。

逆推定計算式の保存 モデルを逆推定するための計算式、逆推定の標準誤差、および個々の逆推定の標準誤差が保存されます。

解の計算式を保存 単純なケースの場合は［逆推定計算式の保存］と同じです。しかし、このコマンドを使うと、複数の変数をもつ予測式を扱うことができ、その予測式の変数に値を代入することができます。正しく機能するのは、演算子と関数が可逆で、それぞれ計算式の中に1回しか出てこない場合のみです。

このコマンドを選択してダイアログボックスが開いたら、解を求める変数を指定します。また、結果のデータテーブルの計算式で参照される列の名前を変更することもできます。列名ではなく、数値を指定することもできます。数値を指定した場合、それらの数値を代入して計算式が解かれます。

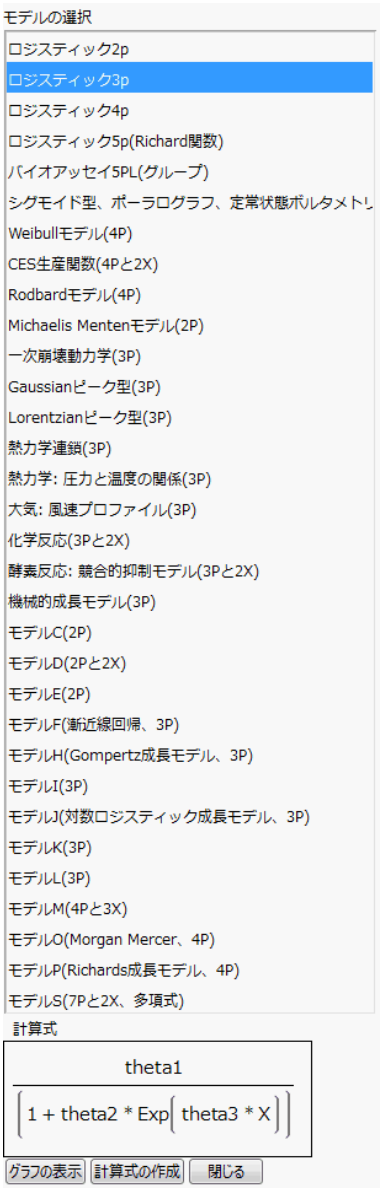
メモ：標準誤差、信頼区間、および仮説検定が正しくなるのは、最小2乗推定が行われた場合か、負の対数尤度が損失関数として指定され最尤推定が行われた場合のみです。

予測式の表示 レポートの上部に、予測モデルまたは損失関数が表示されます。

モデルライブラリを使用した列の作成

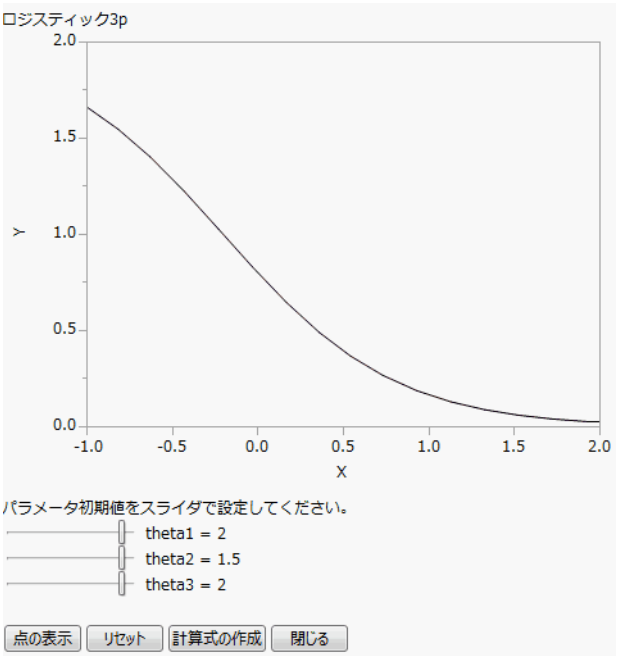
モデルライブラリは、パラメータおよびその初期値が設定された計算式を作成するのに便利です。「非線形回帰」起動ウィンドウの［**モデルライブラリ**］ボタンをクリックすると、ライブラリが開きます。リストでモデルを選択すると、その計算式が**「計算式」**ボックスに表示されます（図13.7）。

図13.7「非線形モデルライブラリ」ダイアログボックス



［**グラフの表示**］をクリックすると、モデルに説明変数が1つしかない場合は2次元上に曲線を、2つある場合は3次元上に曲面プロットを表示します。説明変数（*X*）が3つ以上あるモデルでは、グラフが作成されません。グラフウィンドウでは、スライダを使うか値を入力することによって、パラメータの初期値を変更できます。図13.8を参照してください。

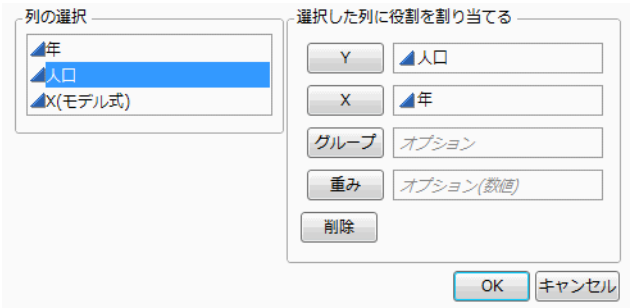
図13.8 モデルライブラリのグラフ例



［**リセット**］ボタンを押すと、パラメータの初期値がデフォルトの値に戻ります。

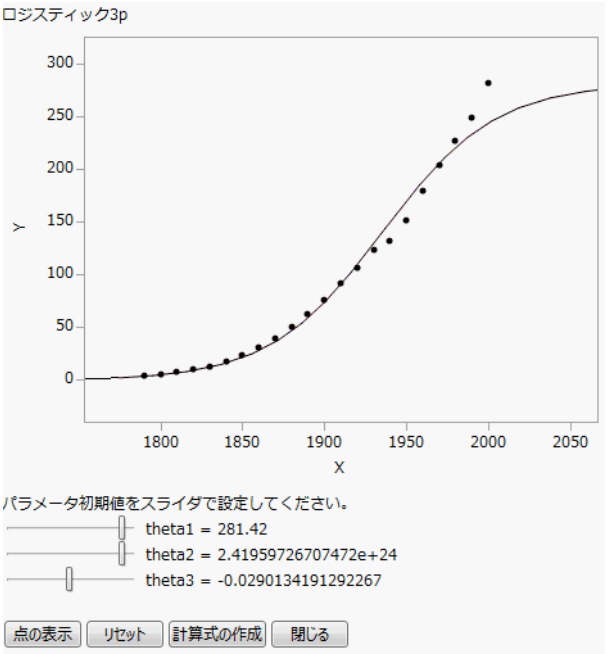
プロットに実際のデータ点を表示するには、［**点の表示**］をクリックします。図13.9のようなウィンドウが開いたら、任意の列に［*X*］と［*Y*］の役割を割り当て、必要に応じてオプションの［グループ］変数を指定します。［グループ］に列を指定すると、カテゴリカル変数の水準ごとにモデルをあてはめることができます。このウィンドウで［グループ］の列を指定した場合は、プラットフォームの起動ウィンドウでも［グループ化］の列を指定してください。

図13.9 役割の選択



用意されているほとんどのモデルにおいて、あらかじめ決められた定数が初期値に使われます。データ点を表示することにより、モデルがどれほどデータにあてはまるかを確認しながら、パラメータ値を調整することができます。「US population.jmp」の例で点を表示すると、図13.10のようになります。

図13.10 点の表示



［**点の表示**］をクリックした後で［**計算式の作成**］をクリックすると、モデルライブラリで選択したモデルに基づいた名前の新しいデータテーブル列が作成されます。列には計算式が含まれ、そのパラメータ値は最後に設定した初期値になっています。

メモ：［**グラフの表示**］ボタンまたは［**点の表示**］ボタンをクリックする前に［**計算式の作成**］をクリックした場合は、XとYの役割と、オプションとしてグループ変数を指定するよう求められます。図13.9を参照してください。変数を指定するとプロットに戻り、必要に応じてパラメータの初期値を調整できます。そこで再び［**計算式の作成**］をクリックすると、新しい列が作成されます。

データテーブル内に計算式が作成されたら、「非線形回帰」起動ダイアログボックスで計算式の列を［**X, 予測式列**］に指定し、分析を続けます。

非線形モデルライブラリのカスタマイズ

モデルライブラリは、「NonlinLib.jsl」というビルトインスクリプトによって作成されます。このスクリプトは、JMPのインストールフォルダ（Windowsの場合）またはアプリケーションパッケージ（Macintoshの場合）にある「Resources¥Builtins」フォルダに含まれています。このスクリプトを変更すると、非線形回帰モデルライブラリをカスタマイズできます。

モデルを追加するには、`Listofmodellist#`というリストに3つの行を加える必要があります。この3つの行は次のような情報であり、これらをリスト内に指定します。

- モデル名（引用符で囲んだ文字列）
- モデルの計算式（式）
- モデルのスケール（範囲）

たとえば、次の式で表される「単純な指数成長モデル」というモデルを追加するとしましょう。

$$y = b_1 e^{kx}$$

この場合、「NonlinLib.jsl」スクリプトのListofmodellist#リストに以下の行を挿入します。

```
{// 単純な指数成長モデル
      " 単純な指数成長モデル ",
      Expr(Parameter({b1=2, k=0.5}, b1*exp(k * :X))),
      lowx = -1; highx = 2; lowy = 0;  highy = 2},
```

以下の点に注意してください。

- 第1行は、リストの開始を意味する開く括弧とコメント（オプション）から成ります。第2行に入力した文字列がモデル名としてモデルライブラリウィンドウに表示されます。
- `lowx`、`highx`、`logy`、および`highy`は、表示されるグラフの範囲を表します。
- 上記の例では最後にカンマがついていますが、`Listofmodellist#`リストの最後のエントリになる場合には、カンマは必要ありません。
- 説明変数が3つ以上あるモデルでは、最後の行（グラフの範囲を指定する行）に"Not Available"という文字列を引用符で囲んで入力します。

  モデルライブラリにあるモデルを削除するには、`Listofmodellist#`リストから該当する行を削除します。

「非線形回帰」プラットフォームの別例

この節では、例を挙げて「非線形回帰」プラットフォームの幅広い用途をご紹介します。

最尤法の例: ロジスティック回帰

この例では、「非線形回帰」プラットフォームを使って損失関数を最小化する方法について説明します。損失関数に負の対数尤度関数を指定すると、最尤法による推定値が求められます。

サンプルデータの「Nonlinear Examples」フォルダにある「Logistic w Loss.jmp」データテーブルは、損失関数を使ってロジスティック回帰をあてはめる例です。**「Y」**列の値は、イベントが発生した場合が「1」、発生しなかった場合が「0」です。**「モデル Y」**列には線形モデル、**「損失」**列には損失関数が含まれます。この例の損失関数は、各オブザベーションの負の対数尤度、つまり応答の観測値についてその観測値が得られる確率を計算し、その対数の符号を逆にしたものです。

以下の手順に従って、モデルを実行します。

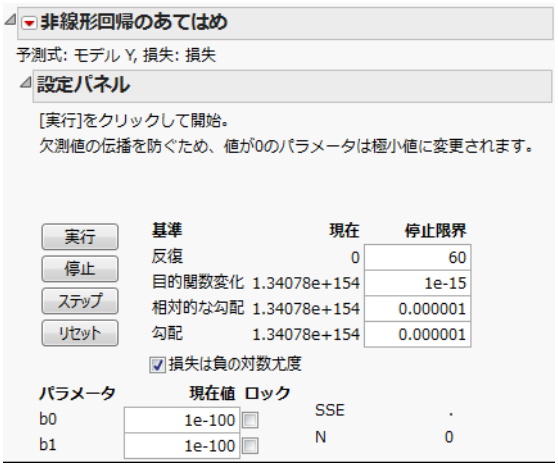
1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Nonlinear Examples¥Logistic w Loss.jmp」を開きます。
2. ［**分析**］>［**発展的なモデル**］>［**非線形回帰**］を選択します。
3. **「モデル Y」**を［**X, 予測式列**］に指定します。
4. **「損失」**を［**損失**］に指定します。

図13.11「非線形回帰」起動ウィンドウ

5. ［OK］をクリックします。

「非線形回帰のあてはめ」設定パネルが表示されます。

図13.12「非線形回帰のあてはめ」設定パネル



6. ［実行］をクリックします。

パラメータ推定値が「解」レポートに表示されます。図13.13を参照してください。

図13.13「解」レポート

「解」レポートの「損失」の値は、パラメータ推定値において計算された負の対数尤度です。

2項分布に対するプロビットモデルの例

「Ingots2.jmp」サンプルデータテーブルは、製造過程で加熱時間とソーキング時間を変えながら、圧延に耐える状態に仕上がったインゴットの数を記録したものです。応答変数の**「数(適合)」**は2項分布に従う変数であり、検査されたインゴットの数（**「数(全体)」**）、加熱時間、およびソーキング時間によって変化しています。2項分布のプロビットモデルにおいて、最尤法でパラメータ推定値を得るには、次のものを使用します。

- 数値微分
- 損失として、負の対数尤度
- Newton-Raphson法

圧延に耐えるインゴットの平均数は、与えられた加熱時間とソーキング時間においてインゴットが良い状態に仕上がる確率に、検査された数を掛けたものです。プロビットモデルをあてはめるために、サンプルデータの**「モデル変数 (P)」**列には次の計算式が設定されています。

```
Normal Distribution(b0+b1*Heat+b2*Soak)
```

Normal Distribution関数の引数は、2変数の線形モデルになっています。

損失関数として、次式で表わされるような、2項分布に基づき計算した負の対数尤度を定義しています。

```
-( 適合数 *Log(p)+( 全体数 - 適合数 )*Log(1-p))
```

以下の手順に従ってモデルをあてはめます。

1. ［**分析**］>［**発展的なモデル**］>［**非線形回帰**］を選択します。
2. **「モデル変数 (P)」**を［**X, 予測式列**］に指定します。
3. **「損失関数」**を［**損失**］に指定します。
4. ［**数値微分のみ**］チェックボックスをオンにします。
5. ［**OK**］をクリックします。
6. ［**実行**］をクリックします。

ここでは、数値微分を反復計算で使用しました（図13.14）。

図13.14「Ingots2」データの解

解

損失	DFE	平均損失	平均損失平方根
47.479945327	16	2.9674966	1.7226423

パラメータ	推定値	近似標準誤差
b0	-2.8934153	0.51256572
b1	0.0399554554	0.01202329
b2	0.0362537934	0.15017139

解法: 数値 SR1

Poisson損失関数の例

Poisson分布は、度数データのモデル化によく使われます。

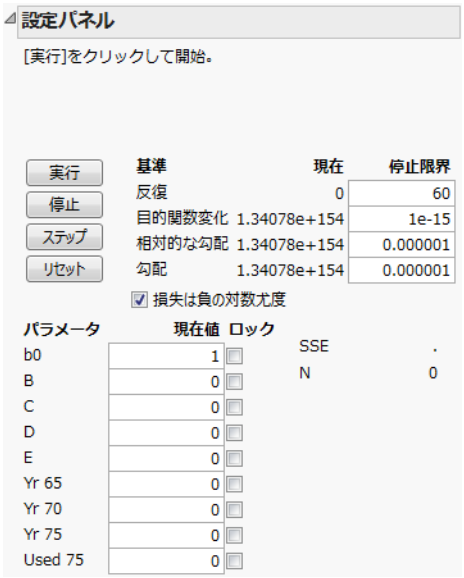
$$P(Y = n) = \frac{e^{-\mu}\mu^{n}}{n!}, n = 0, 1, 2, \ldots$$

μは1つのパラメータか、または多数のパラメータを持つ線形モデルです。このモデルを変換して、反復重み付き最小2乗法であてはめる方法は、たくさんの本や論文で取り上げられています（Nelder and Wedderburn 1972）。JMPでは、もっと単純にモデルを直接あてはめることができます。ここでは、例として、McCullagh and Nelder（1989）が取り上げている、波が原因で生じた貨物輸送船事故の数のデータを用います。

このデータは「Ship Damage.jmp」データテーブルにまとめられています。モデル計算式は**「モデル」**列、損失関数（負の対数尤度）は**「Poisson 損失関数」**列にあります。モデルをあてはめるには、次の手順に従います。

1. ［**分析**］>［**発展的なモデル**］>［**非線形回帰**］を選択します。
2. **「モデル」**を［**X, 予測式列**］に指定します。
3. **「Poisson 損失関数」**を［**損失**］に指定します。
4. ［**OK**］をクリックします。
5. b0 の**「現在値」**（初期値）を「1」に、その他のパラメータを「0」に設定します（図 13.15）。

図13.15 新しいパラメータの入力



6. ［**実行**］をクリックします。
7. ［**信頼限界**］ボタンをクリックします。

図 13.16 のような「解」レポートが表示されます。結果には、パラメータ推定値、信頼区間、要約統計量が含まれています。

図13.16 Poisson損失の「解」表の例

解

損失	DFE	平均損失	平均損失平方根
68.281234087	25	2.7312494	1.6526492

パラメータ	推定値	近似標準誤差	下側信頼限界	上側信頼限界
b0	-6.405914771	0.21744445	-6.8430512	-5.9896833
B	-0.543353982	0.17758996	-0.881379	-0.183552
C	-0.687418388	0.32904722	-1.3764541	-0.0745299
D	-0.075979835	0.2905789	-0.6715296	0.47523936
E	0.325581683	0.23587934	-0.143451	0.78520431
Yr 65	0.6971496668	0.1496412	0.40752821	0.99512758
Yr 70	0.8184263427	0.16977314	0.48728046	1.15369087
Yr 75	0.4534435316	0.23317069	-0.0123098	0.90388895
Used 75	0.3844880718	0.11827198	0.15341509	0.61742312

解法: 解析 NR

メモ: 標準誤差、信頼区間、および仮説検定が正しくなるのは、最小2乗推定が行われた場合か、負の対数尤度が損失関数として指定され最尤推定が行われた場合のみです。

パラメータの範囲を設定する例

まず、「曲線のあてはめ」機能を使ってモデルをあてはめ、そこで予測式をデータテーブルに保存した後、「非線形回帰」プラットフォームでその予測式を利用することもできます。この方法では、操作が多くなり、ユーザによる設定が必要となりますが、どんな非線形回帰モデルでも柔軟にあてはめることができます。

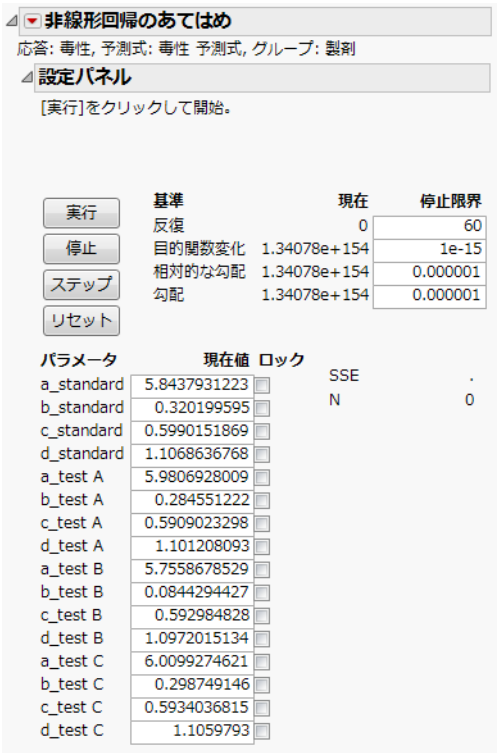
次の例を行うにあたり、まず、「曲線のあてはめ」章の「「曲線のあてはめ」機能の使用例」（174 ページ）の手順に従い、モデルをあてはめてください。そして、次の手順で、「曲線のあてはめ」で予測式を保存し、パラメータの範囲を「非線形回帰」プラットフォームで設定することができます。

1. 「ロジスティック 4P」の赤い三角ボタンのメニューから［**予測式の保存**］>［**パラメトリックな予測式の保存**］を選択します。

   データテーブルに**「毒性 予測式」**という新しい列が表示されます。
2. ［**分析**］>［**発展的なモデル**］>［**非線形回帰**］を選択します。
3. **「毒性」**を［**Y, 応答変数**］に指定します。
4. **「毒性 予測式」**を［**X, 予測式列**］に指定します。
5. **「製剤」**を［**グループ化**］に指定します。
6. ［**OK**］をクリックします。

   「非線形回帰のあてはめ」ウィンドウが表示されます（図13.17）。設定パネルに、パラメータ値と固定オプションが表示されます。各パラメータの前に表示されているアルファベットは、「曲線のあてはめ」プラットフォームの「予測モデル」でのモデル式で使われている変数に対応しています。

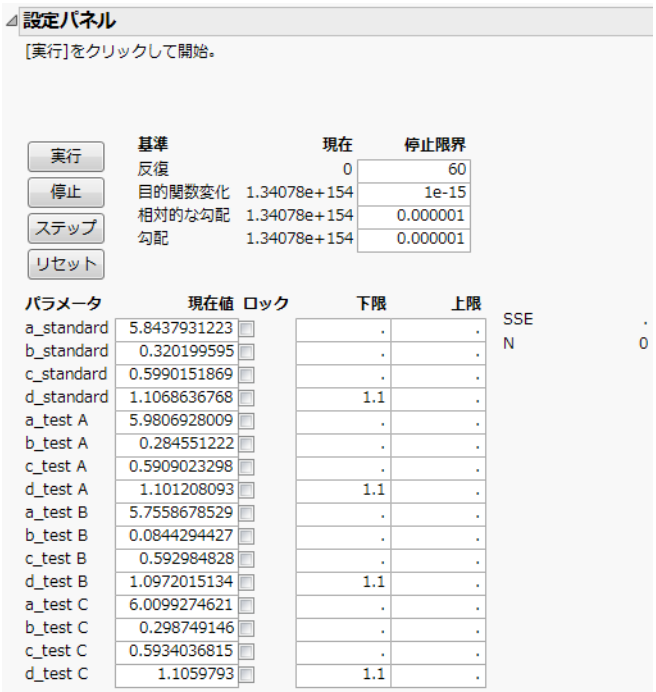
図13.17「非線形回帰のあてはめ」設定パネル



ヒント： 手持ちの情報からパラメータの値がわかっている場合は、パラメータを固定することができます。

7. **「非線形回帰のあてはめ」**の赤い三角ボタンのメニューを開き、**［パラメータの範囲］**を選択します。
   パラメータの隣に上限と下限を設定するためのオプションが表示されます。
8. 図13.18のようにパラメータの下限を設定します。ここでは、過去の経験から、薬剤の毒性の最大値が1.1以上であることがわかっているものとします。

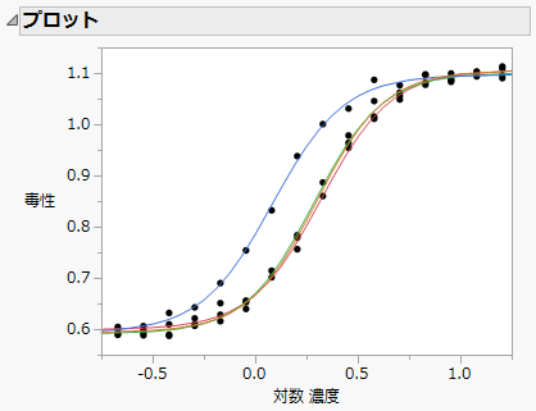
図13.18 パラメータの範囲の設定



9. ［**実行**］をクリックします。

「**解**」レポートには、最終的なパラメータ推定値とその他の適合度統計量が表示されます（図13.19）。プロットには、あてはめたモデルが表示されます。

図13.19 非線形回帰のあてはめプロットとパラメータ推定値



パラメータ	推定値		最小値	最大値
a_standard	5.8437931223		2.9219	8.76569
b_standard	0.320199595		0.1601	0.4803
c_standard	0.5990151869		0.29951	0.89852
d_standard	1.1068636768		0.55343	1.6603
a_test A	5.9806928009		2.99035	8.97104
b_test A	0.284551222		0.14228	0.42683
c_test A	0.5909023298		0.29545	0.88635
d_test A	1.101208093		0.5506	1.65181
a_test B	5.7558678529		2.87793	8.6338
b_test B	0.0844294427		0.04221	0.12664
c_test B	0.592984828		0.29649	0.88948
d_test B	1.0972015134		0.5486	1.6458
a_test C	6.0099274621		3.00496	9.01489
b_test C	0.298749146		0.14937	0.44812
c_test C	0.5934036815		0.2967	0.89011
d_test C	1.1059793		0.55299	1.65897

プロットの下に、パラメータの限界と推定値を調整するオプションが表示されます（図13.19）。

統計的詳細

この節では、「非線形回帰」プラットフォームの統計的詳細とその他の特記事項を紹介します。

プロファイル尤度信頼限界

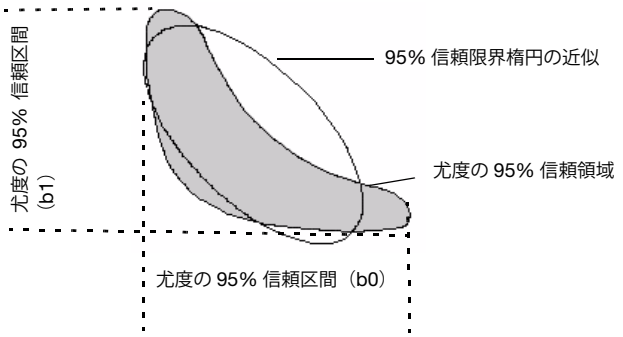
「非線形回帰」プラットフォームで計算される、パラメータの上側信頼限界は、解の最小SSEよりも特定の分だけ大きいSSEをもつという制約の中で、その他のパラメータを自由に動かしたときに、当該のパラメータ値が最大となる点です。下側信頼限界は、同様の条件で、パラメータ値が最小となる点です。これら2つのSSEの差が、*F*分布（もしくはカイ2乗分布）の値になるように信頼限界を求めます。このようにして求められた信頼区間は、**尤度信頼区間**または**プロファイル尤度信頼区間**と呼ばれます（Bates and Watts 1988, Ratkowsky 1990）。

プロファイル信頼限界は、**目標SSE**（goal SSE）から決定されます。目標SSEとは、F検定の結果、与えられたα水準において、解のSSEと有意に異なると判断される誤差平方和（または損失関数和）です。損失関数を負の対数尤度として指定したときは、F分布ではなくカイ2乗分布の分位点が使用されます。各パラメータの上側信頼限界を見つける際、パラメータ値はSSEが目標SSEに達するまで引き上げられます。ただし、該当するパラメータが目標SSEになるように変更されるに合わせて、他のパラメータは最小2乗推定値になるように調整されます。プロファイル尤度の定義から考えれば、すべてのパラメータごとに、反復計算による最適化を複数回行えば求めることができます。ただし、JMPの内部計算では、すべてのパラメータごとに、Johnston and DeLongによって開発された1セットだけの反復計算法が使用されています。『SAS/STAT 9.1 vol. 3』pp. 1666-1667を参照してください。

図13.20にある輪郭は目標SSE（または負の対数尤度）を表し、塗りつぶされた領域はSSE（または負の対数尤度）がそれ以下になる領域を表します。

- 漸近標準誤差によって近似される領域は楕円です。また、この近似による信頼区間は、楕円の極値（水平接線と垂直接線での値）になります。
- プロファイル信頼限界の場合、近似による楕円ではなく、真の領域の極値におけるパラメータ値が求められます。

図13.20 パラメータの信頼限界の図



プロファイル尤度の信頼区間は、近似標準誤差から計算された信頼区間より信用できます。プロファイル尤度の信頼限界が見つからなかった場合、その信頼限界の計算は終了し、次の信頼限界の計算へと移ります。なかなか収束しないときは、次のような作業を行ってください。

- αを大きくする。区間が短くなり、計算できる可能性が高くなります。
- 信頼限界の収束基準を緩める。

損失関数が定義されたときの仕組み

「非線形回帰」プラットフォームでは、デフォルトの残差平方和（誤差平方和）以外の関数を最小化または最大化することができます。この節では、その処理について数学的に説明します。

$f(\beta)$というモデルを使うとしましょう。「非線形回帰」プラットフォームでは、次の式で定義される損失関数の合計が最小化されます。

$$L = \sum_{i=1}^{n} \rho(f(\beta))$$

各行の損失関数ρ(•)は、データテーブル内にある他の変数の関数である場合があります。また、損失関数を1次微分および2次微分した値はゼロでない必要があります。デフォルトの損失関数ρ(•)は残差の2乗で、次のような式で表されます。

$$\rho(f(\beta)) = (y - f(\beta))^2$$

モデルに自分自身で定義した損失関数を指定するには、データテーブル内で変数を作成し、損失関数を構成します。そして、「非線形回帰」プラットフォームを起動し、損失関数を含んだ列を［損失］に指定します。

非線形回帰における最小化の計算では、モデル内にあるρ(•)の最初の2つの微分を取り、次のように勾配と近似ヘッセ行列が求められます。

$$L = \sum_{i=1}^{n} \rho(f(\beta))$$

$$\frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \rho(f(\beta))}{\partial f} \frac{\partial f}{\partial \beta_j}$$

$$\frac{\partial^2 L}{\partial \beta_j \partial \beta_k} = \sum_{i=1}^{n} \left[\frac{\partial^2 \rho(f(\beta))}{(\partial f)^2} \frac{\partial f}{\partial \beta_j} \frac{\partial f}{\partial \beta_k} + \frac{\partial \rho(f(\beta))}{\partial f} \frac{\partial^2 f}{\partial \beta_k \partial \beta_j} \right]$$

パラメータに関して$f(\bullet)$が線形なら、最後の式の第2項はゼロになります。ゼロにならなくても、合計が第1項に比べて比較的小さくなるようであれば、次の式を使っても良いでしょう。

$$\frac{\partial^2 L}{\partial \beta_j \partial \beta_k} \cong \sum_{i=1}^{n} \frac{\partial^2 \rho(f(\beta))}{(\partial f)^2} \frac{\partial f}{\partial \beta_j} \frac{\partial f}{\partial \beta_k}$$

ρが残差の2乗のときは、残差の合計が小さいため、第2項は小さくなるはずです。また、切片項があるときは0になります。最小2乗法の場合は、この項がGauss-Newton法とNewton-Raphson法とで異なります。

メモ: 標準誤差、信頼区間、および仮説検定が正しくなるのは、最小2乗推定が行われた場合か、負の対数尤度が損失関数として指定され最尤推定が行われた場合のみです。

微分した式について

「非線形回帰」プラットフォームでは、一般的な演算を含む計算式に、解析的な微分が使用されます。この節では、どのような微分式が求められるかを紹介します。

「Nonlinear Example」フォルダにある「Negative Exponential.jmp」サンプルデータでは、**「非線形」**列の実際の計算式は次のようになっています。

```
Parameter({b0=0.5, b1=0.5},b0*(1-Exp(-b1*X)))
```

計算式エディタではParameter演算の部分が表示されません。しかし、列にはこの形で保存され、「非線形回帰」起動ダイアログボックスにもこの形で表示されます。「**b0**」と「**b1**」という2つのパラメータには初期値が与えられていて、反復計算の初期値として使われます。

「非線形回帰」プラットフォームでは、プラットフォーム用に計算式のコピーが作成され、それが編集されてパラメータが抽出されます。また、パラメータへの参照が、パラメータの推定場所にマップされます。「非線形回帰」プラットフォームは、予測式をパラメータについて微分した式を内部的に求めます。[**微分した式の表示**]コマンドを使うと、その微分された式がログに表示されます。

予測モデル:

```
b0 * First(T#1=1-(T#2=Exp(-b1*X)), T#3=-(-1*T#2*X))
```

この予測モデルを、パラメータについて微分した式は、次のとおりです。

```
{T#1, T#3*b0}
```

微分は、次のように行われます。

- 補助式を繰り返し計算する手間を省くため、予測モデルはスレッドに分割され、微分計算に必要な補助式の値を保存する割り当てが与えられます。割り当てには`T#1`、`T#2`、という名前がつきます。
- 予測モデルで追加の補助式を計算する必要が生じると、最初の引数式の値を戻す「`First`」関数が使用され、その他の引数も計算されます。この場合、微分のために追加の割り当てが必要になります。
- 微分表自体は、あてはめられるパラメータごとに式をまとめたリストです。先ほどの例では、**b0** についてのモデルの微分は`T#1`です。予測モデル内でのそのスレッドは`1-(Exp(-b1*X))`です。**b1**について微分した式は`T#3*b0`で、先ほど示した割り当てに代入すると`-(1*Exp(-b1*X)*X)*b0`となります。いろいろな最適化処理が行われるものの、演算の組み合わせは常に最適とは限りません。たとえば、`T#3`の式では、二重の負の符号が使われています。

損失関数を指定すると、計算式エディタはパラメータについての微分を取ります。モデルがある場合は、モデルについて1次微分および2次微分した式を取ります。

関数の解析的な微分ができないときは、`NumDeriv`関数を使って数値的な微分が行われます。その場合はプラットフォームに、関数の変化量を計算するために使用するδが表示されます。適切な数値微分を得られるように、いろいろな値のδを試してみる必要があるかもしれません。

ヒント

モデルを式で表す方法はたくさんありますが、それらの式の効率は大きく違います。Ratkowsky（1990）は、その著作の中でいろいろな式を比較しています。

計算式に繰り返し出てくる補助式がある場合は、一時変数へ割り当てるといいでしょう。その後、計算式内でそれを参照します。たとえば、すでに取り上げたモデル計算式の1つに次の式があります。

```
If(Y==0, Log(1/(1+Exp（モデル）)), Log(1-1/(1+Exp（モデル）)));
```

式の一部を一時的なローカル変数に割り当てると、次のような単純な式になります。

```
temp=1/(1+Exp（モデル）);
If(Y==0, Log((temp), Log(1-(temp));
```

微分機能は、割り当てや条件式があっても問題なく機能します。

効果的な非線形回帰モデルに関するメモ

多項式は、**中心化**することを強くお勧めします。

パラメータに関して線形であるような多項式の項がある場合は、必ず中心化するようにしてください。これにより、最適化における数値的な状態が改善されます。たとえば

$$a_1 + b_1 x + c_1 x^2$$

という式は、次のように変換します。

$$a_2 + b_2(x - \bar{x}) + c_2(x - \bar{x})^2$$

2つのモデルは、パラメータを変換した点を除いて等価ですが、2つ目のモデルは、モデルが非線形な場合、ずっと簡単にあてはめることができます。

次に示すように、パラメータの変換式は簡単です。

$$\begin{aligned} a_1 &= a_2 - b_2\bar{x} + c_2\bar{x} \\ b_1 &= b_2 - 2c_2\bar{x} \\ c_1 &= c_2 \end{aligned}$$

それでも反復回数が最大数に達してしまう場合は、最大反復回数を増やすか、収束基準のいずれかを緩めます。

すべての問題に対応できる万能な最適化手法というのは存在しません。JMPの「非線形回帰」プラットフォームには、[**Newton**]、[**準 Newton BFGS**]、[**準 Newton SR1**]、[**数値微分のみ**] といったオプションが用意されているため、広範な問題を解決することができます。

デフォルト設定では解が収束しなかった場合でも、他のいろいろな設定を試してみることにより、収束の可能性は高まります。

モデルの中には、パラメータの初期値に対して非常に敏感なものがあります。そのため、初期値を変えてみると効果があるかも知れません。初期値を編集し、[**リセット**] をクリックして結果を見てみましょう。プロットが役に立つこともよくあります。スライダを動かして曲線のあてはまりを改善してみましょう。パラメータプロファイルが役に立つこともありますが、データセットが大きい場合には計算に時間がかかるかもしれません。

第14章

Gauss過程

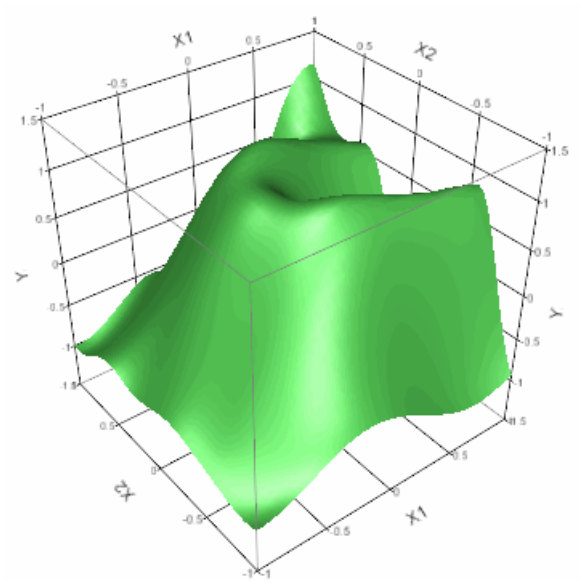
空間的モデルによるデータの補間や平滑化

「Gauss過程」プラットフォームは、複数の独立変数と1つの応答変数との間の関係をモデル化します。独立変数と応答変数は共に連続尺度である必要があります。Gauss過程モデル（Gaussian process model）は、有限要素法のようなコンピュータによるシミュレーション実験などの分野で、データを完璧に補間するモデルとして広く利用されています。Gauss過程モデルは、確率的な誤差を持たないモデル、つまり入力変数の値が同じであれば必ず出力変数の値も同じであるようなデータを扱います。

「Gauss過程」プラットフォームは、データに対して空間相関モデル（spatial correlation model）をあてはめます。このモデルにおいては、2つのオブザベーション間において、独立変数の値から計算される距離が長いほど、応答変数の相関が弱くなります。

このプラットフォームの目的の1つは予測式を求め、さらなる分析や最適化に役立てることです。

図14.1 Gauss過程の予測曲面の例

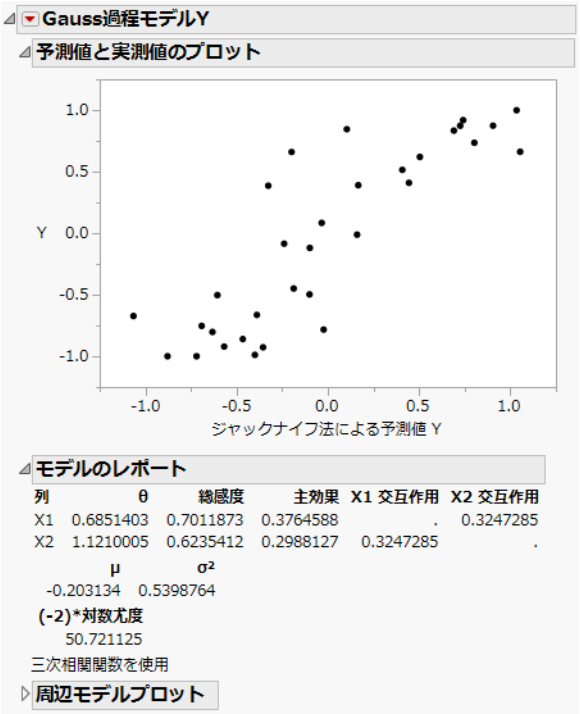


Gauss過程の例

この例では、2つの説明変数からなるモデル式によって、1つの応答変数が誤差無く決定されているデータを使用します。なお、データ点の計画には、Space Filing計画が使われています。「Gauss過程」プラットフォームでは、**X1**と**X2**が**Y**をどの程度説明できるかを調べることができます。Yの式は、列の計算式で確認できます。

1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「2D Gaussian Process Example.jmp」を開きます。
2. ［**分析**］>［**発展的なモデル**］>［**Gauss過程**］を選択します。
3. 「**X1**」と「**X2**」を選択し、［**X**］をクリックします。
4. 「**Y**」を選択し、［**Y**］をクリックします。
5. 「**相関構造**」として［**三次**］を指定します。
6. JMP PRO［**高速Gauss過程**］オプションをオフにします。
7. ［**OK**］をクリックします。

図14.2「Gauss過程」レポート

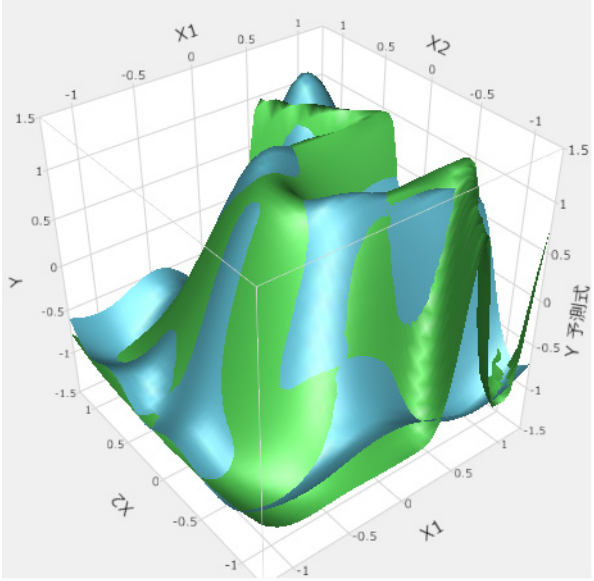


メモ: 最小化を行うルーチンで異なる開始点が使われるために、結果が異なる場合があります。また、使用した相関構造、および、ナゲットパラメータを含めたかどうかによっても結果は異なります。

あてはめられた曲面と元の曲面を視覚的に比較してみましょう。

8. 「Gauss過程モデルY」の赤い三角ボタンをクリックし、[**予測式の保存**]を選択します。
9. [**グラフ**] > [**曲面プロット**]を選択します。
10. 「**X1**」から「**Y 予測式**」までを選択し、[**列**]をクリックします。
11. [**OK**]をクリックします。
12. 「曲面」列で「Y 予測式」に対し[**両面**]を選択します。

図14.3 Yの予測値と実測値の3D曲面プロット

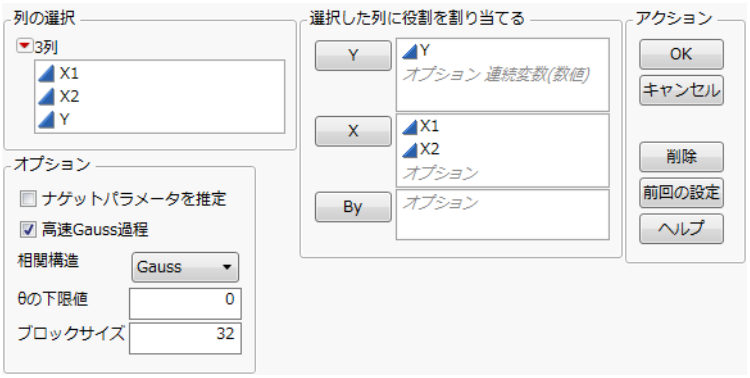


これらの2つの曲面はよく似ています。X1とX2が応答Yに与える影響は目で確認できます。プロットを回転させて、さまざまな角度から見てみましょう。周辺モデルプロットも、因子が応答に与える影響の理解に役立ちます。

「Gauss過程」プラットフォームの起動

「Gauss過程」プラットフォームを起動するには、[**分析**] > [**発展的なモデル**] > [**Gauss過程**]を選択します。

図14.4「Gauss過程」の起動ウィンドウ



Y 分析対象とする連続尺度の列を指定します。

X 説明変数として使う列を指定します。JMP Proでは、［高速Gauss過程］オプションをオンにすれば、カテゴリカルな変数を使用できます。

ナゲットパラメータを推定 モデルにナゲットパラメータ（リッジパラメータ）を追加します。ナゲットパラメータは、応答変数のデータに誤差が含まれている場合に、誤差を考慮した予測モデルをあてはめるのに役立ちます。ナゲットパラメータを含めた場合は、データ点のすべてを通る完璧な補間ではなく、誤差を考慮した平滑化が行われます。

JMP PRO **高速Gauss過程** 高速Gauss過程アルゴリズムを使用するオプション。高速Gauss過程は、Gauss過程を小さなブロックに分割し、計算時間の短縮を図ります。ブロックに分割することで、複数のCPUの使用と並列計算が可能になります。

メモ：オブザベーションが2,500以上ある場合は、高速Gaussアルゴリズムが必要です。

高速Gauss過程の詳細については、Parker（2015）を参照してください。

相関構造 モデルの相関構造を選択します。「Gauss過程」プラットフォームは、データに対して空間相関モデル（spatial correlation model）をあてはめます。このモデルにおいては、2つのオブザベーション間において、独立変数の値から計算される距離が長ければ長いほど、応答変数の相関が弱くなることが仮定されています。

［**Gauss**］を選択すると、Gauss相関関数が使われます。この関数は、データ点間の距離が離れていても、その相関は完全にはゼロになりません。

［**三次**］を選択すると、2つの点の距離が一定以上離れている場合、相関がゼロになります。この手法は、3次スプラインによる補間を一般化したものです。

JMP PRO 高速Gauss過程アルゴリズムは3次の相関構造をサポートしません。

θの下限値 あてはめたモデルで使用するθの最小値を設定します。デフォルトの値は0です。θの値は、通常の回帰モデルにおける傾きパラメータに似ています。θの値が小さいときは、変数が予測値に与える影響が小さいことを示します。

JMP PRO **ブロックサイズ** 高速Gauss過程アルゴリズムで使用される計算ブロック1つあたりのオブザベーション数。ブロックあたりのオブザベーション数は25以上、データセットの行数は最大2,500でなければなりません。

「Gauss過程モデル」レポート

最初の「Gauss過程」レポートには、予測値と実測値のプロットとモデルのレポートが表示されます。各因子の周辺モデルプロットは最初、非表示になっています。

予測値と実測値のプロット

「予測値と実測値のプロット」は、Y軸に実際のY値、X軸にジャックナイフ法による予測値を示したものです。適合度の指標の1つとして、点がプロットの対角線（Y = X）にどの程度沿っているかが挙げられます。

このジャックナイフ法は、1行ずつ除外したデータに対する再あてはめを行ってはいません。その意味で、厳密なジャックナイフ法ではありません。各行は、その行のYを予測するモデルからは除外されますが、推定された相関パラメータには、その行による寄与が含まれます。データを完璧に補間するGauss過程の場合、このジャックナイフ法による予測値は観測された応答値に等しくなりません。

モデルのレポート

「モデルのレポート」は一種の分散分析表です。モデルの推定結果が表示されます。この分散分析表では、変動が関数に基づく手法で計算されます。

θ Gauss過程のモデルパラメータの推定値。「「Gauss過程」プラットフォームの統計的詳細」（230ページ）を参照してください。

総感度 各因子の、主効果とすべての交互作用効果の和。これは、因子とその2次交互作用が応答変数に与える影響の度合いを表します。

実験空間全体について予測式の変動を積分することにより、モデル全体の変動が計算されます。

主効果 各因子の関数的な主効果は、周辺予測値の変動をその因子だけで積分して求めます。主効果は、モデル内の各因子の関数的な主効果と全体の変動の比です。

交互作用 関数的な交互作用効果は、主効果と同じように算出されます。

JMP PRO **カテゴリカル変数** モデルにカテゴリカルな因子が含まれる場合は、カテゴリカルな因子ごとに相関行列が作成されます。非対角成分は、Gauss過程のモデルパラメータの推定値に該当します。「カテゴリカルな説明変数を使ったモデル」（231ページ）を参照してください。

μ と σ^2 平均と分散のモデルパラメータ。

ナゲット ナゲットの推定値。ナゲットの値は、「Gauss過程」起動ウィンドウで［**ナゲットパラメータを推定**］オプションを選択した場合に計算されます。また、共分散行列が特異になるのを防ぐ目的でJMPがナゲットパラメータを追加するときもあります。

(-2)*対数尤度 最大化された対数尤度を(-2)倍したもの。

周辺モデルプロット

モデルの各因子に対して周辺モデルプロットが作成されます。このプロットは、他のすべての因子を平均値に設定したときの、ある因子の各水準の応答を示します。

「Gauss過程」プラットフォームのオプション

「Gauss過程」の赤い三角ボタンのメニューには、レポートを自分のニーズに合わせてカスタマイズするためのオプションが用意されています。

プロファイル 標準的なプロファイルを開きます。詳細については、『プロファイル』の「プロファイル」章を参照してください。

等高線プロファイル 等高線プロファイルを開きます。詳細については、『プロファイル』の「等高線プロファイル」章を参照してください。

曲面プロファイル 曲面プロファイルを開きます。詳細については、『プロファイル』の「曲面プロット」章を参照してください。

予測式の保存 アクティブなデータテーブルに新しい予測式の列を作成します。

分散計算式の保存 アクティブなデータテーブルに新しい分散計算式の列を作成します。

JMP PRO **予測式を発行** 予測式を作成し、それを「計算式デポ」プラットフォームの計算式列のスクリプトとして保存します。「計算式デポ」レポートが開いていない場合は、このオプションによって「計算式デポ」レポートが作成されます。「計算式デポ」章（167ページ）を参照してください。

JMP PRO **分散計算式を発行** 分散計算式を作成し、それを「計算式デポ」プラットフォームの計算式列のスクリプトとして保存します。「計算式デポ」レポートが開いていない場合は、このオプションによって「計算式デポ」レポートが作成されます。「計算式デポ」章（167ページ）を参照してください。

ジャックナイフ法による予測値を保存 データテーブル内にジャックナイフ法による予測値を保存します。これらの値は、予測値と実測値のプロットにおけるX軸の値です。

以下のオプションについて詳しくは、『JMPの使用法』の「JMPのレポート」章を参照してください。

ローカルデータフィルタ 現在のレポートにおいてデータをフィルタリングするローカルデータフィルタを表示するか、非表示にします。

やり直し 分析を繰り返すオプションや、やり直すオプションを含みます。また、［自動再計算］オプションを選択すると、このオプションに対応しているプラットフォームにおいて、データテーブルに加えた変更が、該当するレポートに即座に反映されるようになります。

スクリプトの保存 レポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。

Byグループのスクリプトを保存 By変数の全水準に対するレポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。起動ウィンドウでBy変数を指定した場合のみ使用可能です。

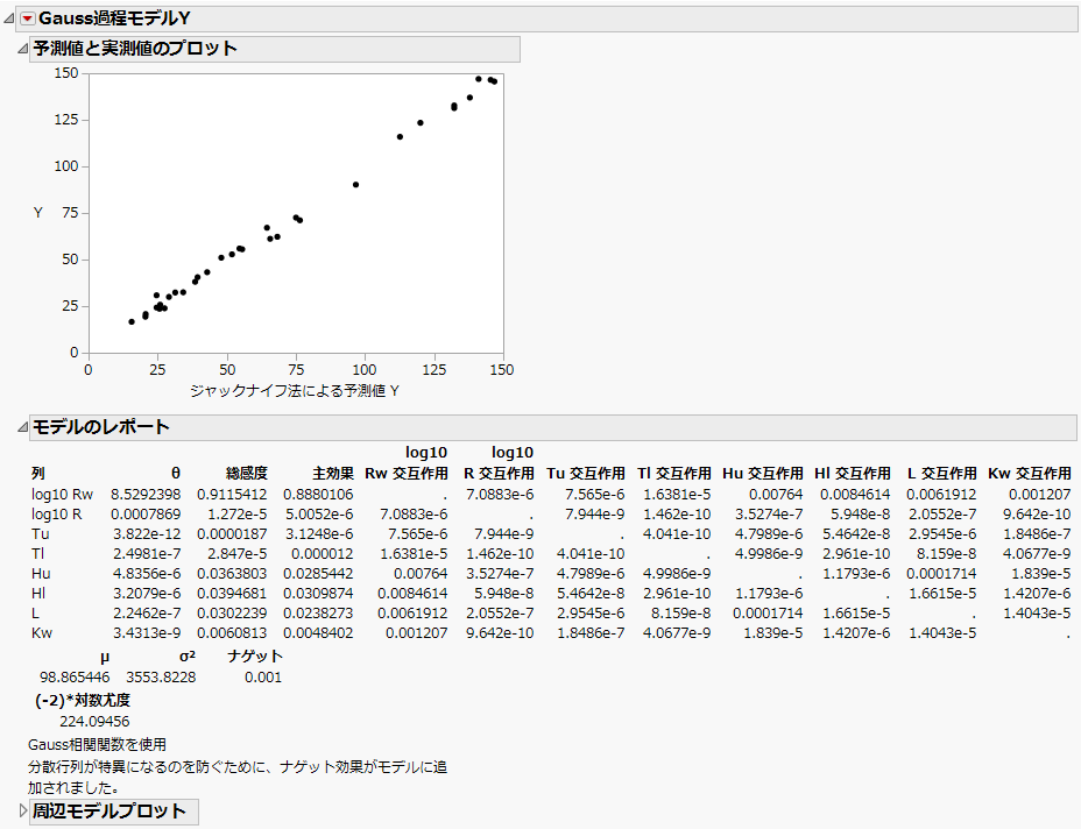
「Gauss過程」プラットフォームの別例

Gauss過程のモデルの例

この例では、地面から2つの帯水層に掘った試錐孔からの水流を示すデータを使用します。各因子を指定してGauss過程を実行すると、それらの因子が応答Yに与える影響を理解できます。

1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Design Experiment¥Borehole Latin Hypercube.jmp」を開きます。
2. ［**分析**］>［**発展的なモデル**］>［**Gauss過程**］を選択します。
3. 「**log10 Rw**」から「**Kw**」までを選択し、［**X**］をクリックします。
4. 「**Y**」を選択し、［**Y**］をクリックします。
5. JMP Pro の場合は、分析の実行時間を短縮するため、［高速Gauss過程］を選択したままにします。
6. ［**OK**］をクリックします。

図14.5「Borehole Latin Hypercube」のレポート



「予測値と実測値のプロット」を見ると、点がY = Xの線に沿っていることから、Gauss過程の予測モデルが真の関数をよく近似していると判断できます。「モデルのレポート」では、最初の因子である「log10 Rw」の総感度が最も高くなっています。「log10 Rw」の総感度の推定値は、応答の変動の90%以上を説明しています。θの値が小さい因子は、予測式にほとんど（またはまったく）影響しません。

メモ：画面に表示される推定値は、高速Gauss過程アルゴリズムを使った図14.5の値とは異なる場合があります。

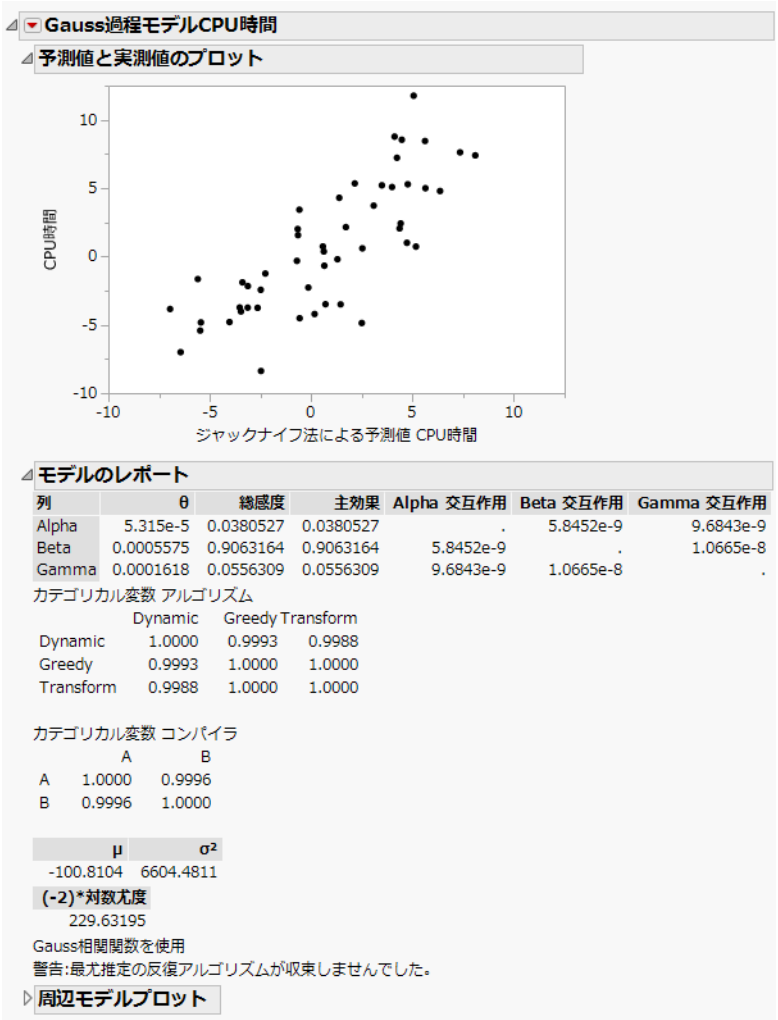
JMP PRO カテゴリカルな説明変数を使ったGauss過程モデルの例

この例では、「Algorithm Data.jmp」サンプルデータを使用します。データは、実験回数50のSpace Filling計画からシミュレートしたCPU時間です。「Algorithm Factors.jmp」には、計画の因子と設定がまとめられています。この計画には、連続尺度の因子が3つ、カテゴリカルな因子が2つあります。目標は、連続尺度の因子とカテゴリカルな因子の両方を含むGauss過程モデルで**「CPU時間」**を予測することです。

1. ［ヘルプ］>［サンプルデータライブラリ］を選択し、「Design Experiment¥Algorithm Data.jmp」を開きます。
2. ［分析］>［発展的なモデル］>［Gauss過程］を選択します。
3. 「Alpha」から「コンパイラ」までを選択し、［X］をクリックします。
4. 「CPU時間」を選択し、［Y］をクリックします。
5. JMP PRO［高速Gauss過程］オプションを選択したまま、分析を実行します。［OK］をクリックします。

メモ：カテゴリカルな因子があるモデルには、［高速 Gauss 過程］オプションを使う必要があります。詳細は、「カテゴリカルな説明変数を使ったモデル」（231 ページ）の節を参照してください。

図14.6「Algorithm Data」のレポート



「予測値と実測値のプロット」を見ると、「CPU時間」の予測値と実測値の間に強い相関があることがわかります。これは、Gauss過程の予測モデルが真の関数をよく近似していることを示します。「モデルのレポート」では、説明変数である**「Beta」**の総感度が最も高くなっています。これは、連続尺度の説明変数の中で、**「Beta」**が応答の変動を最も多く説明していることを示します。このレポートとは別に、カテゴリカルな説明変数の**「アルゴリズム」**と**「コンパイラ」**に対し、それぞれ「カテゴリカル変数」という行列が作成されています。これは、それぞれの変数の水準間に見られる相関を示した相関行列です。行列の非対角成分はτパラメータです。

「Gauss過程」プラットフォームの統計的詳細

連続尺度の説明変数を使ったモデル

Gauss過程モデルに連続尺度の説明変数だけが含まれる場合は、Gaussと三次の相関構造が使用されます。

Gauss相関構造は、距離に対するべき乗を2とした、積-指数（product exponential）型の関数で表されます。このモデルでは、Yが平均μで、共分散行列$\sigma^2\mathbf{R}$の正規分布に従うと仮定されます。**R**行列の要素は、次のように定義されます。

$$r_{ij} = \exp\left(-\sum_{k=1}^{K} \theta_k (x_{ik} - x_{jk})^2\right)$$

この式で、

K = 連続尺度の説明変数の数

θ_k = k番目の説明変数に対するθパラメータ

x_{ik} = 個体iに対するk番目の説明変数の値

x_{jk} = 個体jに対するk番目の説明変数の値

三次の相関構造でも、Yが平均μで、共分散行列$\sigma^2\mathbf{R}$の正規分布に従うと仮定されます。**R**行列は次の要素で構成されます。

$$r_{ij} = \prod_k \rho(d;\theta_k)$$

この式で、

$$d = x_{ik} - x_{jk}$$

$$\rho(d;\theta) = \begin{cases} 1-6(d\theta)^2+6(|d|\theta)^3, & |d| \le \frac{1}{2\theta} \\ 2(1-|d|\theta)^3, & \frac{1}{2\theta} < |d| \le \frac{1}{\theta} \\ 0, & \frac{1}{\theta} < |d| \end{cases}$$

詳細については、Santer（2003）を参照してください。三次相関構造で使用されるθパラメータは、文献で使用されているパラメータの逆数です。逆数を使えば、θがモデルに影響を与えない場合のρの値が無限でなく0になるためです。

JMP PRO カテゴリカルな説明変数を使ったモデル

Gauss過程モデルにカテゴリカルな説明変数が含まれる場合は、相関構造としてGauss相関構造が使用されます。**R**行列の要素は次のように定義されます。

$$r_{ij} = \left(\prod_{p=1}^{P} \tau_{p_{ij}} \right) \exp\left(-\sum_{k=1}^{K} \theta_k (x_{ik} - x_{jk})^2 \right)$$

この式で、

K = 連続尺度の説明変数の数

P = カテゴリカルな説明変数の数

θ_k = k番目の連続尺度の説明変数に対するθパラメータ

x_{ik} = 個体iに対するk番目の連続尺度の説明変数の値

x_{jk} = 個体jに対するk番目の連続尺度の説明変数の値

$\tau_{p_{ij}}$ = 個体iに対する説明変数pの観測された水準と個体jに対する説明変数pの観測された水準との相関

カテゴリカル変数の水準の組み合わせごとにτパラメータがあります。そして、τ_{ij}は、個体iにおけるカテゴリカル変数の水準と、個体jにおけるカテゴリカル変数の水準との組み合わせに対応します。このため、共分散行列の要素のr_{ij}は、個体iのカテゴリカル変数の水準と、個体jのカテゴリカル変数の水準に依存します。詳細については、Qian et al.（2008）を参照してください。

分散計算式のパラメータ化

カテゴリカルな説明変数が含まれていない場合は、保存される分散計算式は、前々節で説明した**R**の計算式が使用されます。カテゴリカルな説明変数が含まれる場合、**R**の各要素を求めるのに、次の計算式が使われます。

$$r_{ij} = \exp\left(-\sum_{k=1}^{K} \theta_k (x_{ik} - x_{jk})^2 - \sum_{p=1}^{P} \phi_{p_{ij}}\right)$$

この式において、$\phi_{p_{ij}} = -\ln(\tau_{p_{ij}})$です。その他の変数は前に定義したとおりです。

モデルのあてはめの詳細

モデルパラメータは最尤法であてはめられます。あてはめたパラメータは、プラットフォームのレポートに表示されます。パラメータには以下のとおりです。

- μは、Gauss過程の平均
- σ^2は、Gauss過程の分散
- θは、**R**の定義におけるθ_kの値に相当します。
- カテゴリカル変数の行列の非対角成分は、**R**の定義におけるτ_{pij}の値に相当します。

メモ：「分散行列が特異になるのを防ぐために、ナゲット効果がモデルに追加されました。」というメモが表示されたときは、分散行列を逆行列が存在するようにするためにリッジパラメータが加えられています。

第15章

時系列分析

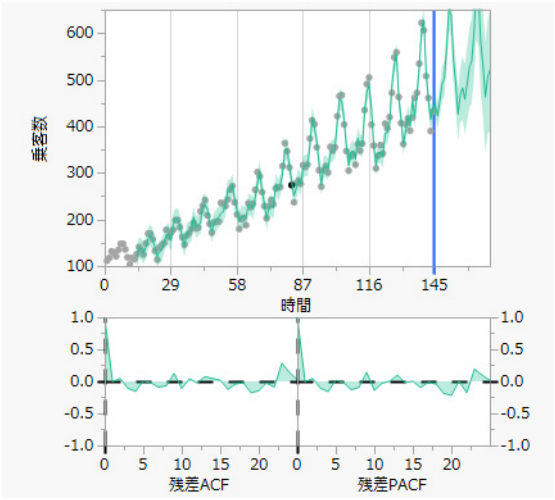
時系列モデルや伝達関数モデルのあてはめ

「時系列分析」プラットフォームでは、一変量の時系列データに対する分析や予測を行えます。時系列データとは、一定の時間間隔を置いて測定された観測値のセットを指します。通常、時間的に近い観測値には相関があります。時系列分析は、観測値間の依存関係を利用して未来の時系列を予測します。

時系列データに共通する特性として、季節性、トレンド、自己相関が挙げられます。「時系列分析」プラットフォームには、これらの特性を扱うためのオプションが用意されています。バリオグラムや自己相関プロット、偏自己相関プロット、スペクトル密度プロットなどを使い、時系列を予測するのにどのモデルが適しているかを調べることができます。また、分解の手法がいくつか用意されているため、データから季節性の変動や一般的な変動を取り除き、分析しやすい状態にできます。また、より洗練されたARIMAモデルをあてはめ、季節性と長期的なトレンドを1つのモデルに取り入れることができます。

入力系列を指定すれば、伝達関数モデルをあてはめることも可能です。

図15.1 予測プロット



「時系列分析」プラットフォームの概要

時系列データとは、一定の時間間隔を置いて測定された観測値のセット（$y_1, y_2, ..., y_N$）を指します。時系列データの例としては、四半期ごとの売上、月間平均気温、太陽黒点の数などが挙げられます。「時系列分析」プラットフォームを使えば、このようなデータに含まれるパターンやトレンドを調べることができます。その後、見つかったパターンやトレンドを利用して未来の時系列を予測します。

時系列データに共通する特性として、季節性、トレンド、自己相関が挙げられます。**季節性**とは、一定の期間に生じるパターンを指します。たとえば、1か月に1回記録するデータの場合、夏のデータはどの年度でも似ている可能性があります。**トレンド**は、時系列の長期的な動きを指します。時間の経過に伴う緩やかな値の増減などです。**自己相関**は、時系列内の各点と、時系列内の以前の値との間に見られる相関の度合いを示します。

「時系列分析」プラットフォームには、さまざまなモデルと予測手法が用意されています。ただし、そのすべてがトレンドや季節性を扱えるわけではありません。適切なモデルを選ぶためには、時系列がどのような特性を持っているかを特定する必要があります。「時系列分析」プラットフォームでは、バリオグラムや自己相関プロット、偏自己相関プロット、スペクトラル密度プロットなどを使って、時系列の展開の予測に適したモデルの種類を特定できます。差分演算と分解の手法もいくつか用意されているため、データから季節性の変動や一般的な変動を取り除き、分析しやすい状態にできます。

また、より洗練されたモデルをあてはめて、季節性や長期的なトレンドを取り入れることも可能です。このような特徴を持つモデルとして、指数平滑化法の高度な形態であるWinter法の加法型モデルが挙げられます。さらにこのプラットフォームでは、ARIMAモデル（自己回帰和分移動平均モデル）をあてはめることができます。ARIMAモデルは、統計的に見て最も複雑である一方、最も柔軟なモデルだと言えます。高度な指数平滑加法とARIMAモデルは、どちらも解釈が難しいものの、予測ツールとして非常に優れています。

入力系列を指定すれば、伝達関数モデルをあてはめることも可能です。

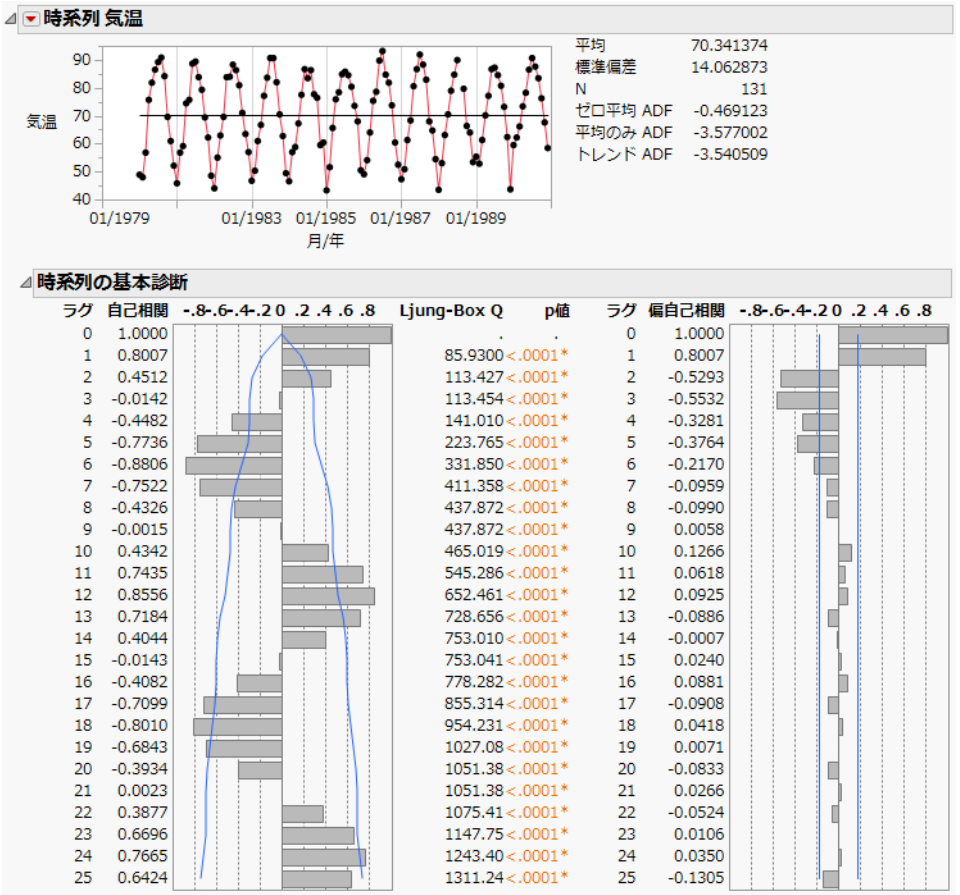
「時系列分析」プラットフォームの例

この例では、1980～1990年の月間最高気温（華氏）をまとめた「Raleigh Temps.jmp」データテーブルを分析します。「時系列分析」プラットフォームを使って時系列を調べ、今後2年間の月間最高気温を予測します。

1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Time Series」フォルダにある「Raleigh Temps.jmp」を開きます。
2. ［**分析**］>［**発展的なモデル**］>［**時系列分析**］を選択します。
3. 「**気温**」を選択し、［**Y, 時系列**］をクリックします。
4. 「**月/年**」を選択し、［**X, 時間ID**］をクリックします。
5. 「**予測する期数**」のボックスに「24」と入力します。

   これは、データにあてはめたモデルを使って予測する未来の期間の数です。今後2年間の気温を予測したいので、24か月に設定します。
6. ［**OK**］をクリックします。

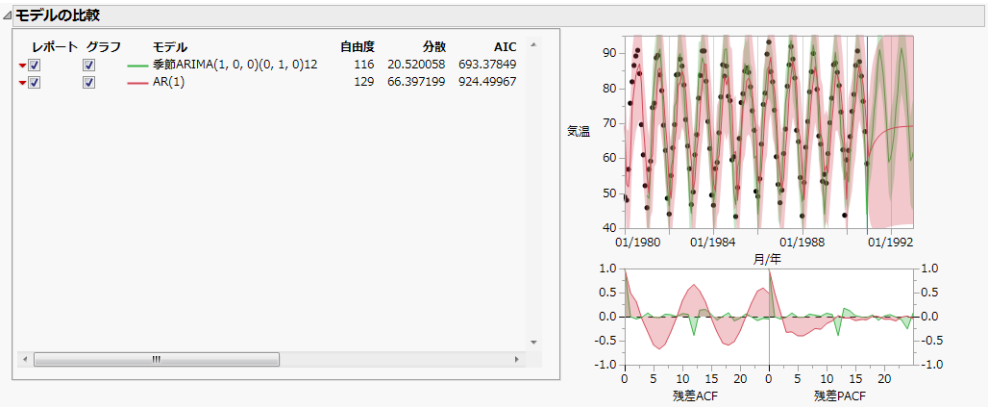
図15.2「Raleigh Temps.jmp」の時系列分析レポート



「時系列」グラフを見ると、時系列が周期的であることがわかります。この周期的な要素は自己相関チャートにも表れています。1ラグ離れている点は、正の相関があり、自己相関の値が0.8007です。点がさらに離れるにつれ、相関は負の相関になり、その後また正に転じます。このパターンが繰り返されています。「時系列」グラフと自己相関チャートに見られるこのようなパターンは、時系列に季節性がある証拠です。

7. 「時系列」の赤い三角ボタンをクリックし、［**ARIMA**］を選択します。
8. 時系列に自己相関の証拠があるため、「*p*, 自己回帰次数」を「1」に設定します。
9. ［**推定**］をクリックします。
10. 「時系列」の赤い三角ボタンをクリックし、［**季節 ARIMA**］を選択します。
11. 時系列に自己相関の証拠があるため、「ARIMA」パネルで「*p*, 自己回帰次数」を「1」に設定します。
12. 時系列に季節性の証拠があるため、「季節 ARIMA」パネルで「*D*, 差分の次数」を「1」に設定します。
13. ［**推定**］をクリックします。
14. 「モデルの比較」表の**「グラフ」**列で、両方のモデルのボックスをオンにします。

図15.3「Raleigh Temps.jmp」の「モデルの比較」表

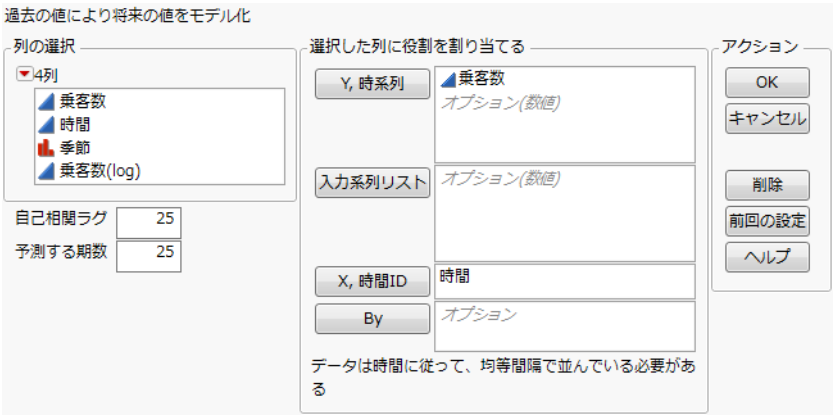


「モデルの比較」表は、AICの値の降順に並べられています。つまり、レポートの一番上に表示されているモデルが最良のモデルです。季節ARIMAモデルのAICの値（689.4）は、通常のARIMAモデルの値（920.5）を大きく下回っています。グラフを見ると、ARIMAモデルは観測されたデータ点をかなりよく予測している一方で、残差が季節ARIMAモデルより大きくなっています。また、狭い予測区間では、季節ARIMAモデルの方が未来の観測値を現実的に予測しています。これらの結果は、時系列に季節的要素の証拠があることと一致しています。

「時系列」プラットフォームの起動

「時系列」プラットフォームを起動するには、[**分析**] > [**発展的なモデル**] > [**時系列**] を選択します。「Seriesg.jmp」データテーブルで「時系列」起動ウィンドウを開くと、図15.4のようになります。

図15.4「時系列分析」起動ウィンドウ



「時系列分析」プラットフォームの起動ウィンドウには、以下のオプションがあります。

Y, 時系列 時系列変数として1つまたは複数の列を指定します。この変数はY軸に表示されます。

入力系列リスト 入力系列変数として1つまたは複数の列を指定します。この変数は「入力系列パネル」に表示され、伝達関数モデルに使用されます。時系列または指示変数の数値でなければなりません。

X, 時間ID 時間軸（X軸）のラベルに使用する変数を指定します。［X, 時間ID］を指定しなかった場合は、行番号が使用されます。

メモ：［X, 時間ID］変数を使用する場合は、［時間の単位］列プロパティを使って時間の単位を指定できます。選択肢には、［年］、［四半期］、［月］、［週］、［日］、［時］、［分］、［秒］があります。これにより、予測値をプロットする際の間隔が決まります。指定しなかった場合は、時間が等間隔の数値データとして扱われます。

注意：［X, 時間ID］に指定した変数の観測値は、均等な間隔で並んでいると仮定されます。しかし、「時系列分析」プラットフォームは、タイムスタンプの数値が増加しているかどうかしか確認しません。観測値の間隔が等しいかどうかは検証されません。

By 変数の列を指定すると、各水準ごとに個別の分析が行われます。複数のBy変数を割り当てた場合、それらのBy変数の水準の組み合わせごとに個別に分析が行われます。

メモ：By変数を使用する場合は、各水準に観測値がいくつあるかによっては自己相関ラグの数を変更する必要があります。ラグの数は1より大きく、水準あたりの観測値数より小さくなければなりません。

自己相関ラグ 自己相関と偏自己相関の計算に使うラグの数を指定します。これは相関計算に使用するタイムラグの最大数です。値は1より大きく、行数より小さくなければなりません。デフォルトでは25に設定されています。

ヒント：通常は、オブザベーションの数を n としたとき、$n/4$ 程度までをラグの最高数とします。

予測する期数 データにあてはめた各モデルで予測する未来の期間の数を指定します。デフォルトでは25に設定されています。

「時系列」レポート

デフォルトの「時系列」レポートには、時系列グラフ、要約統計量、時系列変数の検定、基本診断レポートが表示されます。起動ウィンドウで［入力系列リスト］の列を指定した場合は、「伝達関数分析」と「入力系列パネル」の各レポートも表示されます。「伝達関数分析」レポートと「入力系列パネル」レポートには、最初、「時系列」レポートと同じ情報が表示されます。

時系列グラフ

「時系列」グラフは時間IDと各時系列をプロットします。時間IDが指定されていない場合は、代わりに行番号が使用されます。また、**ADF（Augmented Dickey-Fuller）**検定を使った定常性の検定も行われます。「時系列」グラフの横には、次のような検定と要約統計量が表示されます。

平均　標本の平均。

標準偏差　標本の標準偏差。

N　時系列の長さ。

ゼロ平均 ADF　次の式で表されるゼロ平均のランダムウォークに対する検定。

$$x_t = \phi x_{t-1} + e_t$$

平均のみ ADF　次の式で表されるゼロでない平均のランダムウォークに対する検定。

$$x_t - \mu = \phi(x_{t-1} - \mu) + e_t$$

トレンド ADF　次の式で表される、ゼロでない平均と線形のトレンドを持つランダムウォークに対する検定。

$$x_t - \mu - \beta t = \phi[x_{t-1} - \mu - \beta(t-1)] + e_t$$

「時系列の基本診断」チャート

「時系列の基本診断」チャートに表示される情報は、「時系列」の赤い三角ボタンのメニューでどのオプションを選択するかによって異なります。このメニューの中で「時系列の基本診断」チャートに情報を表示するオプションは、［自己相関］、［偏自己相関］、［バリオグラム］、［AR係数］です。デフォルトでは、自己相関と偏自己相関が表示されます。

自己相関チャート

［自己相関］オプションを選択すると、「時系列の基本診断」チャートに以下の列が表示されます。

ラグ　点間の期数。

メモ：分析の全体像把握のため、ラグの番号は0から始まります。相関の計算をラグ1から始める場合は、グラフの作成前に、JMPの環境設定を変更してください。［**ファイル**］>［**環境設定**］>［**プラットフォーム**］>［**時系列分析**］を選択し、［**自己相関プロットでラグ0を非表示**］チェックボックスをオンにします。

自己相関 第kラグの自己相関は、次式で計算されます。

$$r_k = \frac{c_k}{c_0} \text{この式で、} c_k = \frac{1}{N}\sum_{t=k+1}^{N}(y_t - \bar{y})(y_{t-k} - \bar{y})$$

$\bar{y}$は欠測していないN個の点の平均です。定義上、1番上の自己相関（ラグ0の自己相関）は、常に1になります。

棒グラフの棒の長さは、自己相関を表しています。青色の曲線は、おおよその95%信頼区間を表し、次の式で計算される標準誤差の大まかな近似を2倍（±2標準誤差）したものです。

$$\mathrm{SE}_k = \sqrt{\frac{1}{N}\left(1 + 2\sum_{i=1}^{k-1} r_i^2\right)}$$

Ljung-Box Q 先頭からそのラグまでの複数の自己相関のうち、少なくとも1つが0と有意に異なることを検定する統計量です。この検定が有意であれば、その時系列はホワイトノイズではないと判断できます。なお、これらのグラフや検定は、元の時系列データだけでなく、モデルの残差を診断するときにも使えます。Qは検定統計量です。

p値 Ljung-Box検定のp値。

偏自己相関チャート

［偏自己相関］オプションを選択すると、「時系列の基本診断」チャートに以下の列が表示されます。

ラグ 点間の期数。

偏自己相関 第kラグの偏自己相関。

棒グラフの棒の長さは、偏自己相関を表しています。2本の青い線は95%予測区間の大まかな近似を示しています。これは、次の式で計算される標準誤差の大まかな近似を2倍したもの（±2標準誤差）です。

$$\mathrm{SE}_k = \frac{1}{\sqrt{n}}$$

バリオグラムチャート

［バリオグラム］オプションを選択すると、「時系列の基本診断」チャートに以下の列が表示されます。

ラグ 点間の期数。

バリオグラム バリオグラムでは、k個のラグだけ離れた点との間に見られる差の分散が、1つのラグだけ離れた点との差の分散と比較されます。バリオグラムは、次の計算式を使って自己相関から計算されます。

$$V_k = \frac{1 - r_{k+1}}{1 - r_1}$$

ここで、r_kはラグkにおける自己相関を表します。

AR係数チャート

［AR係数］オプションを選択すると、「時系列の基本診断」チャートに以下の列が表示されます。

ラグ　点間の期数。

AR係数　ここで算出される値は、自己回帰係数だけからなる、高次の自己回帰モデルをあてはめたときに得られる推定値を近似したものです。

「時系列分析」プラットフォームのオプション

「時系列」、「伝達関数分析」、「入力系列」の赤い三角ボタンのメニューには、同じオプションが表示されます。

時系列の診断

グラフ　時系列グラフを制御するオプションを含むサブメニューが開きます。

時系列グラフ　時系列グラフを表示します。

点の表示　時系列グラフ内に点を表示します。

接続線　時系列グラフ内に点を結ぶ線を表示します。

平均線　時系列グラフ内で、時系列の平均を示す水平な線を表示します。

自己相関　「時系列の基本診断」チャートに「自己相関」プロットを表示します。自己相関グラフは、指定されたラグ（時間）だけ離れたすべての点のペアから相関を計算したものです。「自己相関チャート」（238ページ）を参照してください。

ヒント：標本の自己相関グラフは、**標本自己相関関数**とも呼ばれます。

偏自己相関　「時系列の基本診断」チャートに「偏自己相関」グラフを表示します。偏自己相関グラフは、指定されたラグ（時間）だけ離れたすべての点のペアから偏相関を計算したものです。「偏自己相関チャート」（239ページ）を参照してください。

ヒント：自己相関グラフと偏自己相関グラフは、時系列の定常性（時間が経過しても平均と標準偏差が一定であるかどうか）と時系列にあてはめる適切なモデルを決める手掛かりになります。

バリオグラム　「時系列の基本診断」チャートにバリオグラムを表示します。「バリオグラムチャート」（239ページ）を参照してください。

AR係数　「時系列の基本診断」チャートにAR（自己回帰）係数の最小2乗法による推定値のグラフを表示します。「AR係数チャート」（240ページ）を参照してください。

スペクトル密度　スペクトル密度を期間または周波数の関数としてグラフにしたものを表示します。同時にホワイトノイズの検定も行われ、2種類の検定の結果が表示されます。「「スペクトル密度」レポート」（258ページ）および「スペクトラル密度の統計的詳細」（264ページ）を参照してください。

差分と分解

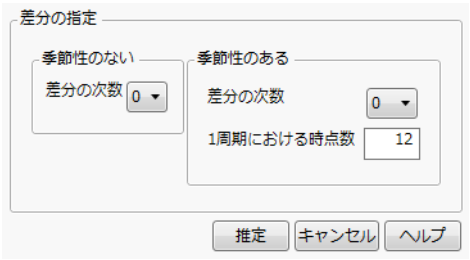
差分　「差分の指定」ウィンドウ（図15.5）が開きます。このウィンドウで、時系列に適用したい差分演算を指定します。時系列内の値の差分を取ると、定常でない時系列を定常化できます。差分は次の式で計算されます。

$$w_t = (1 - \mathrm{B})^d (1 - \mathrm{B}^s)^D y_t$$

tは時間に対する通し番号、**B**は$\mathbf{B}y_t = y_{t-1}$と定義されたラグ演算子です。

メモ：トレンドや季節性がある場合など、時系列の多くは、平均が一定しません。そのように定常でない時系列を、ARMAモデルを始めとする定常性を仮定した時系列モデルで記述することはできません。トレンドや季節性を除外し、差分を取った定常な時系列を作成することで、定常性が仮定されるモデルを使った時系列の記述が可能になります。

図15.5「差分の指定」ウィンドウ



「差分の指定」ウィンドウでは、季節性のない差分の次数であるd、季節性のある差分の次数D、1周期における時点数sを指定します。差分の次数としてゼロを選択すると、その差分は計算されません。差分演算を指定して［**推定**］をクリックするたびに、レポートウィンドウに「差分」レポートが追加されます。詳細は、「「時系列分析」プラットフォームの別例」（259ページ）を参照してください。

分解　分解の手法を含むサブメニューが開きます。時系列の分解は、データに見られる線形のトレンドや季節的周期を取り除きます。それにより、モデルの推定精度が高まります。分解の手法として3つのオプションが用意されています。

線形トレンドの除去　線形回帰モデルを使って原系列の線形トレンドを推定し、データから線形トレンドを取り除きます。レポートウィンドウにトレンドを除去した時系列の「時系列」レポートと線形トレンドに関する情報が表示されます。「「時系列」レポート」（237ページ）および「「線形トレンド」レポート」（250ページ）を参照してください。

周期性の除去 単一の余弦関数（コサイン関数）によって原系列に含まれる周期性要素を推定し、それを原系列から除去します。［周期性の除去］オプションを選択すると、「周期を定義」ウィンドウが開きます。ここでは、1 周期における時点数と、原系列データから定数を引くかどうかを指定します。レポートウィンドウに周期性を除去した時系列の「時系列」レポートと周期に関する情報が表示されます。「「時系列」レポート」（237 ページ）および「「周期性」レポート」（250 ページ）を参照してください。

X11 アメリカ合衆国国勢調査局によって開発された X11 手法を使ってトレンドと季節性の効果を除去します（Shiskin et. al., 1967）。X11 手法の詳細については、「X11 法による分解の統計的詳細」（264 ページ）を参照してください。このオプションを選択すると、「分解方法の選択」ウィンドウが開きます。ここでは、乗法型または加法型の X11 調整を指定します。［**OK**］をクリックすると、レポートウィンドウに「X11」レポートが追加されます。「「X11」レポート」（250 ページ）を参照してください。

X11 オプションは、月次データまたは四半期データでのみ使用できます。「**X, 時間 ID**」列の値が、月または四半期の均等な間隔で並んだ数値でなければならず、隙間や欠測値があってはなりません。これらの条件を満たしていない時間列に対して X11 を適用しようとすると、エラーになります。

メモ：［**線形トレンドの除去**］または［**周期性の除去**］を選択した場合は、データテーブルにトレンドまたは周期性が除去されたデータを含む列が追加されます。オプションを選択した時点でこの列がすでにデータテーブル内に存在する場合は、列が上書きされます。

ヒント：一般に、時系列データを分解するには、まず線形トレンドを除去し、次に 12 か月のような長い周期を除去します。その後、6 か月のような短い周期を除去します。

ラグプロットの表示 時点 *t* +/- *p* における観測値を *X* 軸、時点 *t* における観測値を *Y* 軸にプロットしたラグプロットが表示されます。+/- *p* はラグを指します。このプロットからは、時点 *t* におけるある観測値が時点 *t* +/- *p* における別の観測値とどのように関連しているかがわかります。プロットに明確な構造が見られない場合、観測値の間に関連はありません。プロットに構造が見られる場合は、一定の期間における観測値の間に何らかの関係があります。時系列モデルの作成において、構造を特定することは重要です。

相互相関 （「伝達関数分析」の赤い三角ボタンのメニューのみ。）レポートに相互相関プロットを表示します。プロットの長さは、自己相関プロットの 2 倍（2 × ACF length + 1）です。相互相関プロットは、出力系列と入力系列との相関を表したものです。2 本の青い線は標準誤差を表します。

白色化 （「入力系列」の赤い三角ボタンのメニューのみ。）「白色化の指定」ウィンドウが開き、白色化の次数を設定することができます。白色化は、伝達関数モデルの特定に役立つ手法です。白色化についての詳細は、Box et al.（1994）を参照してください。

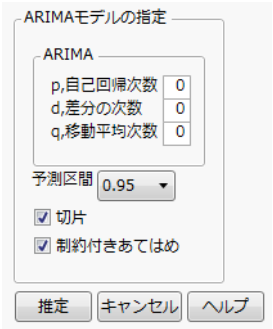
ARIMA モデルと季節 ARIMA モデル

ARIMA モデル 「ARIMA の指定」ウィンドウが開き、あてはめたい ARIMA モデルを指定できます。ARIMA モデルは、過去の値や誤差（**ランダムショック**または**イノベーション**ともいう）を線形結合させ

て将来値を予測するモデルです。ARIMAモデルは、最尤法によってARIMAモデルのパラメータを推定します。「ARIMAモデル」（268ページ）を参照してください。

メモ：ARIMAモデルは、一般にARIMA（*p*,*d*,*q*）と書き表されます。*p*、*d*、*q*のうち、ゼロのものがある場合、その文字は省略されます。たとえば、*p*と*d*がゼロの場合、モデルは移動平均モデルと等しくなり、MA(*q*)と記されます。

図15.6「ARIMAの指定」ウィンドウ



***p*, 自己回帰次数**　自己回帰を表すφ(B)演算子の次数（*p*）。

***d*, 差分の次数**　差分演算子の次数（*d*）。

***q*, 移動平均次数**　移動平均を表す演算子θ(B)の次数（*q*）。

予測区間　予測区間の信頼水準を0～1の値に設定できます。

切片　モデルに切片項μを含むかどうかを指定します。

制約付きあてはめ　このオプションをオンにすると、あてはめ処理に制限が課され、自己回帰パラメータは常に定常性の範囲に、また、移動平均パラメータは反転可能な範囲にとどまるようになります。

ヒント：反復計算でなかなか真の最適値が見つからない場合や、時間がかかり過ぎている場合は、このオプションをオフにしてください。「モデルの要約」表を見ると、あてはめたモデルの定常性や反転可能性がわかります。

モデルの指定を終え、[**推定**]をクリックすると、レポートウィンドウにモデルのレポートが追加されます。「レポート」（249ページ）を参照してください。

季節ARIMAモデル　「季節ARIMAの指定」ウィンドウが開き、あてはめたい季節ARIMAモデルを指定できます。ウィンドウには、「ARIMAの指定」ウィンドウと同じ要素に加え、季節要素も含まれます。「1周期における時点数」オプションでは、1周期に含める観測値の数を指定します。季節ARIMAモデルの詳細については、「季節ARIMAモデル」（269ページ）を参照してください。

メモ：季節 ARIMA モデルは、Seasonal ARIMA（*p,d,q*）（*P,D,Q*）*s* と書き表されます。

モデルの指定を終え、［**推定**］をクリックすると、レポートウィンドウにモデルのレポートが追加されます。「レポート」（249 ページ）を参照してください。

平滑化法モデル

［平滑化法モデル］メニューを選択すると、平滑化法に関係するモデルのサブメニューが表示されます。いずれかのモデルを選択すると、指定ウィンドウが開きます。この指定ウィンドウについては、「平滑化法モデルのウィンドウ」（248 ページ）を参照してください。実行すると、指定したモデルのレポートが追加されます。このレポートについては、「モデルのレポート」（253 ページ）を参照してください。平滑化法モデルは、平均やトレンドなどが時間によって変化するモデルで、一般的なモデルは次式で表されます。

$$y_t = \mu_t + \beta_t t + s(t) + a_t$$

この式で、

μ_t は時間によって変化する平均

β_t は時間によって変化する傾き

$s(t)$ は時間によって変化する季節影響を表す項

a_t はランダムショック

一般的な平滑化法モデルの詳細については、「平滑化法モデルの統計的詳細」（265 ページ）を参照してください。以下の平滑化法モデルが用意されています。

単純移動平均　複数の隣接する点の平均を使って値を推定するモデル。使用する点の数は、平滑化を行うウィンドウとして定義されます。「単純平滑化平均の指定」ウィンドウでは、平滑化を行う時間ウィンドウを指定します。指定が終わると、「単純移動平均」レポートが表示されます。デフォルトでは、時間ウィンドウに含まれる連続した観測値の平均（単純移動平均）がプロットされます。同じプロットに複数の単純移動平均モデルを追加できます。詳細については、「「単純平滑化平均の指定」ウィンドウ」（247 ページ）を参照してください。

1 重指数平滑化法　水準要素を持つモデル。「1 重指数平滑化法」（265 ページ）を参照してください。

2 重指数平滑化法　水準要素とトレンド要素を持つモデル。線形指数平滑化法の特殊形態です。「2 重（Brown）指数平滑化法」（266 ページ）を参照してください。

線形指数平滑化法　水準要素とトレンド要素を持つモデル。「線形（Holt）指数平滑化法」（266 ページ）を参照してください。

ダンプトレンド線形 - 指数平滑化法　水準要素とダンプトレンド要素を持つモデル。このモデルは、線形トレンドより複雑なトレンドが見られる時系列に適しています。「ダンプトレンド線形 - 指数平滑化法」（267 ページ）を参照してください。

季節指数平滑化法　水準要素と季節要素を持つモデル。「季節指数平滑化法」（267 ページ）を参照してください。

Winters法 水準要素、トレンド要素、および季節要素を持つモデル。「Winters法（加法型）」（268ページ）を参照してください。

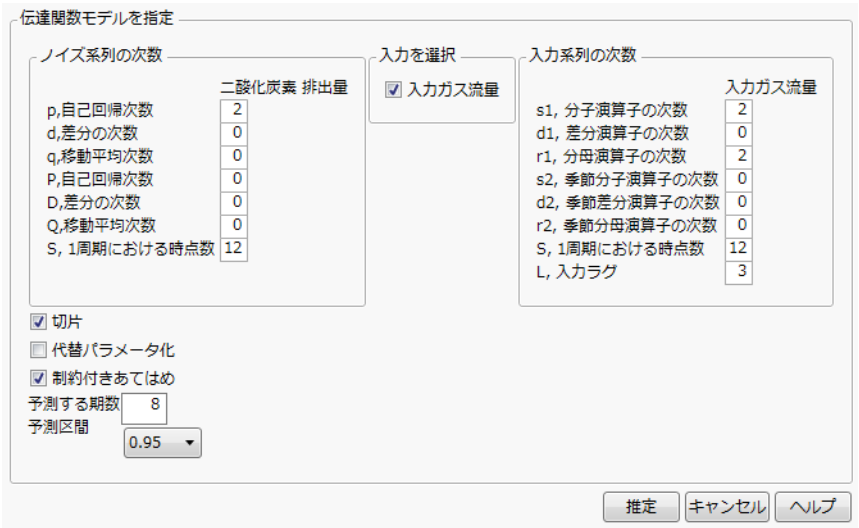
メモ: どの平滑化モデルにも、それぞれ対応する等価なARIMAモデルがあります。しかし、［ARIMA］オプションでARIMAモデルをあてはめることによって平滑化の重みを推定することはできません。平滑化モデルでは、ARIMAモデルとは別の方法でパラメータを制約しているからです。

伝達関数モデル

伝達関数 （「伝達関数モデル」の赤い三角ボタンのメニューのみ。）「伝達関数モデルの指定」ウィンドウが開きます。伝達関数モデルは、ARIMAモデルと同様、探索・あてはめ・比較を反復して行うことにより作成していきます。伝達関数モデルを作成する前のデータを検討する際に、データを白色化（prewhiten）すると効果的なことがあります。「伝達関数の統計的詳細」（270ページ）を参照してください。

メモ: 現時点での「伝達関数モデル」プラットフォームは、欠測値に完全には対応していません。

図15.7「伝達関数モデルの指定」ウィンドウ



「伝達関数モデルの指定」ウィンドウは、以下のセクションに分かれています。

ノイズ系列の次数 ノイズ系列について指定します。小文字のアルファベットは非季節性多項式の係数、大文字のアルファベットは季節性多項式の係数です。

入力を選択 モデルの入力系列を選択します。

入力系列の次数　入力系列について指定します。最初の3つの次数は非季節性多項式、次の4つの次数は季節性多項式に関連します。最後のオプションは入力ラグの次数です。

そのほか、モデルのあてはめを制御するオプションが3つあります。

切片　μがゼロかどうかを指定します。

代替パラメータ化　分子の多項式において、一般的な回帰係数を因数分解してパラメータ化するかどうかを指定します。

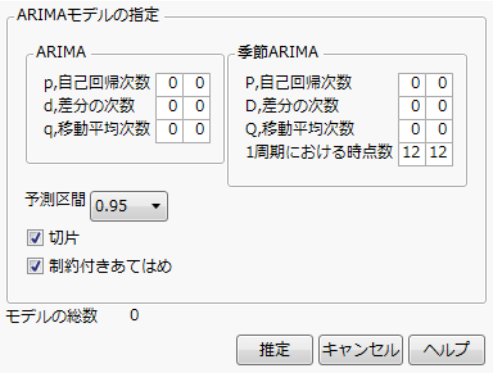
制約付きあてはめ　AR係数およびMA係数に制約をつけます。

予測する期数　予測に使用する予測期間の数を指定します。データテーブルの最後に、Y変数の欠測値と入力変数の非欠測値を含む行がある場合は、これらの行が当初の設定値として使用されます。入力変数の値は、入力変数の将来値として扱われます。

予測区間　予測区間の信頼水準を指定します。

複数のARIMAモデル　「複数のARIMAモデル」ウィンドウが開き、次数の範囲を指定することで複数のARIMAまたは季節ARIMAモデルをあてはめることができます。ウィンドウに範囲を入力すると、それに応じて「モデルの総数」が更新されます。

図15.8「複数のARIMAモデル」の指定ウィンドウ



モデルを指定し、[**推定**]をクリックすると、レポートウィンドウに各モデルのレポートが追加されます。「レポート」（249ページ）を参照してください。

スペクトル密度の保存　スペクトル密度とピリオドグラムを含むテーブルが作成されます。そのテーブルでは、*i+1*番目の行が周波数f_i=i/N（つまり1/Nのi次調波）に該当します。新しいデータテーブルには以下の列があります。

周期　i次調波の周期、$1/f_i$

周波数　i次調波の周波数、f_i

角周波数　i次調波の角周波数、$2\pi f_i$

正弦（sin）　Fourier の正弦係数、a_i

余弦（cos）　Fourier の余弦係数、b_i

ピリオドグラム　ピリオドグラム、$I(f_i)$

スペクトル密度　ピリオドグラムを平滑化したもの

予測する期数　ウィンドウが開き、あてはめたモデルで予測する未来の期間の数を指定します。デフォルト値は、「時系列分析」起動ダイアログボックスで設定した値になっています。値を変更すると、既存の予測結果と未来の予測結果の両方で新しい値が使用されます。

最大反復回数　ウィンドウが開き、ARIMA モデルの推定に使用する最適化計算での最大反復回数を指定します。

以下のオプションについて詳しくは、『JMP の使用法』の「JMP のレポート」章を参照してください。

ローカルデータフィルタ　現在のレポートにおいてデータをフィルタリングするローカルデータフィルタを表示するか、非表示にします。

やり直し　分析を繰り返すオプションや、やり直すオプションを含みます。また、［自動再計算］オプションを選択すると、このオプションに対応しているプラットフォームにおいて、データテーブルに加えた変更が、該当するレポートに即座に反映されるようになります。

スクリプトの保存　レポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。

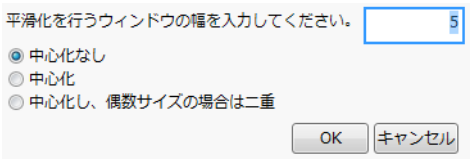
By グループのスクリプトを保存　By 変数の全水準に対するレポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。起動ウィンドウで By 変数を指定した場合のみ使用可能です。

平滑化法モデルの指定ウィンドウ

「単純平滑化平均の指定」ウィンドウ

平滑化法モデルとして［単純移動平均］を選択すると、「単純平滑化平均の指定」ウィンドウが開きます。以下の説明において、単純移動平均（SMA; Simple Moving Average）におけるウィンドウの幅を w とします。また、w 個の連続する観測値の平均を、$f_t=(y_t+y_{t-1}+y_{t-2}+\ldots y_{t-(w-2)}+y_{t-(w-1)})/w$ とします。

図15.9「単純平滑化平均の指定」ウィンドウ



平滑化を行うウィンドウの幅を入力してください。 平滑化を行うウィンドウの幅（w）は、平均を取る連続した点の数を指します。幅が大きいほど、時系列は平滑になります。

中心化なし 平滑化を行うウィンドウは、時系列の推定を行う時点tまでの点で構成されます。つまり、f_tを時点tでの移動平均とします。

中心化 平滑化を行うウィンドウは、時系列の推定を行う時点を中心として設定されます。

- wが奇数の場合、f_tを、時点t-(w-1)/2での移動平均とします。
- wが偶数の場合、f_tを、時点t-(w-1)/2での移動平均とします。ただし、データテーブルに保存されるf_tは、時点t-(w-2)/2での移動平均とします。

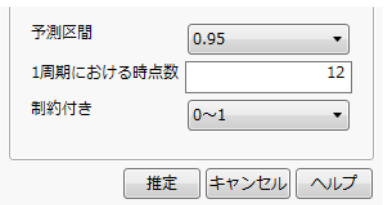
中心化し、偶数サイズの場合は二重 wが偶数の場合、平滑化を行うウィンドウは、時系列の推定を行う時点を中心とすることができません。このオプションは、ほぼ中心化された2つのウィンドウを作成し、その平均を取ります。移動平均は次のように求められます。

$$f_{t-\frac{w}{2}} = \frac{y_t + 2\sum_{i=1}^{w-1} y_{t-i} + y_{t-w}}{2w}$$

平滑化法モデルのウィンドウ

平滑化法モデルのオプションのうち、［単純移動平均］以外のものを選択すると、平滑化法モデルの指定ウィンドウが開きます。ウィンドウのタイトルやオプションの種類は、選択した平滑化法モデルによって異なります。

図15.10 平滑化法モデルの指定ウィンドウ



予測区間 予測区間の予測水準を設定します。

1周期における時点数 （季節指数平滑化法モデルの場合のみ。）季節指数平滑化法モデルの1期間あたりの観測値数を指定します。

制約付き あてはめ処理で平滑化重みに付ける制約の種類を指定できます。以下の種類があります。

0～1 平滑化重みの値を0～1の範囲に制限します。

制約なし パラメータの値が自由に変化します。

定常性 / 反転可能性 該当する ARIMA モデルが定常性および反転可能性を満たすように、パラメータが制約されます。

カスタム 個々の平滑化の重みに対して制約を設定できるように、パネルが拡張されます。各平滑化の重みの名前の隣にあるポップアップメニューを使って、**「境界あり」**、**「固定」**、**「制約なし」**のどれかを選択します。重みを固定したり、境界を設定するには、正または負の実数を指定します。

図15.11 カスタムの平滑化重み

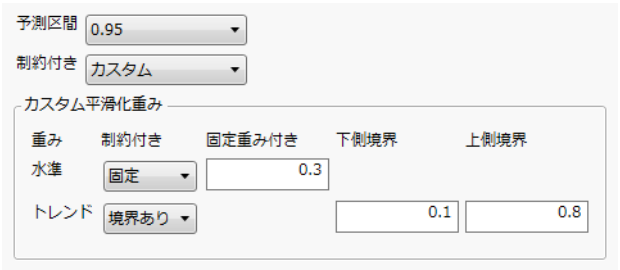


図 15.11 の例では、「水準」重み（α）が 0.3 の値に固定され、「トレンド」重み（γ）の境界が 0.1 と 0.8 に設定されています。そのため、「水準」重みは 0.3 に固定されて、「トレンド」重みは 0.1 ～ 0.8 の範囲内になります。重みを特定の値に固定すれば、その重みに対する予測値と残差が計算できます。

レポート

「差分」レポート

「差分」レポートには、差分を取った時系列の自己相関と偏自己相関のグラフが表示されます。グラフから、差分を取った時系列が定常かどうかが判断できます。

「時系列分析」プラットフォームで使用できる ARIMA モデルと季節 ARIMA モデルは、差分演算に対応しています。これらの 2 つのモデルは、まず差分演算子に従って時系列の差分を取り、次に差分を取った時系列であてはめを行います。ARIMA モデルを作成する際、［差分］オプションを準備ツールとして使用し、差分の次数としてどの値を指定すべきかを判断できます。

「差分」の赤い三角ボタンのメニューには、以下のオプションがあります。

グラフ サブメニューが開き、差分を取った時系列のプロットを制御するオプションが表示されます。「「時系列分析」プラットフォームのオプション」（240 ページ）を参照してください。

自己相関 差分を取った時系列の自己相関が表示されます。

偏自己相関 差分を取った時系列の偏自己相関が表示されます。

バリオグラム 差分を取った時系列のバリオグラムが表示されます。

保存 元のデータテーブルに、差分を取った時系列の値を含む列が保存されます。先頭にあるいくつかの要素は、差分の計算中に失われ、欠測値になります。

分解レポート

この節では、3つの分解オプションで作成されるレポートについて解説します。

- 「「線形トレンド」レポート」(250ページ)
- 「「周期性」レポート」(250ページ)
- 「「X11」レポート」(250ページ)

「線形トレンド」レポート

「線形トレンド」レポートには、データにあてはめられた線形回帰モデルにおけるβ_0とβ_1の推定値が表示されます。

$$\text{Trend}_t = \beta_0 + \beta_1 \textit{時間}$$

線形トレンドを除去した後の時系列は、$D_t = O_t$ - Trend_tです。この式で、O_tは原系列です。

「周期性」レポート

「周期性」レポートには、次式で表される周期性モデルのパラメータ推定値が表示されます。

$$\text{Cycle}_t = C + A \cdot \cos\left(2 \cdot \pi \cdot \left(\left(\frac{1}{U}\right) \cdot t + P\right)\right)$$

モデルのパラメータや変数は、次のように定義されます。

C (オプションの)定数

A 余弦波の振幅

U 1周期における時点数

P 余弦波のフェーズ

t 特定の観測値の行番号から1を引いた値

周期性を除去した後の時系列は、$D_t = O_t$ - Cycle_tです。この式で、O_tは原系列です。

「X11」レポート

どの分解方法を選んだかによって、X11オプションは、「X11 - 乗法的」レポートか「X11 - 加法的」レポートを追加します。そのレポートには、次に示す4つのプロットが含まれています。

原系列と調整済み系列 X11で調整した時系列が元の時系列のO_tに重ねられます。X11で調整した値は、乗法型調整の場合はO_t / S_t、加法型調整の場合は$O_t - S_t$です。

D10 - 最終季節要因　季節要因要素S_tを時系列にプロットします。

D12 - 最終トレンド　トレンド要素C_tを時系列にプロットします。

D13 - 最終不規則系列　不規則要素I_tを時系列にプロットします。

X11 レポートのオプション

X11 レポートの赤い三角ボタンのメニューには、次のようなオプションがあります。

各表の表示　X11 法の要約表を表示します。この要約表は、Shiskin et.al.（1967）に記述されているとおりです。表は、表15.1のような5つのカテゴリ（B～F）に分類されます。

列の保存　季節調節済み系列、季節要素（S_t）、トレンド要素（C_t）、不規則要素（I_t）といった4つの列をデータテーブルに保存します。

すべての列を保存　［各表の表示］オプションで作成された表のすべての列をデータテーブルに保存します。

表15.1 X11 出力表のカテゴリ

接頭辞	カテゴリの内容
B	季節要素、トレンド要素、不規則要素の予備推定値
C	季節要素、トレンド要素、不規則要素の中間推定値
D	季節要素、トレンド要素、不規則要素の最終推定値
E	分析表
F	要約指標

X11 法の各表についての詳細は、Shisken et.al.（1967）、または『SAS/ETS 13.1 User's Guide』（「The output from PROC X11」で検索してください）を参照してください。

「モデルの比較」レポート

モデルのあてはめが終わると、レポートウィンドウに「モデルの比較」レポートが表示されます。このレポートは、「モデルの比較」表とモデルのプロットで構成されます。新しいモデルをあてはめるたびに、「モデルの比較」表に別の色を使った新しい行が追加されます。「モデルの比較」表は、各モデルの適合度統計量をまとめたもので、同じ時系列にあてはめた複数のモデルを比較できます。モデルは、AIC の降順に並べられています。各適合度統計量の定義については、「「モデルの要約」表」（253 ページ）を参照してください。「モデルの比較」表には「重み」という統計量も表示されます。この「重み」は、「AIC 重み」を正規化したものです。「AIC 重み」は、次式で計算されます。

$$\text{AIC 重み} = \exp[-0.5(\text{AIC} - \text{最良の AIC})] / \sum_{k=1}^{K} (\exp[-0.5(\text{AIC}_k - \text{最良の AIC})])$$

Kはモデルの合計数、AIC_kはモデルkのAICです。また、「最良のAIC」は、あてはめたモデルのAICのうち、最小となっているAICです。

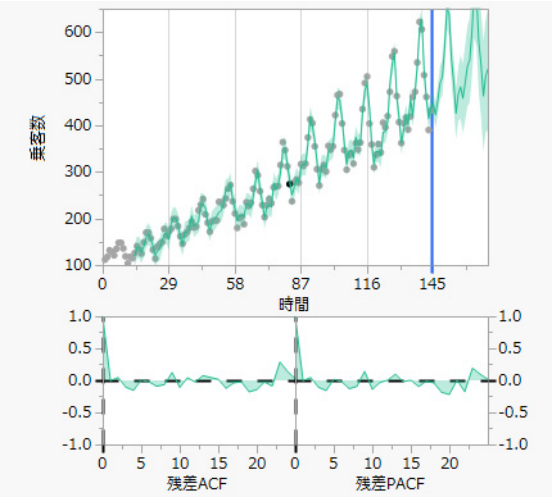
図15.12 モデルの比較

モデルの比較

レポート	グラフ	モデル	自由度	分散	AIC	SBC	R2乗	-2 対数尤度	重み	.2 .4 .6 .8	MAPE	MAE
☑	☐	季節ARIMA(0, 1, 1)(0, 1, 1)12	128	138.49122	1020.9047	1029.5302	0.990	1014.9047	1.000000		3.197010	8.990978
☑	☐	ARIMA(1, 1, 1)	140	978.9437	1394.1215	1403.0101	0.932	1388.1215	0.000000		8.687058	24.364166
☑	☐	ARMA(1, 1)	141	989.13835	1407.7483	1416.6577	0.919	1401.7483	0.000000		9.704486	25.348507
☑	☐	AR(1)	142	1134.3871	1426.1794	1432.1190	0.909	1422.1794	0.000000		9.873726	26.556219
☑	☐	MA(1)	142	4264.6121	1616.8626	1622.8022	0.693	1612.8626	0.000000		23.865796	53.716974

［**レポート**］チェックボックスを使うと、どのモデルのレポートを表示するかを指定できます。「モデルの比較」表の右側に2つのグラフが表示されます。そのうちの上側に描かれているグラフは、元データ、予測値、予測限界をプロットしたものです。その下に、残差の自己相関関数（ACF）と偏自己相関関数（PACF）のプロットが表示されます。グラフにどのモデルを表示するかは、［**グラフ**］のチェックボックスで指定します。

図15.13 モデルのプロット



「モデルの比較」レポートのオプション

「モデルの比較」レポートでは、各モデルの赤い三角ボタンのメニューに以下のオプションがあります。

新規あてはめ　モデルの指定ウィンドウが開きます。設定値を変更し、新しいモデルをあてはめることができます。

1回シミュレーション　k期先まで、1回のシミュレーションを行います。シミュレーションの結果は、「モデルの比較」の時系列グラフに表示されます。kの値を変更するには、「時系列」の赤い三角ボタンをクリックし、［予測する期数］を選択します。

複数回シミュレーション *k*期先まで、指定の数だけシミュレーションを行います。シミュレーションの結果は、「モデルの比較」の時系列グラフに表示されます。*k*の値を変更するには、「時系列」の赤い三角ボタンをクリックし、［予測する期数］を選択します。

モデルシミュレーションの削除 そのモデルのシミュレーション結果を削除します。

すべてのシミュレーションの削除 すべてのモデルのシミュレーション結果を削除します。

シミュレーションの生成 そのモデルに対してシミュレーションを生成し、結果をデータテーブルに保存します。乱数シード値、シミュレーション数、予測する期数を指定します。

シード値の設定 予測値のシミュレーションに使うシード値を指定します。

モデルのレポート

時系列分析のモデル化オプションには、モデルを時系列にあてはめるコマンドや、あてはめたモデルを使って時系列の将来値を予測するコマンドがあります。また、選択したモデルが適切なものかどうかを判断するために統計量および残差を計算するオプションもあります。モデル化オプションは何度でも選択できます。モデルを選択するたびに、「モデルの比較」表にモデルが追加され、「時系列」レポートウィンドウにあてはめの結果と予測のレポートが表示されます。「モデルの比較」表で「レポート」チェックボックスがオンになっているモデルに対し、レポートが作成されます。レポートのタイトルにモデルの名前が含まれます。

デフォルトでは、以下のレポートが表示されます。

- 「モデルの要約」表
- 「パラメータ推定値」表
- 予測プロット
- 残差
- 反復計算の履歴

「モデルの要約」表

「モデルの要約」表は、モデルのあてはめ統計量をまとめたものです。以下の説明において、*n*は系列の長さ、*k*はモデルのパラメータ数です。

自由度 誤差自由度、*n*-*k*。

誤差平方和 残差の平方和。

分散推定値 無条件の平方和（SSE）を自由度の数（*n*-*k*）で割ったもの。分散の推定値はSSE / (*n* – *k*)で求められます。これはランダムショックa_tの分散の推定値です。これについては、「ARIMA モデル」（268ページ）の節で説明します。

標準偏差 分散推定値の平方根。ランダムショックであるa_tの標準偏差の推定値です。

赤池の情報量規準（AIC） AICの値が小さいほど、モデルはよくあてはまっています。AICは次の式で計算されます。

$$\text{AIC} = -2 \times \text{対数尤度} + 2k$$

Schwarzのベイズ規準（SBC、BIC） SBCの値が小さいほど、モデルはよくあてはまっています。Schwarzのベイズ規準はベイズ情報量規準に相当します。SBCは次の式で計算されます。

$$\text{SBC} = -2 \times \text{対数尤度} + k\ln(n)$$

R2乗 R2乗は次の式で計算されます。

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

この式で、

$$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2$$

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$\hat{y}_i$ は1期先予測値、

$\bar{y}_i$ は平均y_iです。

モデルが時系列にあまりよくあてはまっていないと、モデルの誤差の平方和（SSE）が合計平方和のSSTより大きくなります。その結果、R^2がマイナスになることがあります。

自由度調整R2乗 調整R2乗は次の式で計算されます。

$$1 - \left[\frac{(n-1)}{(n-k)}(1-R^2)\right]$$

MAPE 平均絶対誤差率。次の式で計算されます。

$$\frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

MAE 平均絶対誤差。次の式で計算されます。

$$\frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|$$

-2対数尤度 パラメータ推定値に対する尤度関数値の自然対数を-2倍したもの。値が小さいほどよくあてはまっています。『基本的な回帰モデル』の付録「統計的詳細」を参照してください。

定常性　自己回帰演算子が定常性を満たしているかどうか、つまり、$\phi(z) = 0$ のすべての根が単位円の外にあるかどうかがわかります。

反転可能性　移動平均演算子が反転可能性（可逆性）を満たしているかどうか、つまり、$\theta(z) = 0$ のすべての根が単位円の外にあるかどうかがわかります。

メモ： ϕ 演算子と θ 演算子の定義については、「ARIMA モデル」（268 ページ）の節で説明しています。

「パラメータ推定値」表

各あてはめごとに、パラメータ推定値を示す「パラメータ推定値」表が作成されます。モデルの種類によってパラメータの種類は異なります。パラメータについてはそれぞれの時系列モデルの節を参照してください。「パラメータ推定値」表には、以下の列があります。

項　パラメータの名前。各モデルの節に説明があります。モデルには、**切片**や平均項を含むものがあり、その場合は関連する**定数推定値**も表示されます。定数推定値の定義については、ARIMA モデルの説明を参照してください。

因子　（乗法型の季節 ARIMA モデルの場合のみ。）そのパラメータを含むモデルの因子。乗法型の季節モデルでは、因子 1 は季節性がなく、因子 2 は季節性があります。

ラグ　（ARIMA モデルと季節 ARIMA モデルの場合のみ。）ラグの次数。つまり、各パラメータに適用される遅れ演算子の次数です。

推定値　時系列モデルのパラメータ推定値。

標準誤差　パラメータ推定値の標準誤差を推定した値。これらの推定値は、検定や予測区間の計算に使用されます。

t 値　各パラメータがゼロであるという帰無仮説の検定統計量。パラメータの検定統計量は、パラメータ推定値とその標準誤差の比です。帰無仮説が成立している場合、この統計量は Student の *t* 分布に近似的に従います。一般に、*t* 値の絶対値が 2 より大きいと、そのパラメータ推定値は有意であると判定できます。絶対値が 2 というのは、0.05 の有意水準にほぼ相当するからです。

p 値 (Prob>|t|)　各パラメータに対して計算された *p* 値。帰無仮説が真のときに、まったくの偶然だけで、計算された t 値より（絶対値が）大きい *t* 値が得られる確率を示します。

定数推定値　切片または平均項を含むモデルに対して表示されます。定数推定値の定義については、ARIMA モデルに関する説明を参照してください。

Mu　（ARIMA モデルと季節 ARIMA モデルの場合のみ。）ARIMA モデルまたは季節 ARIMA モデルの切片の推定値。

予測プロット

モデルごとに、予測値のプロットが作成されます。予測プロットは、時系列の観測値と予測値をグラフにしたものです。縦の線で 2 つの領域に分割されています。左側の領域には、1 期先予測値が観測値と一緒にプロットされています。右側の領域には、2 期以上先の予測値が、その予測区間と一緒に示されます。

予測する期数を変更するには、プラットフォームの起動ウィンドウにある「予測する期数」ボックスの設定値を変更するか、レポートのポップアップメニューにある［予測する期数］を選択して新しい数を入力します。

残差

「残差」レポートのグラフは、あてはめたモデルの残差の値を示します。値は、時系列の観測値から1期先予測値を引いたものです。残差の自己相関レポートと偏自己相関レポートも表示されます。これらのレポートで、あてはめたモデルが適切かどうかを判断できます。モデルが適切というのは、残差プロットがゼロの線を中心にほぼ正規分布に従い、残差の自己相関と偏自己相関がゼロに近いことです。

反復計算の履歴

パラメータ推定の計算では、対数尤度を最大にするようなパラメータ推定値を求めるために、反復処理が行われます。「反復計算の履歴」レポートはモデルごとに表示され、各反復における目的関数の値を示します。これを見ると、反復計算における問題点が分かるかもしれません。データに不適切なモデルをあてはめようとすると、反復を続けても計算が収束しないことがあります。「反復の履歴」表には以下の値が表示されます。

反復　反復回数。

反復計算の履歴　各ステップごとの目的関数の値。

ステップ　反復ステップの種類。

目的関数基準　目的関数の勾配の基準。

モデルレポートのオプション

それぞれのモデルレポートにある赤い三角ボタンのメニューには、以下のオプションがあります。

点の表示　予測グラフにデータ点を表示します。

予測区間の表示　予測グラフに予測区間を表示します。

列の保存　新しいデータテーブルを作成し、列にモデルの結果を保存します。

予測式の保存　新しいデータテーブルを作成し、データと予測式を保存します。

SAS ジョブの作成　ARIMA モデルに関する分析を SAS で再現するための SAS コードを作成します。

SAS でサブミット　ARIMA モデルに関する分析を再現するコードを SAS でサブミット（実行）します。SAS サーバーに接続していない場合は、接続手順の指示が表示されます。

残差統計量　表示される残差統計量の種類を制御します。各種類については「「時系列分析」プラットフォームのオプション」（240ページ）に説明がありますが、これらのオプションは残差系列に適用されます。

「伝達関数モデル」レポート

伝達関数モデルをあてはめると、「モデルの比較」表にその伝達関数モデルが追加されます。「モデルの比較」表で［**レポート**］のチェックボックスがオンになっている伝達関数モデルについては、「伝達関数モデル」レポートが作成されます。「伝達関数モデル」レポートは、次のようなレポートで構成されます。

- モデルの要約
- パラメータ推定値
- 残差
- 対話式予測
- 反復計算の履歴

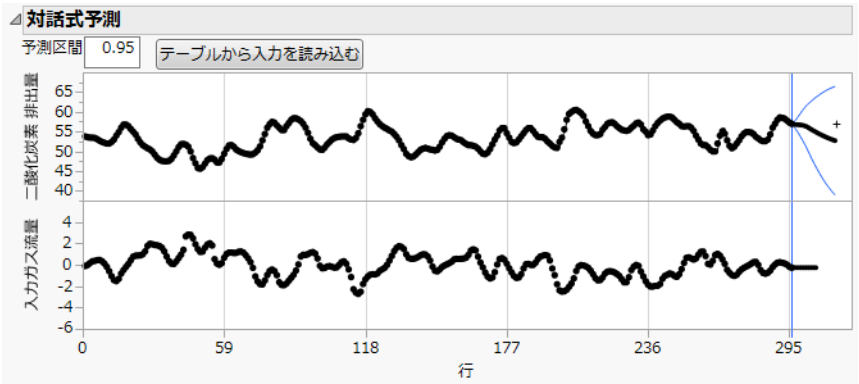
「モデルの要約」、「パラメータ推定値」、「残差」、「反復計算の履歴」に含まれる情報は、「時系列」レポートのものと同じです。これらのレポートの詳細については、「モデルのレポート」（253ページ）を参照してください。「パラメータ推定値」表の後にモデルの式が表示されます。式中の **B** は遅れ演算子です。

対話式予測

「対話式予測」レポートには、指定した予測区間の予測グラフが表示されます。予測値の予測区間が青色で表示されます。予測区間の信頼水準を変更するには、グラフの上にある「予測区間」ボックスに数値を入力します。

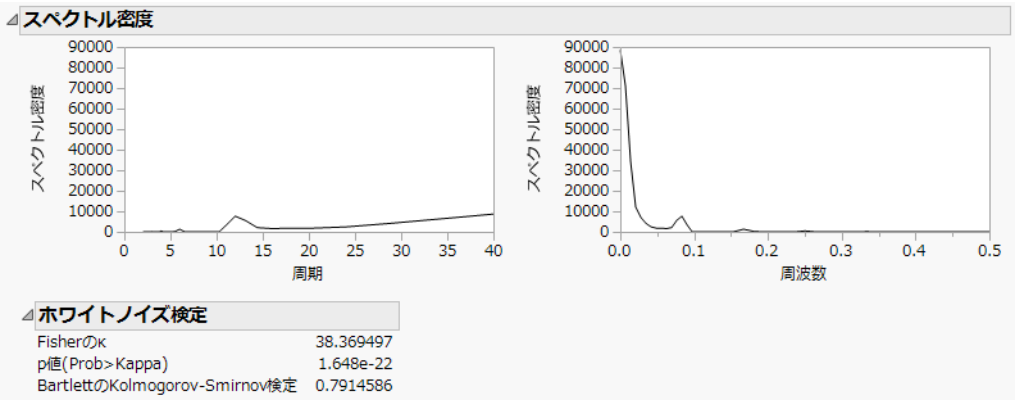
グラフ内の予測区間の数を変更するには、プラス記号をドラッグします。予測区間内で入力値を変更するには、［テーブルから入力を読み込む］ボタンを使うか、グラフ上の点を別の値にドラッグします。変更内容は出力グラフの予測区間に反映されます。

図15.14「対話式予測」グラフ



「スペクトル密度」レポート

図15.15「スペクトル密度」レポートと「ホワイトノイズ検定」レポート



「ホワイトノイズ検定」レポートには、以下の統計量が表示されます。

Fisherの κ 統計量 Fisherのカッパ（κ）統計量は、「時系列データが分散1の正規分布から取り出されたものである」とする帰無仮説を、「時系列に何らかの周期がある」とする対立仮説に照らし合わせて検定するときに使います。κはピリオドグラムの最大値である $I(f_i)$ とその平均値の比です。

p値 (Prob>Kappa) 帰無仮説が真である場合に、より大きなκの値が観測される確率。次の式で計算されます。

$$Pr(k > \kappa) = 1 - \sum_{j=0}^{q} (-1)^j \binom{q}{j} \left[\max\left(1 - \frac{jk}{q}, 0\right)\right]^{q-1}$$

この式で、

N が偶数の場合は q=N/2、N が奇数の場合は q=(N-1)/2、

κはκの観測値です。

この確率が有意水準αより小さいとき、帰無仮説は棄却されます。

BartlettのKolmogorov-Smirnov検定 この検定は、正規化された累積ピリオドグラムを、(0, 1)区間にわたる一様分布の累積分布関数と比較します。検定統計量は、累積ピリオドグラムと一様分布の累積分布関数との差の絶対値のうちで一番大きな値です。その値が $a/(\sqrt{q})$ を超える場合は、時系列が独立で同一な正規分布から取り出されたとする仮説が棄却されます。a=1.36とa=1.63は、それぞれ5%と1%の有意水準に該当します。

「時系列分析」プラットフォームの別例

ここでは、「SeriesP.jmp」データテーブルを使用して時系列分析を行います。まず、時間IDに適した新しい列を作成します。

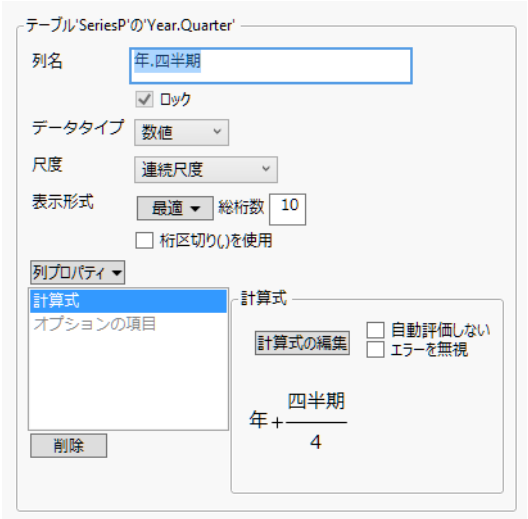
適切な［時間ID］列の作成

1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Time Series」フォルダにある「SeriesP.jmp」を開きます。

   「SeriesP.jmp」データテーブルには、応答を観測した期間を表す**「年」**列と**「四半期」**列があります。しかし、「時系列」プラットフォームのX軸には、等間隔で並んだ一意の時点を含む1つの列しか使えません。［時間ID］を指定しない場合は、自動的に行番号で代用されます。行番号では期間がわかりにくいため、**「年」**と**「四半期」**を使って時間ID列を作成しましょう。
2. ［**列**］>［**列の新規作成**］を選択します。「列名」ボックスに、**「年.四半期」**と入力します。
3. ［**列プロパティ**］>［**計算式**］を選択します。
4. **「年」**を選択し、プラス記号をクリックします。
5. **「四半期」**を選択し、除算記号をクリックします。「4」と入力し、**Enter**キーを押します。
6. ［**OK**］をクリックします。

   入力の完了した「列の新規作成」ウィンドウは、図15.16のようになります。

図15.16 列の新規作成



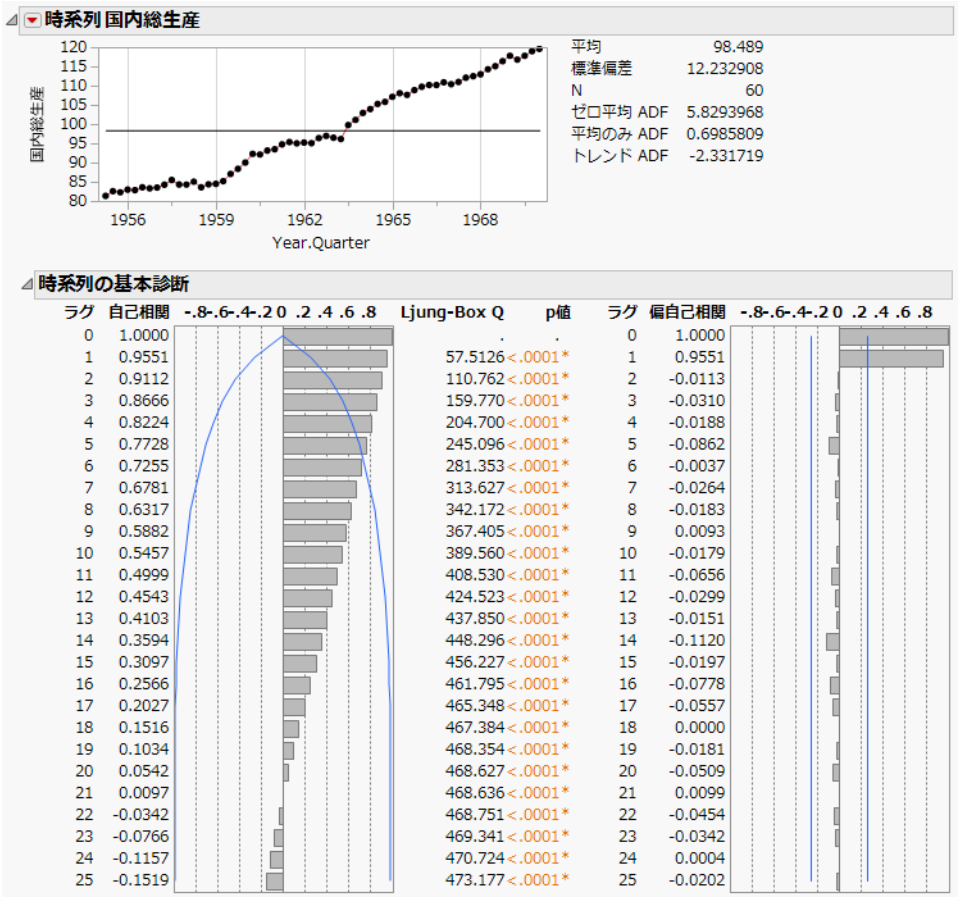
7. ［**OK**］をクリックします。

時系列分析

データテーブルに適切な時間ID列ができたので、分析に進みます。

1. ［**分析**］>［**発展的なモデル**］>［**時系列分析**］を選択します。
2. 「**国内総生産**」を選択し、［**Y, 時系列**］をクリックします。
3. 「**年.四半期**」を選択し、［**X, 時間ID**］をクリックします。
4. ［**OK**］をクリックします。

図15.17「SeriesP.jmp」の「時系列」レポート

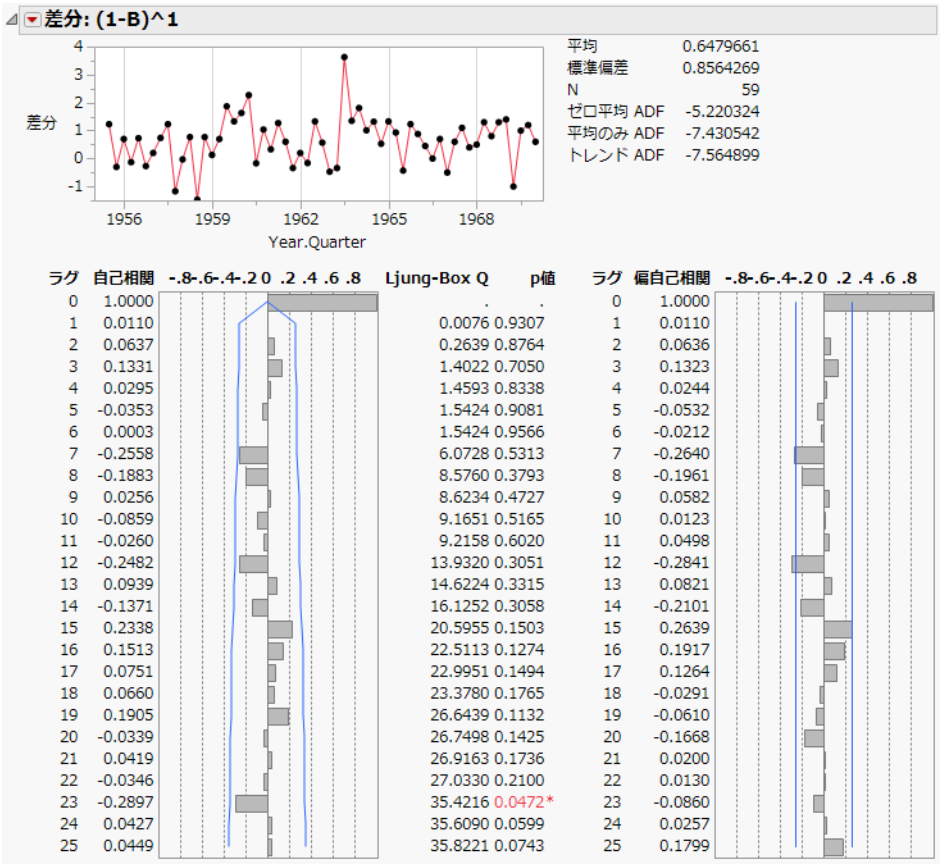


ラグ	自己相関	Ljung-Box Q	p値	ラグ	偏自己相関
0	1.0000	.	.	0	1.0000
1	0.9551	57.5126	<.0001*	1	0.9551
2	0.9112	110.762	<.0001*	2	-0.0113
3	0.8666	159.770	<.0001*	3	-0.0310
4	0.8224	204.700	<.0001*	4	-0.0188
5	0.7728	245.096	<.0001*	5	-0.0862
6	0.7255	281.353	<.0001*	6	-0.0037
7	0.6781	313.627	<.0001*	7	-0.0264
8	0.6317	342.172	<.0001*	8	-0.0183
9	0.5882	367.405	<.0001*	9	0.0093
10	0.5457	389.560	<.0001*	10	-0.0179
11	0.4999	408.530	<.0001*	11	-0.0656
12	0.4543	424.523	<.0001*	12	-0.0299
13	0.4103	437.850	<.0001*	13	-0.0151
14	0.3594	448.296	<.0001*	14	-0.1120
15	0.3097	456.227	<.0001*	15	-0.0197
16	0.2566	461.795	<.0001*	16	-0.0778
17	0.2027	465.348	<.0001*	17	-0.0557
18	0.1516	467.384	<.0001*	18	0.0000
19	0.1034	468.354	<.0001*	19	-0.0181
20	0.0542	468.627	<.0001*	20	-0.0509
21	0.0097	468.636	<.0001*	21	0.0099
22	-0.0342	468.751	<.0001*	22	-0.0454
23	-0.0766	469.341	<.0001*	23	-0.0342
24	-0.1157	470.724	<.0001*	24	0.0004
25	-0.1519	473.177	<.0001*	25	-0.0202

時系列には、線形に近い増加傾向が見られます。さらに、自己相関チャートを見ると、接近している点間に強い相関があることがわかります。ラグが1、2、3の点は、自己相関の値がそれぞれ0.9551、0.9112、0.8666です。

5. 「時系列 国内総生産」の赤い三角ボタンをクリックし、［**差分**］を選択します。
6. 季節性のない差分の次数を「1」に設定し、［**推定**］をクリックします。

図15.18「SeriesP.jmp」の「差分」レポート



「差分」レポートは、どのモデルが原系列へのあてはめに適しているかの判断に役立ちます。差分のプロットを見ると、差分を取った時系列には原系列で見られたような傾向がないことがわかります。そのため、ラグ1の差分が適切だと考えられます。また、トレンドの除外後も季節性の兆候は見られません。これらの理由から、原系列にあてはめるモデルは、線形のトレンドを扱えることが重要で、季節性を扱える必要はありません。指数平滑化法のモデルとARIMAモデルが適切でしょう。

7. 「時系列 国内総生産」の赤い三角ボタンをクリックし、**［平滑化法モデル］＞［線形指数平滑化法］**を選択します。
8. **［推定］**をクリックします。
9. 「時系列 国内総生産」の赤い三角ボタンをクリックし、**［複数のARIMAモデル］**を選択します。このオプションでは、$(p,d,q)(P,D,Q)$の値範囲で複数のARIMAモデルをあてはめることができます。
10. 「ARIMA」パネルで以下の範囲を設定します。
    - 差分のレポートから、ラグ1の差分が適切だと考えられるため、差分の次数dを1に固定します。それには、範囲を1～1に設定します。

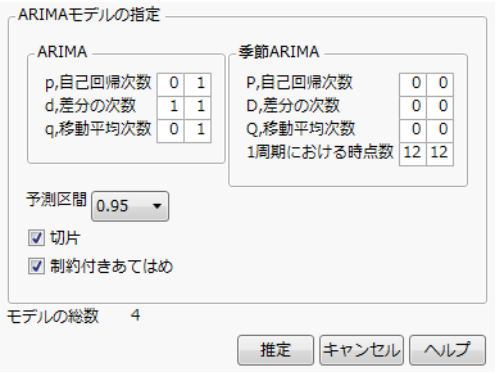
– 原系列に自己相関の証拠が見られたため、自己回帰次数pを0～1に設定します。
– 移動平均次数qを0～1に設定します。

メモ：ほとんどの場合、pとqは小さな値に抑えるだけで十分です。

– 時系列には季節性の証拠が見られなかったため、P、D、Qは0のままにします。

このように設定すると、合計4つのモデルがあてはめられます。

図15.19「複数のARIMAモデル」の指定ウィンドウ



11.［**推定**］をクリックします。

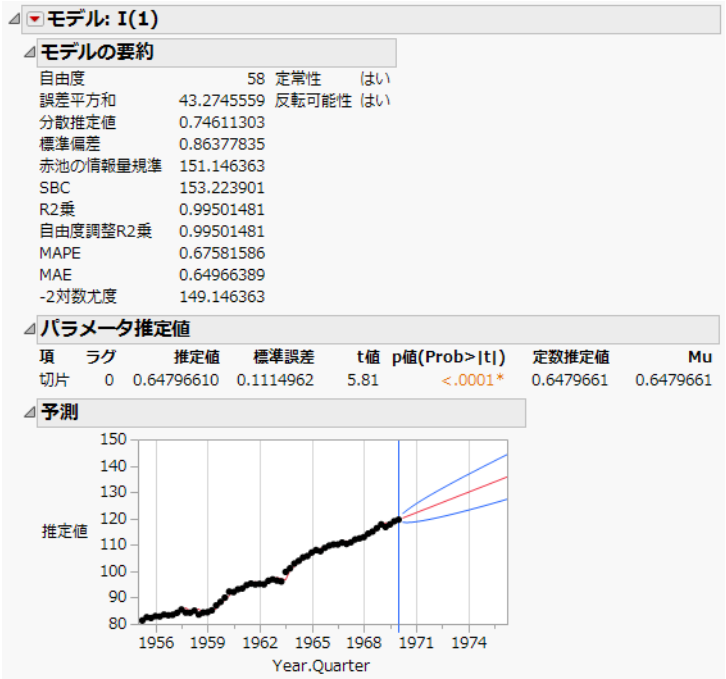
図15.20 モデルの比較

モデルの比較

レポート	グラフ	モデル	自由度	分散	AIC	SBC	R2乗	-2 対数尤度	重み	.2 .4 .6 .8	MAPE	MAE
☑	☐	I(1)	58	0.746113	151.14636	153.22390	0.995	149.14636	0.477865		0.675816	0.649664
☑	☐	ARI(1, 1)	57	0.7591096	153.13924	157.29432	0.995	149.13924	0.176424		0.676216	0.649960
☑	☐	IMA(1, 1)	57	0.7591199	153.14002	157.29510	0.995	149.14002	0.176355		0.676179	0.649933
☑	☐	ARIMA(1, 1, 1)	56	0.748633	153.91882	160.15143	0.995	147.91882	0.119474		0.668189	0.641981
☑	☐	線形(Holt)指数平滑化法	56	0.7878954	155.66571	159.78660	0.994	151.66571	0.049882		0.716452	0.686737

「モデルの比較」表は、AIC規準に従って最良と判断されたモデルから順に並んでいます。この例では、ARIMA(0,1,0)モデル（レポートでは「I(1)」と表記）が原系列に最もよくあてはまっています。「I(1)」モデルが「最良」にはなっていますが、どのモデルも適合度統計量に大きな差はありません。どれも適切だと考えることができます。

図15.21 ARIMA(0,1,0)のモデルレポート



「I(1)」のモデルレポートに予測グラフが表示されています。青色の線は予測区間を示します。「国内総生産」の予測値は線形的に増加しています。

「時系列分析」プラットフォームの統計的詳細

この節では、「時系列分析」プラットフォームの以下の要素に関する統計的詳細を紹介します。

スペクトラル密度の統計的詳細

Fourier級数の係数を最小2乗法によって推定する場合は、次のような式になります。

$$a_t = \frac{2}{N}\sum_{i=1}^{N} y_t \cos(2\pi f_i t)$$

$$b_t = \frac{2}{N}\sum_{i=1}^{N} y_t \sin(2\pi f_i t)$$

ここで、$f_i = i/N$ です。これらの係数を組み合わせたピリオドグラム $I(f_i) = \frac{N}{2}(a_i^2 + b_i^2)$ は、周波数 f_i での波の強度を表します。

ピリオドグラムを平滑化して1/(4π)を掛けると、スペクトル密度になります。

X11法による分解の統計的詳細

X11法は、原系列を乗法型または加法型で分解します。トレンド、季節性、不規則性という3つの要素に分解するために、反復的な計算が行われます。トレンド要素には、長期的なトレンドや長期的な周期性が含まれます。不規則要素には、トレンドや季節性で説明できない変動の効果が含まれます。X11手法の歴史的経緯については、『SAS/ETS 13.1 User's Guide』（「Historical Development of X-11」で検索してください）に概要があります。

乗法型調整は、次のモデルをあてはめます。

$$O_t = C_t \cdot S_t \cdot I_t$$

この式で、

O_tは、原系列

C_tは、トレンド要素

S_tは、季節要素

I_tは、不規則要素

乗法型モデルでの季節調整済み系列は、O_t / S_t。

加法型調整は、次のモデルをあてはめます。

$$O_t = C_t + S_t + I_t$$

加法型モデルでの季節調整済み系列は、$O_t - S_t$。

平滑化法モデルの統計的詳細

平滑化法モデルは、次のように定義されます。

$$y_t = \mu_t + \beta_t t + s(t) + a_t$$

この式で、

μ_t は時間によって変化する平均

β_t は時間によって変化する傾き

$s(t)$ は時間によって変化をする季節影響を表す項

a_t はランダムショック

トレンドのないモデルは $\beta_t = 0$、季節性のないモデルは $s(t) = 0$ となります。時間によって変化する各項の推定量は、次のように定義されます。

平滑化された水準 L_t で μ_t を推定

平滑化されたトレンド T_t で β_t を推定

平滑化された季節効果 S_{t-j} $(j = 0, 1, \ldots, s-1)$ で $s(t)$ を推定

すべての平滑化モデルにおいて、これらの推定値の変化を漸化式で記述したものが定義されています。平滑化の式は、**平滑化の重み**と呼ばれるモデルパラメータをもとに作成されます。

α、水準に対する平滑化の重み

γ、トレンド T_t に対する平滑化の重み

φ、トレンドダンプに対する平滑化の重み

δ、季節効果に対する平滑化の重み

これらのパラメータに共通していることは、重みが大きいほど最近のデータに与える影響が大きく、重みが小さいほど最近のデータに与える影響が小さい点です。

1重指数平滑化法

1重指数平滑化法のモデルは、$y_t = \mu_t + \alpha_t$ という式で表されます。

平滑化の式 $L_t=\alpha y_t+(1-\alpha)L_{t-1}$ は、1個の平滑化重み α を使って定義されています。この平滑化法モデルは、次のような制約をもつ ARIMA(0,1,1) モデルと等価です。

$(1-B)y_t = (1-\theta B)\alpha_t$ この式で、$\theta = 1-\alpha$

移動平均モデルの形式に書き換えると、次のような式になります。

$$y_t = a_t + \sum_{j-1}^{\infty} \alpha a_{t-j}$$

2重（Brown）指数平滑化法

2重指数平滑化法のモデルは、$y_t = \mu_t + \beta_1 t + a_t$ という式で表されます。

平滑化式は、1つの重み α を使って次のように定義されます。

$$L_t = \alpha y_t + (1-\alpha) L_{t-1}$$

$$T_t = \alpha(L_t - L_{t-1}) + (1-\alpha) T_{t-1}$$

このモデルは、次のような制約をもつ ARIMA(0,1,1)$(0,1,1)_1$ モデルと等価です。

$(1-B)^2 y_t = (1-\theta B)^2 a_t$ ここで $\theta_{1,1} = \theta_{2,1}$ および $\theta = 1-\alpha$

移動平均モデルの形式に書き換えると、次のような式になります。

$$y_t = a_t + \sum_{j=1}^{\infty} (2\alpha + (j-1)\alpha^2) a_{t-j}$$

線形（Holt）指数平滑化法

線形指数平滑化法のモデルは、$y_t = \mu_t + \beta_t t + a_t$ という式で表されます。

平滑化式は、平滑化重みの α と γ を使って次のように定義されます。

$$L_t = \alpha y_t + (1-\alpha)(L_{t-1} + T_{t-1})$$

$$T_t = \gamma(L_t - L_{t-1}) + (1-\gamma) T_{t-1}$$

このモデルは、次のような制約をもつ ARIMA(0,2,2) モデルと等価です。

$(1-B)^2 y_t = (1-\theta B - \theta_2 B^2) a_t$ ここで $\theta = 2-\alpha-\alpha\gamma$ および $\theta_2 = \alpha - 1$

移動平均モデルの形式に書き換えると、次のような式になります。

$$y_t = a_t + \sum_{j=1}^{\infty} (\alpha + j\alpha\gamma) a_{t-j}$$

ダンプトレンド線形 - 指数平滑化法

ダンプトレンド線形 - 指数平滑化法のモデルは、$y_t = \mu_t + \beta_t t + a_t$ という式で表されます。

平滑化の式は、平滑化の重み α、γ、φ を使って次のように定義されます。

$$L_t = \alpha y_t + (1-\alpha)(L_{t-1} + \varphi T_{t-1})$$

$$T_t = \gamma(L_t - L_{t-1}) + (1-\gamma)\varphi T_{t-1}$$

このモデルは、次のような制約をもつ ARIMA(1,1,2) モデルと等価です。

$$(1-\varphi B)(1-B)y_t = (1-\theta_1 B - \theta_2 B^2)a_t$$

この式で、

$$\theta_1 = 1 + \varphi - \alpha - \alpha\gamma\varphi$$

$$\theta_2 = (\alpha - 1)\varphi$$

移動平均モデルの形式に書き換えると、次のような式になります。

$$y_t = \alpha_t + \sum_{j=1}^{\infty}\left(\frac{\alpha + \alpha\gamma\varphi(\varphi^j - 1)}{\varphi - 1}\right)\alpha_{t-j}$$

季節指数平滑化法

季節指数平滑化法のモデルは、$y_t = \mu_t + s(t) + a_t$ という式で表されます。

平滑化の式は、平滑化重みの α と δ を使って次のように定義されます。

$$L_t = \alpha(y_t - S_{t-s}) + (1-\alpha)L_{t-1}$$

$$S_t = \delta(y_t - L_{t-s}) + (1-\delta)S_{t-s}$$

このモデルは、次のような制約をもつ季節 ARIMA（0,1,*s*+1）（0,1,0）$_s$ モデルと等価です。

$$(1-B)(1-B^s)y_t = (1-\theta_1 B - \theta_2 B^2 - \theta_3 B^{s+1})a_t$$

この式で、

$$\theta_1 = 1 - \alpha$$

$$\theta_2 = (1-\delta)(1-\alpha)$$

$$\theta_3 = (1-\alpha)(\delta - 1)$$

移動平均モデルの形式に書き換えると、次のような式になります。

$$y_t = a_t + \sum_{j=1}^{\infty} \psi_j a_{t-j} \text{ この式で、} \psi = \begin{cases} \alpha \text{ for } j \bmod s \neq 0 \\ \alpha + \delta(1-\alpha) \text{ for} j \bmod s = 0 \end{cases}$$

Winters法（加法型）

Winters法の加法型モデルは、$y_t = \mu_t + \beta_t t + s(t) + a_t$ という式で表されます。

平滑化の式は、重みのα、γ、δを使って次のように定義されます。

$$L_t = \alpha(y_t - S_{t-s}) + (1-\alpha)(L_{t-1} + T_{t-1})$$

$$T_t = \gamma(L_t - L_{t-1}) + (1-\gamma)T_{t-1}$$

$$S_t = \delta(y_t - L_t) + (1-\delta)S_{t-s}$$

このモデルは、次の式で表される季節ARIMA（0,1,s+1）（0,1,0）sモデルと等価です。

$$(1-B)(1-B^2)y_t = \left(1 - \sum_{i=1}^{s+1} \theta_i B^i\right) a_t$$

移動平均モデルの形式に書き換えると、次のような式になります。

$$y_t = a_t + \sum_{j=1}^{\infty} \Psi_j a_{t-j}$$

この式で、

$$\psi = \begin{cases} \alpha + j\alpha\gamma, & j \bmod s \neq 0 \\ \alpha + j\alpha\gamma + \delta(1-\alpha), & j \bmod s = 0 \end{cases}$$

ARIMAモデルの統計的詳細

ARIMAモデル

応答系列$\{y_i\}$に対するARIMAモデルの一般式は次のとおりです。

$$\phi(B)(w_t - \mu) = \theta(B)a_t$$

この式で、
tは時間を示す通し番号

Bは$By_t = y_{t-1}$と定義された遅れ演算子

$w_t = (1-B)^d y_t$ は差分をとった後の応答系列

μは切片項または平均項

$\phi(B)$ と$\theta(B)$は、それぞれ自己回帰演算子と移動平均演算子で、次のように表されます。

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p$$

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q$$

この式で、
a_t はランダムショックの並び

a_tは、互いに独立し、平均0で等分散の正規分布に従うと仮定されます。

モデルは、次のように書くこともできます。

$$\phi(B) w_t = \delta + \theta(B) a_t$$

定数の推定値δは、次の関係式で求められます。

$$\delta = \phi(B)\mu = \mu - \phi_1\mu - \phi_2\mu - \ldots - \phi_p\mu$$

季節ARIMAモデル

［**季節ARIMA**］モデルの差分演算子、自己回帰次数演算子、移動平均演算子は、季節性に関する多項式と、季節性とは関係ない多項式の積となります。

$$w_t = (1-B)^d (1-B^s)^D y_t$$

$$\varphi(B) = (1 - \varphi_{1,1}B - \varphi_{1,2}B^2 - \ldots - \varphi_{1,p}B^p)(1 - \varphi_{2,s}B^s - \varphi_{2,2s}B^{2s} - \ldots - \varphi_{2,Ps}B^{Ps})$$

$$\theta(B) = (1 - \theta_{1,1}B - \theta_{1,2}B^2 - \ldots - \theta_{1,q}B^q)(1 - \theta_{2,s}B^s - \theta_{2,2s}B^{2s} - \ldots - \theta_{2,Qs}B^{Qs})$$

上の式で、sは1周期に含める観測値の数です。それぞれの係数における1番目の添え字は、季節性を表すかどうかの番号（1は非季節因子、2は季節因子）を表し、2番目の添え字はラグ数を表します。

伝達関数の統計的詳細

入力系列の数を m としたとき、標準的な伝達関数モデルは次のように表せます。

$$Y_t - \mu = \frac{\omega_1(B)}{\delta_1(B)} X_{1,\, t-d1} + \ldots + \frac{\omega_m(B)}{\delta_m(B)} X_{m,\, m-dm} + \frac{\theta(B)}{\varphi(B)} e_t$$

この式で、

Y_t は出力系列

$X_1 \sim X_m$ は m 個の入力系列

e_t はノイズ系列

$X_{1,\, t-d1}$ は t より $d1$ 期前の系列 X_1

μ はモデルの平均

$\varphi(B)$ および $\theta(B)$ は、ARIMA モデルの自己回帰多項式と移動平均多項式

$\omega_k(B)$ および $\delta_k(B)$ は、個々の伝達関数の分子および分母関数（または多項式）で、k は $1 \sim m$ 個の入力系列に対する通し番号。

上記のモデルの多項式には、非季節性を表す部分と、季節性を表す部分の両方を含めることができます。いずれか1つだけを含めることもできますし、または季節 ARIMA のように非季節性と季節性の積から成るものを含めることもできます。モデルを指定する際、含めたくない部分がある場合は、その部分の値をデフォルトの0のままにします。

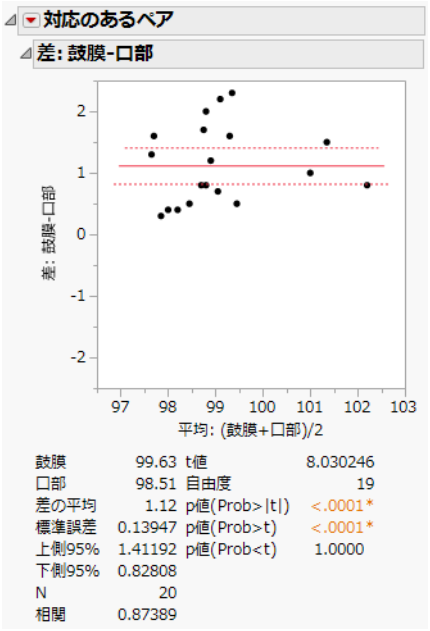
第16章

対応のあるペア分析
同一対象に対する測定値を比較する

「対応のあるペア」プラットフォームでは、相関がある2変数間の平均を比較し、その差を評価します。たとえば、同じ患者から測定した、治療前後の血圧を比較します。応答に相関があることを考慮した統計手法として、**対応のあるt検定**が使用されます。

このプラットフォームでは、ペアの平均と差をプロットした散布図が描かれます。また、両側検定、下側の片側検定、上側の片側検定という3つの対立仮説に対する検定が実行されます。なお、3列以上の応答を指定すると、それらの組み合わせに対して、処理が行われます。また、グループ列を指定した場合、単純な反復測定分散分析モデルに基づき、検定が行われます。

図16.1 対応のあるペア分析の例



「対応のあるペア」プラットフォームの概要

「対応のあるペア」プラットフォームは、各行において対応関係がある2つの応答変数を、対応のある *t* 検定によって比較します。対応のあるデータの例としては、同じ患者で反復測定した処置前後データが挙げられます。また、同一の対象を、2つの異なる計器で測定したデータも該当します。

「対応のあるペア」プラットフォームを使用するには、応答データが2列に保存されていなければいけません。すべての測定値が1列に入っているときは、次のいずれかの操作を実行してください。

- ［**テーブル**］メニューの［**列の分割**］オプションで列を2つに分けます。その後、「対応のあるペア」プラットフォームを実行してください。
- 応答データが2列に含まれている場合には、まず、その差を計算する3つ目の列を作成し、次に、差の列の平均が0かどうかを「一変量の分布」プラットフォームで検定することでも、同じ分析を実行できます。
- 応答データが1列で保存されている場合は、2元配置分散分析でも、同じ分析を実行できます。このとき、1つの因子は、2つの応答のいずれであるかを示すものです。もう1つの因子は、各個体を示すものです。「二変量の関係」の「一元配置」プラットフォームでブロック変数（個体の列）を指定するか、「モデルのあてはめ」プラットフォームで二元配置分散分析を実行してください。応答を区別する因子に対する検定が、対応のある *t* 検定と同じ結果になります。

メモ: データに対応があるときは、対応のない *t* 検定を行わないようにしてください。データを1列に積み重ね、ブロック変数を指定せずに、「二変量の関係」で一元配置分散分析を行うのは、適切ではありません。それは、応答の間にある相関を無視して検定を行うと、応答に負の相関がある場合には効果を過大評価し、応答に正の相関がある場合には効果を過小評価してしまうからです。

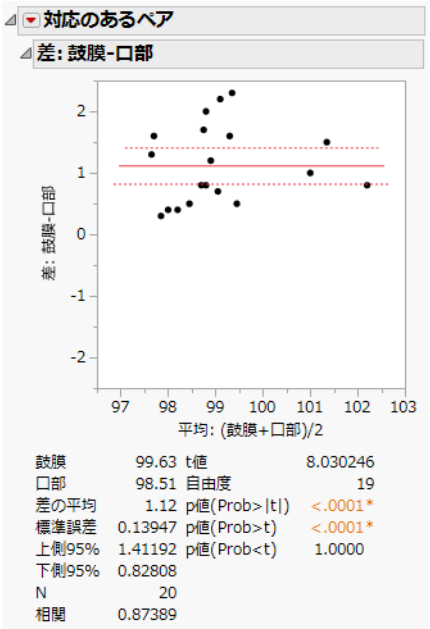
対応のあるペアの比較例

この例では、「Therm.jmp」サンプルデータを使用します。このデータには、20名の体温を、口腔体温計と鼓膜体温計（耳式体温計）で計測した結果が記録されています。2種類の体温計で計測された体温が等しいかどうかを調べます。ここでは、20名の個体差には興味がなく、2つの体温計における測定値の差に興味があります。

1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Therm.jmp」を開きます。
2. ［**分析**］>［**発展的なモデル**］>［**対応のあるペア**］を選択します。
3. 「**口部**」と「**鼓膜**」を選択し、［**Y, 対応のある応答**］をクリックします。
4. ［**OK**］をクリックします。

   レポートウィンドウが表示されます。

図16.2「対応のあるペア」レポートウィンドウ



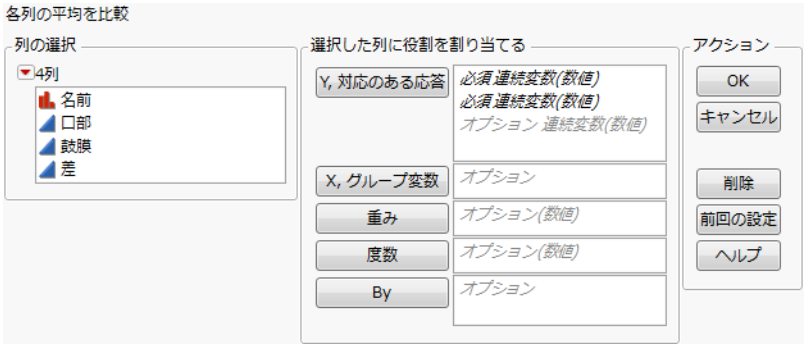
この結果から、鼓膜体温計の方が口腔体温計に比べ、平均して1.12度高い測定値が得られることがわかります。「p値(Prob>|t|)」が小さい値であるため、この差は偶然の結果ではなく、統計的に有意と判断できます。

なお、この分析は、どちらの体温計が正しいかを判定するものではなく、体温計間に差があることを示しているにすぎません。

「対応のあるペア」プラットフォームの起動

「対応のあるペア」プラットフォームを起動するには、［**分析**］>［**発展的なモデル**］>［**対応のあるペア**］を選択します。

図16.3「対応のあるペア」起動ウィンドウ



Y, 対応のある応答　2つの応答列を指定します。3つ以上の応答変数を分析する方法については、「複数のY列」（274ページ）を参照してください。

X, グループ変数　グループ間の差を比較するためのグループ変数を指定します。詳細は、「グループ効果を含めた分析」（276ページ）を参照してください。

重み　この役割を割り当てた列の数値は、分析において各行の重みとして使用されます。

度数　この役割を割り当てた列の数値は、分析において各行の度数として使用されます。

By　By変数の水準ごとに、個別に分析が行われます。

起動ウィンドウの詳細については、『JMPの使用法』の「はじめに」章を参照してください。

[OK]をクリックすると、「対応のあるペア」レポートウィンドウが開きます。「「対応のあるペア」レポート」（275ページ）を参照してください。

複数のY列

対応するペア分析では複数の応答変数を処理できます。応答変数の数が奇数の場合、可能な組み合わせがすべて分析対象となります。次の表は、応答変数の数が3つの場合の例を示しています。

Y1-Y2	Y1-Y3
	Y2-Y3

応答変数の数が偶数の場合は、可能な組み合わせをすべて分析するかどうかを確認するメッセージが表示されます。すべての組み合わせを分析しない場合は、隣接する応答変数をペアとして分析します。次の表は、応答変数の数が4つの場合の例を示しています。

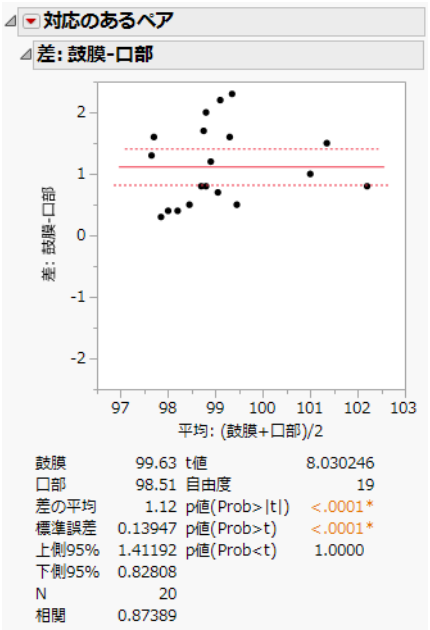
Y1-Y2	Y3-Y4

「対応のあるペア」レポート

「対応のあるペアの比較例」（272ページ）の手順に従って作業すると、図16.4のレポートウィンドウが作成されます。

「対応のあるペア」レポートには、Tukeyの差-平均プロット（Tukey diffrence-mean plot）、要約統計量、対応のあるt検定の結果が表示されます。「「差」のプロットとレポート」（275ページ）を参照してください。［**X, グループ変数**］を指定した場合は、「グループごと」レポートも表示されます。「グループ効果を含めた分析」（276ページ）を参照してください。

図16.4「対応のあるペア」レポートの例



メモ: 赤い三角ボタンをクリックするとメニューが開き、最初に表示されるレポートウィンドウにレポートを追加するための各種オプションが表示されます。「「対応のあるペア」プラットフォームのオプション」（276ページ）を参照してください。

「差」のプロットとレポート

「差」プロットには差と平均が表示されます。「差」プロットについて、次の点を確認してください。

- 水平線が平均の差を示し、その上下に95%信頼区間を表す点線が表示されています。0が信頼区間の中に挟まれている場合は、平均が0.05の水準において有意には異ならないことを意味します。この例では、信頼区間が0の水平線より上に位置しているため、差が有意であることがわかります。

- 参照枠を追加した場合は、ペアの平均は垂直線で表されます。参照枠の詳細については、「「対応のあるペア」プラットフォームのオプション」（276ページ）を参照してください。

「差」レポートには、各応答変数の平均、差の平均、差の信頼区間が表示されます。また、対応のあるt検定の結果も表示されます。

グループ効果を含めた分析

メモ：「グループごと」レポートは、［**X, グループ変数**］を指定した場合にのみ表示されます。

グループ効果を含めた分析は、単純な反復測定分析です。（「モデルのあてはめ」プラットフォームで**「MANOVA」**手法を使用して反復測定分析を行うと同じ結果が出ます。）

差の平均　対応のある2列の差を計算し、それらのグループ別平均を比較したものです。言い換えると、これは個体内因子と個体間因子の交互作用（一次因子と二次因子の交互作用）を表しています。

平均の平均　対応のある2列の平均を計算し、それらのグループ別平均を比較したものです。言い換えると、個体間因子（一次因子）の主効果を表しています。

グループ別検定　次の2つに関して、グループごとの平均が異なるかを、*F*検定によって検定します。

- **「差の平均」**は、応答の差がグループによって異なるかどうかを検定します。
- **「平均の平均」**は、応答の平均がグループによって異なるかどうかを検定します。

関連情報

- 「グループ効果を含めたペア比較の例」（277ページ）

「対応のあるペア」プラットフォームのオプション

「対応のあるペア」の赤い三角ボタンのメニューには、以下のオプションがあります。

平均値と差のプロット　「平均値と差」のプロットの表示／非表示を切り替えます。このプロットの詳細については、「「差」のプロットとレポート」（275ページ）を参照してください。

行ごとに差をプロット　「行番号ごとの差」のプロットの表示／非表示を切り替えます。

参照枠　「平均値と差」のプロットにおける参照枠の表示／非表示を切り替えます。この参照枠は、図中に傾いて表示される矩形です。参照枠は、かなり押しつぶされて表示される場合もあります。赤い縦線は、「平均の平均」を表します。参照枠は、差の範囲がデータ範囲の半分より大きい場合にはデフォルトで表示されます。

Wilcoxonの符号付順位検定 「Wilcoxonの符号付順位検定」の表示／非表示を切り替えます。Wilcoxonの符号付順位検定は、対応のあるデータの差に適用されます。これは、対応のあるt検定のノンパラメトリック版で、正の値を取る差と負の値を取る差の大きさを比較するものです。この検定では、0の差はPratt法を用いて処理します。また、差の分布が対称であると仮定します。詳細については、『基本的な統計分析』の「一変量の分布」章を参照してください。Lehman（2006）、Conover（1999、350ページ）、Cureton（1967）も参照してください。

符号検定 符号検定の表示／非表示を切り替えます。これは、対応のあるt検定のノンパラメトリック版です。差の符号（正負）だけを検定に使用します。

α水準の設定 分析で使用するα水準を変更します。レポートとプロット内の信頼区間に影響します。

以下のオプションについて詳しくは、『JMPの使用法』の「JMPのレポート」章を参照してください。

ローカルデータフィルタ 現在のレポートにおいてデータをフィルタリングするローカルデータフィルタを表示するか、非表示にします。

やり直し 分析を繰り返すオプションや、やり直すオプションを含みます。また、［自動再計算］オプションを選択すると、このオプションに対応しているプラットフォームにおいて、データテーブルに加えた変更が、該当するレポートに即座に反映されるようになります。

スクリプトの保存 レポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。

Byグループのスクリプトを保存 By変数の全水準に対するレポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。起動ウィンドウでBy変数を指定した場合のみ使用可能です。

グループ効果を含めたペア比較の例

この例では、「Dogs.jmp」サンプルデータを使用して、「対応のあるペア」でグループを含めた分析と、「MANOVAのあてはめ」（「モデルのあてはめ」を使用）を用いる方法を紹介します。両者の検定結果は、同じです。

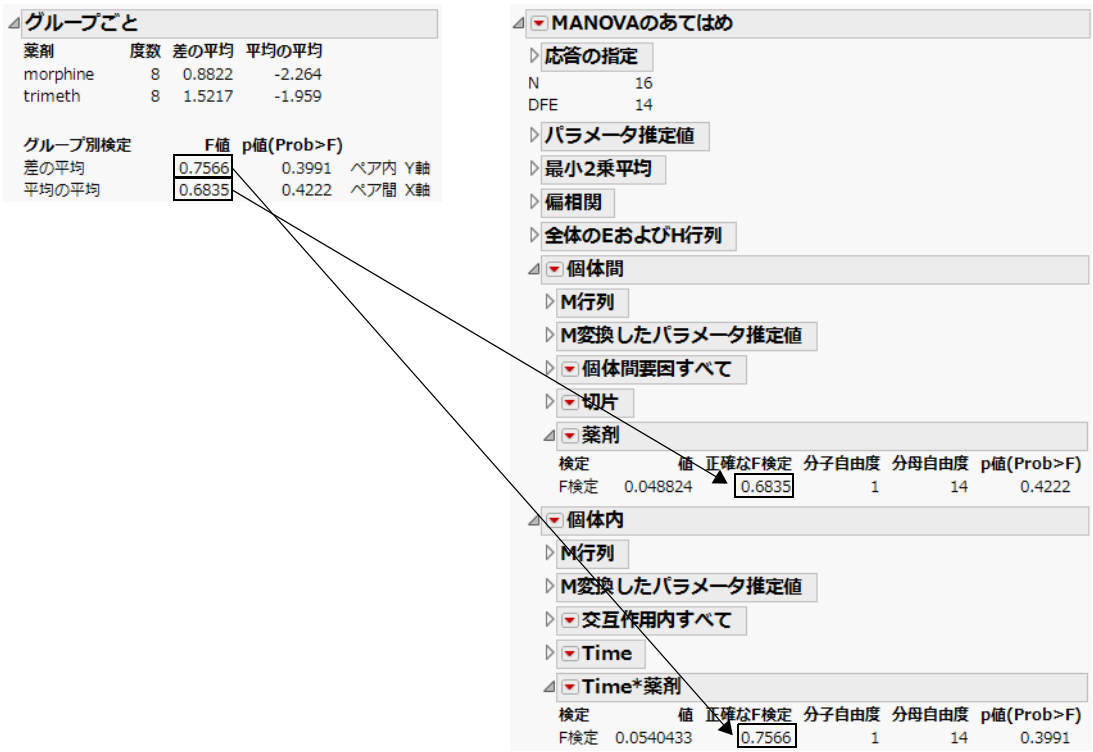
1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Dogs.jmp」を開きます。
2. ［**分析**］>［**発展的なモデル**］>［**対応のあるペア**］を選択します。
3. 「**Log(ヒスタミン0)**」と「**Log(ヒスタミン1)**」を選択し、［**Y, 対応のある応答**］をクリックします。
4. 「**薬剤**」を選択し、［**X, グループ変数**］をクリックします。
5. ［**OK**］をクリックします。

図16.5の左側のレポートが表示されます。

次に、「モデルのあてはめ」で同じデータテーブルを使用してレポートを作成します。

1. ［分析］>［モデルのあてはめ］を選択します。
2. 「Log(ヒスタミン0)」と「Log(ヒスタミン1)」を選択し、［Y］をクリックします。
3. 「薬剤」を選択し、［追加］をクリックします。
4. 「手法」から「MANOVA」を選択します。
5. ［実行］をクリックします。
6. 「応答の指定」レポートで、［応答の選択］メニューから［反復測定］を選択します。
7. ［OK］をクリックします。

図16.5「対応のあるペア」の「グループごと」レポートと「モデルのあてはめ」の反復測定データに対するMANOVAレポートの例



「グループごと」レポートの「差の平均」の「F値」が、「個体内」レポートの「時間*薬剤」の「正確なF検定」の値に対応しています。「グループごと」レポートの「平均の平均」の「F値」が、「個体間」レポートの「薬剤」の「正確なF検定」の値に対応しています。

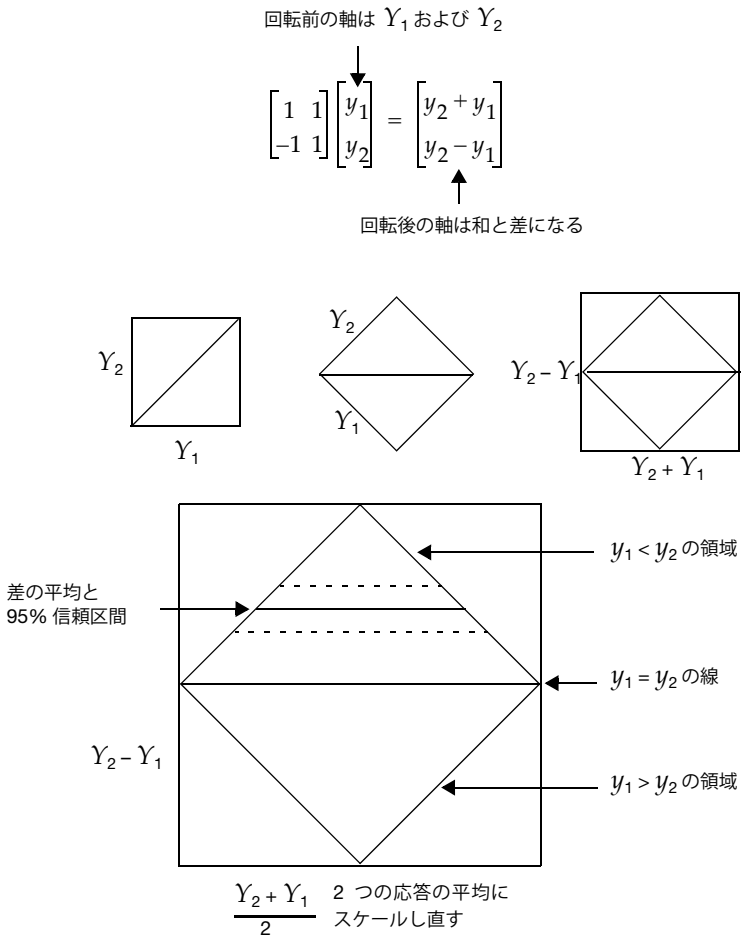
「対応のあるペア」プラットフォームの統計的詳細

ここでは、対応のあるペア分析の統計的詳細について説明します。

対応のあるペアのグラフ

このプラットフォームのグラフは、X軸に2変数の平均、Y軸に2変数の差をプロットしています。「Tukeyの平均-差プロット」（Tukey mean-difference plot）と呼ばれています（Cleveland, 1994）。これは、2変数の散布図を45度回転させたものと同じです。45度の回転と再スケールによって元の座標軸が差と平均に変換されます。

図16.6 45度の回転による差と平均への変換例

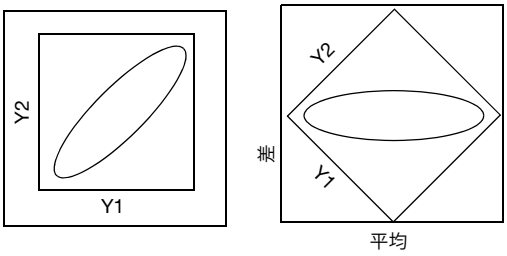


応答の相関

対応のあるデータが同じ個体から複数回にわたって測定されたものである場合は、多くの場合、正の相関があります。ただし、一方が他方を阻害する応答の場合などには、負の相関があるかもしれません。

図16.7は、2つの応答変数に正の相関があると差の分散（Y軸上）が小さくなることを示しています。負の相関があるときは楕円が反対の方向に伸び、回転させたグラフではY軸上での分散が大きくなります。

図16.7 回転前後の正の相関の例



第17章

応答のスクリーニング
大規模データにある多数の応答変数を検定する

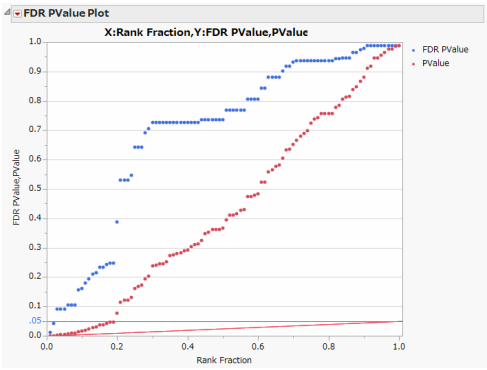
近年になり、1つの工業部品や生物検体について、多数の項目が1度に測定されるようになりました。このような大規模データを分析するには、新しい統計手法が必要です。多数の応答に対して統計的検定を行う場合には、それを考慮した適切な手法が必要です。

「応答のスクリーニング」プラットフォームは、応答変数や説明変数が多数ある場合に、それらに対する検定の処理を一度に行います。検定結果や要約統計量は、データテーブルとしても出力されるため、それらの結果をさらに検討できます。生のp値だけではなく、FDR（False Discovery Rate; 偽発見率）を制御するように調整されたp値も計算されます。FDRを考慮した多重性調整は、本当は差がないのに「差がある」と誤って判断してしまう確率を制御する手法です。なお、FDR調整p値をプロットするときには、解釈を簡単にするため、対数スケールが使われます。

大規模なデータはあまり綺麗ではなく、外れ値や欠測値を含む場合がよくあります。「応答のスクリーニング」では、外れ値や欠測値を処理するオプションがあります。ロバスト推定を使うと、外れ値からあまり影響を受けずに推定が行えます。欠測値に対するオプションを使用すれば、欠測値を計算に含めることができます。こうした機能があるため、データの品質をあまり気にせずに、すぐに分析を行うことができます。

実質的には無意味なほど差が小さくても、標本サイズが大きくなると、統計的有意にはなりやすくなり、差があると判断してしまう可能性が高くなります。「応答のスクリーニング」では、どの程度の大きさがあれば差が実質的に意味があるかを指定し、それに対する検定を行うことができます。また、指定した絶対値を上回る差がないことを、つまり、平均が実質的には等しいことを確認したい場合もあるでしょう。このような同等性検定も、「応答のスクリーニング」では実行できます。

図17.1「応答のスクリーニング」プロットの例



「応答スクリーニング」プラットフォームの概要

「応答のスクリーニング」プラットフォームは、応答変数や説明変数が多数ある場合に、それらに対する検定の処理を一度に行います。各応答変数に対する各説明変数の個別の検定を、一度に行います。このような統計処理には、主に2つの問題があります。1つは、統計的検定を何回も実行しなければならないので、検定の多重性という問題が生じます。もう1つは、外れ値や欠測値を処理しなければいけないということです。「応答のスクリーニング」は、これら2つの問題に対処しています。

「応答のスクリーニング」の機能は、独立したプラットフォームとして、または「モデルのあてはめ」プラットフォームの手法として呼び出すことができます。独立したプラットフォームとしては、表17.1に示すように、「二変量の関係」プラットフォームと同じ検定を実行します。「モデルのあてはめ」の手法としては、線形モデルやロジスティックモデルをあてはめて、各効果に対して検定を行います。

状況に応じた推測に対応するように、「応答のスクリーニング」には次の機能があります。

データテーブル 結果は、グラフだけではなく、データテーブルとしても出力されます。出力されたデータテーブルをもとに、検定結果をさらに検討したり、p値の大きさで並べ替えたり、特定の大きさ以下のp値を抜き出したり、いろいろなグラフを描いたりできます。このデータテーブルには、生のp値や、FDR調整したp値などの統計量が出力されます。

FDR 検定を何度も行うと、まったくの偶然だけで有意となってしまう確率が大きくなります。このため、生のp値を何らかの方法で調整する必要があります。「応答のスクリーニング」では、**FDR**（False Discovery Rate; 偽発見率）を制御するように、多重性の調整を行ったp値を算出します。FDRとは、「有意とされた仮説の中における、実際には有意でない仮説の割合」の期待値です（Benjamini and Hochberg 1995, Westfall et al. 2011）。

実質的な差の検定 実質的には無意味なほど差が小さくても、標本サイズが大きくなると、統計的には有意にはなりやすくなり、「差がある」と判断してしまう可能性が高くなります。この問題に対処するため、差がどれぐらいの大きさあれば**実質的**に意味があるかを定義できます。そして、検定では、そこで指定された大きさ以上の差があると判定されたものだけが検出されます。

実質的な同等性の検定 多数の因子について検討する場合、応答に対する効果が実質的に同等である因子に着目する場合があります。そのような場合には、実質的に同等とみなす差を決めて、同等性の検定を行います。

大規模なデータはあまり綺麗ではなく、外れ値や欠測値を含む場合がよくあります。「応答のスクリーニング」では、外れ値や欠測値を処理するオプションがあります。こうした機能があるため、データの品質をあまり気にせずに、すぐに分析を行うことができます。

ロバスト推定 データに外れ値があると、推定値に対する標準誤差が大きくなり、有意になりにくくなる場合があります。［ロバスト］オプションを選択すると、Huber M推定が実行されます。この推定は、外れ値を手動で除外することなしに、その影響を少なくすることができます。

欠測値のオプション 「応答のスクリーニング」プラットフォームには、カテゴリカルな説明変数における欠測値をカテゴリとして扱うオプションがあります。

表17.1「応答のスクリーニング」で実行される分析

応答	因子	二変量の関係	説明
連続尺度	カテゴリカル	一元配置	分散分析
連続尺度	連続尺度	二変量	単回帰
カテゴリカル	カテゴリカル	分割表	カイ2乗
カテゴリカル	連続尺度	ロジスティック	単純ロジスティック回帰

「応答のスクリーニング」プラットフォームは、レポート（「応答のスクリーニング」レポート）とデータテーブル（「PValues」テーブル）を生成します。「モデルのあてはめ」における［応答のスクリーニング］手法の場合は、レポート（「応答のスクリーニングのあてはめ」レポート）と2種類のデータテーブル（「PValues」と「Y Fits」）が生成されます。

JSLコマンド`Summarize Y by X`は、「応答のスクリーニング」プラットフォームと同じ働きをしますが、プラットフォームウィンドウは生成されません。詳細については、『スクリプト構文リファレンス』の「Summarize YByX」を参照してください。

「応答スクリーニング」の例

「Probe.jmp」サンプルデータテーブルには、5800のウエハについて387の特性を調べたデータが（**「Responses」**列グループに）記録されています。**「ロット ID」**列と**「ウエハ番号」**列は、一意にウエハを識別します。工程の変化（**「工程」列**）の前後で値が違っている特性を調べたいと思います。

1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Probe.jmp」を開きます。
2. ［**分析**］>［**スクリーニング**］>［**応答のスクリーニング**］を選択します。

   「応答のスクリーニング」起動ウィンドウが表示されます。
3. **「Responses」**列グループを選択し、［**Y, 応答変数**］をクリックします。
4. **「工程」**を選択し、［**X**］をクリックします。
5. **「最大対数価値」**ボックスに「100」と入力します。

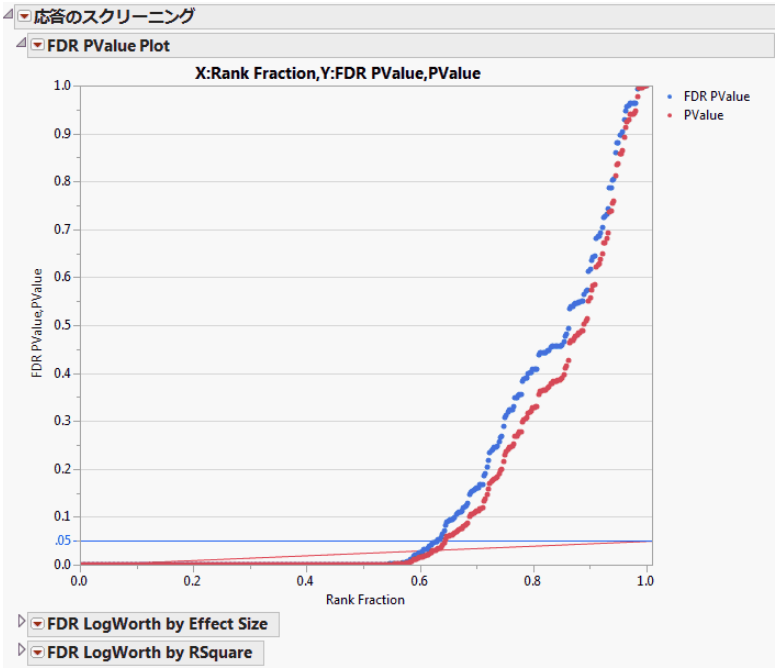
   対数価値が100以上だと、*p*値は極端に小さくなります。「最大対数価値」の値を設定することで、プロットの尺度を調整できます。
6. ［**OK**］をクリックします。

「応答のスクリーニング」レポートと、補足情報のデータテーブルが併せて表示されます。レポート（図17.2）には、「FDR PValue Plot」と、他にも2つプロットレポートがあります。データテーブルには、［**Y, 応答変数**］として指定した387個の列ごとに1行ずつ結果が含まれています。

「FDR PValue Plot」には、387個の各検定について、「FDR Pvalue」（FDR調整した*p*値）と、「PValue」（生の*p*値）がプロットされています。グラフのY軸がこれら2種類のp値を、X軸が「Rank Fraction」（順位の割合）を表しています。「PValue」は、調整を行っていない通常の*p*値です。「FDR PValue」は、偽発見率（false discovery rate）を制御するように調整された*p*値です。「FDR PValue」は青、「PValue」は赤でプロットされます。「Rank Fraction」は、「FDR PValue」を小さい方から順番に（有意性が高い順で）順位付けしています。

グラフにおける青の水平線と赤の右上がりの直線は両方とも、FDR調整したp値に対する5%の閾値を示します。FDR調整*p*値が青い線を下回る場合、FDRを5%に抑えたなかで有意となっています。同じように、通常の*p*値が赤い線を下回る場合、FDRを5%に抑えたなかで有意となります。このように、プロットのどちらの*p*値を見ても、FDR法で有意かどうかを確認できます。

図17.2「工程」に対する387回の検定の「応答のスクリーニング」レポート



「FDR PValue Plot」から、60%以上の検定で有意差が認められていることがわかります。通常の*p*値だと「有意差あり」と判断されるのに、FDR調整した*p*値だと「有意差なし」となる検定は、赤い線より上で、青い線より下にある赤い点が該当します。そのような検定は少ないことが分かります。

「工程」に対して有意差が認められる特性を見極めるためには、プロット上で該当する点の周りをドラッグし、四角く囲みます。これで、これらの点に対応する行が「PValues」テーブルで選択されます。その第1列を見れば、特性の名前がわかります。または、「PValues」テーブルで該当する行を選択することもできます。

「PValues」データテーブル（図17.3）には、「**Responses**」グループの応答ごとに1行ずつ、計387行あります。応答は第1列（「**Y**」）に含まれています。各応答が「X」列（つまり、「**工程**」）の効果に照らして検定されています。

図17.3「PValues」データテーブル（一部）

PValues 2
Original Data　Probe
FDR LogWorth by Effect Size
FDR LogWorth by RSquare
FDR PValue Plot
Fit Selected Items

列(15/0)
Y
X
Count
PValue
LogWorth
FDR PValue
FDR LogWorth
Effect Size
Rank Fraction
YMean
SSE
DFE
MSE
F Ratio
RSquare

行
すべての行　387
選択されている行　0
除外されている行　0
表示しない行　0
ラベルのついた行　0

	Y	X	Count	PValue	LogWorth	FDR PValue	FDR LogWorth
1	DELL_RPNBR	工程	5794	8.044596e-13	12.094495751	2.39467e-12	11.620754237
2	DELL_RPPBR	工程	5794	0.4782555718	0.3203399617	0.5449559037	0.2636386382
3	DELW_M1	工程	5789	0.1806001982	0.7432817773	0.2459236742	0.6091996613
4	DELW_M2	工程	5728	4.19956e-322	100	0	100
5	DELW_NBASE	工程	5778	0.388034756	0.4111293732	0.4570716144	0.3400157489
6	DELW_NEMIT	工程	5788	7.306034e-56	55.136318307	4.065967e-55	54.390836173
7	DELW_NENBNI	工程	5792	4.19956e-322	100	0	100
8	DELW_NSINK	工程	5782	4.19956e-322	100	0	100
9	DELW_PBASE	工程	5786	2.908704e-31	30.536300511	1.227409e-30	29.911010679
10	DELW_PCOLL	工程	5785	4.19956e-322	100	0	100
11	DELW_PEMIT	工程	5789	0.6709186367	0.1733301442	0.7257260746	0.1392272729
12	DELW_PSINK	工程	5789	1.666427e-40	39.778213644	7.709736e-40	39.112960512
13	DELW_RPNBR	工程	5789	9.008231e-24	23.045360484	3.391334e-23	22.469629432
14	DELW_RPPBR	工程	5793	1.169725e-26	25.931916295	4.583412e-26	25.338811147
15	DELW_SICR	工程	5787	0.1180928699	0.9277763229	0.1667193458	0.7780140026
16	M1_COMB_VG…	工程	5800	0.3778451214	0.4226861809	0.4548355067	0.3421456395
17	M1_TRENCH_V…	工程	5800	0.0002786205	3.5549868893	0.0005193703	3.2845228853
18	M2/M1_CAP_V…	工程	5800	0.6294372314	0.201047572	0.6866587979	0.1632590111
19	M2_COMB_BB…	工程	5800	0.9988767052	0.0004881149	0.999486039	0.0002232678
20	M2_COMB_VG…	工程	5800	0.2990592705	0.5242427304	0.3827958663	0.4170327607
21	NISO_TUB-TR…	工程	5800	1.58784e-164	100	1.38575e-163	100
22	NISO_TUB-TU…	工程	5800	0.4892691678	0.3104521513	0.5509658664	0.2588753059
23	PS_RPNBR	工程	5790	0.0000237641	4.6240777354	0.0000493267	4.3069182395
24	RCON_NEM	工程	5795	0.3685513548	0.4335019879	0.4464470671	0.3502300258
25	RCON_PEM	工程	5794	1.095985e-24	23.96019535	4.208583e-24	23.375864125

残りの列には、「**Y**」と「**X**」の検定に関する情報が記録されています。この例で使われている検定は、一元配置の分散分析です。このデータテーブルには他にも、*p*値、LogWorth（*p*値の対数価値）、FDR調整した*p*値、FDR LogWorth（FDR調整*p*値の対数価値）が含まれています。このデータテーブルにおいて、各種の統計量を並べ替えたり、行を選択したり、目的の統計量のグラフを作成したりすることができます。

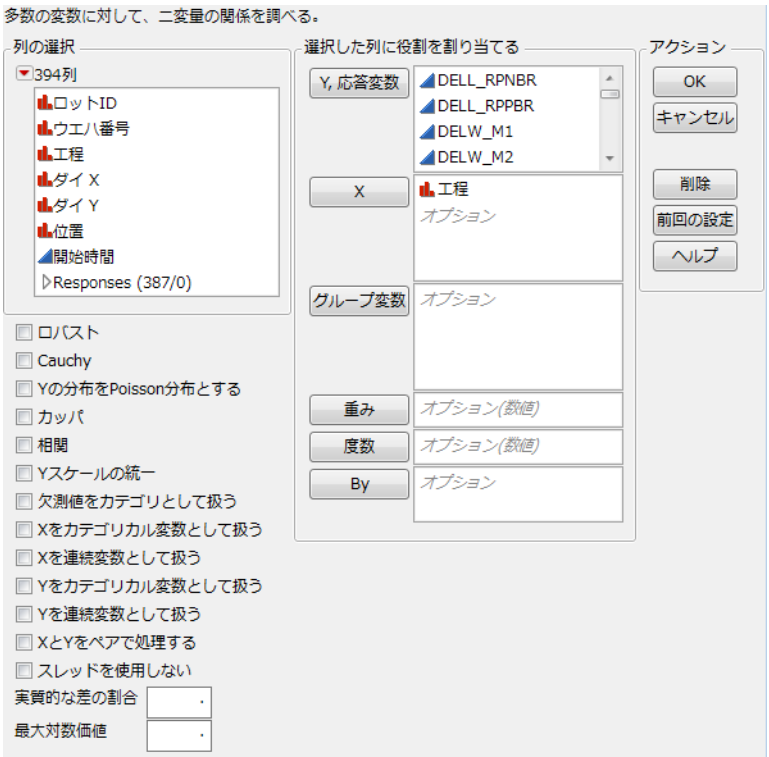
起動ウィンドウで「最大対数価値」を100に設定したため、LogWorthとFDR LogWorthの値のうち、*p*値が1e-100以下である値は「100」として表示されています。また、FDR LogWorthの値が2を超えるセルは、背景色に色が付けられます。

レポートと「PValues」テーブルの詳細については、「「応答のスクリーニング」レポート」（288ページ）を参照してください。

「応答のスクリーニング」プラットフォームの起動

「応答のスクリーニング」プラットフォームを起動するには、[**分析**] > [**スクリーニング**] > [**応答のスクリーニング**] を選択します。

図17.4「応答のスクリーニング」起動ウィンドウ



起動ウィンドウの役割

Y, 目的変数　測定値が含まれている、応答変数の列を指定します。

X　応答変数との関係を調べたい、説明変数の列を指定します。

グループ変数　ここで指定した列の水準ごとに、該当する行が個別に分析されますが、結果は1つのテーブルとレポートに表示されます。

重み　この役割を割り当てた列の値は、分析において各行に対する重みとして使用されます。詳細については、『基本的な回帰モデル』の「モデルの指定」章、「重み」の節を参照してください。

度数　この役割を割り当てた列の値は、各行の度数として使用されます。これにより、予め集計されたデータを扱うことができます。詳細については、『基本的な回帰モデル』の「モデルの指定」章、「度数」の節を参照してください。

By　ここで指定した列の水準ごとに、対応するYとXが分析され、結果が個別のテーブルとレポートに表示されます。

起動ウィンドウのオプション

ロバスト 応答が連続尺度である場合、Huberのロバスト法による推定を行います。この推定方法は、外れ値による影響が少ないです。外れ値がない場合、HuberのM推定の結果は、最小2乗推定のものと近くなります。このオプションを選択した場合は、計算時間がかかります。

Cauchy このオプションでは、誤差がCauchy分布に従うと仮定されます。Cauchy分布は正規分布よりも裾が広く、その結果、外れ値が推定に与える影響が小さくなります。このオプションは、データにある外れ値の割合が大きい場合に有用です。しかし、データが正規分布に近く、外れ値が少ない場合は、このオプションの推定結果は間違ったものになる可能性があります。[Cauchyのあてはめ]オプションは、最尤推定によってパラメータ推定値を算出します。

Yの分布をPoisson分布とする 各応答変数（Y）を、Poisson分布に従う度数としてあてはめます。検定は、カテゴリカルなXに対してのみ実行されます。このオプションは、応答変数が度数である場合に適しています。

カッパ YとXが両方ともカテゴリカルで水準が同じ場合、カッパ統計量が求められます。「Kappa」という新しい列がデータテーブルに追加されます。カッパ統計量は、YとXの一致度を表す指標です。

相関 XとYが両方ともカテゴリカルであっても、相関を計算し、「Corr」という新しい列をデータテーブルに追加します。

このオプションをオンにすると、カテゴリ値を順番に並べて通し番号を付けて、その通し番号からPearson の積率相関係数を求めますが、このように計算された相関は次のようなデータに対するSpearmanの順位相関係数になっています。

– XとYが順序尺度のデータ

– XとYが名義尺度で、値が自然な順序通りに定義されているデータ

もし、XとYが両方とも二値の変数である場合は、このように順位から計算されたPearsonの相関係数は、Kendallのτbにも一致します。相関の絶対値が大きければ、関係があることを示唆しています。逆に相関の絶対値が小さければ、関係が薄いことを示唆しています。

Yスケールの統一 レポートの[選択した項目の二変量関係]オプションを使用して個々の分析を実行する際に、すべての応答変数（Y）のスケールを統一します。

欠測値をカテゴリとして扱う カテゴリカルなX変数について、Xの欠測値をカテゴリとして扱います。

Xをカテゴリカル変数として扱う 列に設定されている尺度を無視して、すべてのX列を名義尺度として扱います。

Xを連続変数として扱う 列に設定されている尺度を無視して、すべてのX列を連続尺度として扱います。

Yをカテゴリカル変数として扱う 列に設定されている尺度を無視して、すべてのY列を名義尺度として扱います。

Yを連続変数として扱う 列に設定されている尺度を無視して、すべてのY列を連続尺度として扱います。

XとYをペアで処理する ［Y, 応答変数］と［X］のリストでの順序に従ってY列とX列をペアにして、それらのペアだけに検定を行います。1番目のY変数と1番目のX変数とをペアにして、2番目のY変数と2番目のX変数とをペアにして、といった具合で組み合わせます。

スレッドを使用しない マルチスレッド処理を行いません。

実質的な差の割合 実質的に意味のある差を、仕様限界の範囲に対する割合として指定します。［仕様限界］が列プロパティとして指定されていない場合は、応答の標準偏差を6倍したものが、仕様限界の範囲として設定されます。このとき、標準偏差の推定値は、四分位範囲（IQR）から、$\hat{\sigma} = (IQR)/(1.3489795)$ という式によって求められます。

「実質的な差の割合」が未指定の場合は、デフォルト値の0.10が使用されます。実質的有意差の検定と同等性検定では、ここで指定された値を基に、実質的な差を判断します。「平均の比較のデータテーブル」（296ページ）を参照してください。

最大対数価値 対数価値（p 値の $-\log_{10}$）を示すプロットの尺度の調整に使われます。対数価値プロットの尺度が極端なものにならないよう、ここで指定した最大対数価値を超える対数価値は、この値に変更してプロットに示されます。例については、「「最大対数価値」オプションの例」（304ページ）を参照してください。

OK 分析を実行し、結果を表示します。

キャンセル 起動ウィンドウを閉じます。

削除 役割から変数を削除します。

前回の設定 前回実行したモデルの指定内容を起動ウィンドウに読み込みます。

ヘルプ 「応答のスクリーニング」起動ウィンドウに関するヘルプトピックを開きます。

「応答のスクリーニング」レポート

「応答のスクリーニング」レポートには、グラフビルダーによって描かれたグラフがいくつか表示されます。これらはFDR調整したp値のグラフです。詳細は、「FDR（False Discovery Rate; 偽発見率）」（311ページ）を参照してください。

デフォルトで表示されるグラフは、「FDR PValue Plot」（FDRのプロット）、「FDR LogWorth by Effect Size」（FDR対数価値と効果の大きさのプロット）、「FDR LogWorth by RSquare」（FDR対数価値とR2乗値のプロット）です。起動ウィンドウで［ロバスト］オプションを選択した場合は、これらのグラフのほかに、ロバスト推定による同様のプロットも併せて表示されます。さらに、「Robust LogWorth by LogWorth」（ロバスト対数価値と、通常の対数価値）のプロットも描かれ、ロバスト推定による影響を評価できます。各プロットの赤い三角ボタンのメニューから、グラフビルダーの標準オプションを選択できます。詳細については、『グラフ機能』の「グラフビルダー」章を参照してください。

FDR PValue Plot

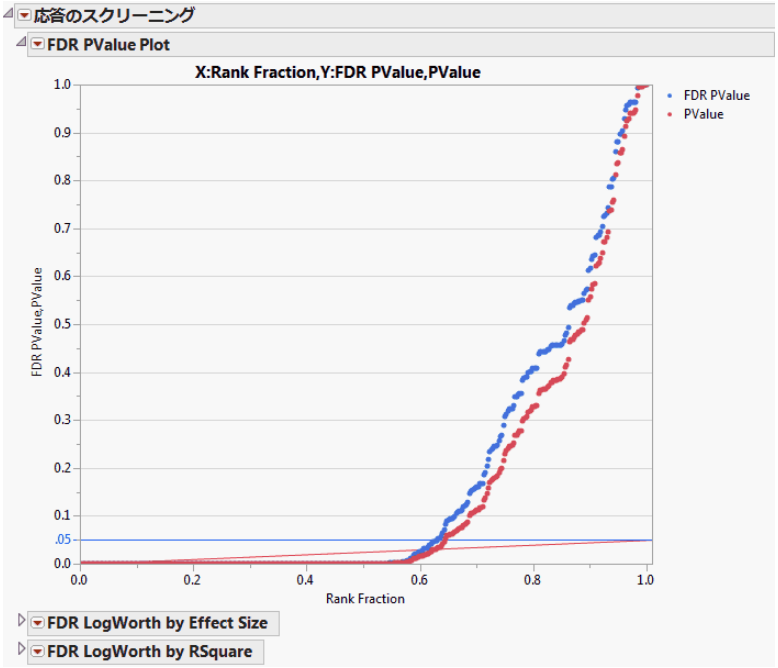
「FDR PValue Plot」レポートのプロットは、Y軸が「FDR PValues」と「PValues」、X軸が「Rank Fraction」です。「Rank Fraction」は、p値を有意差の大きい順で順位付けしています。「FDR PValues」は青、「PValues」は赤でプロットされます。

青い水平線は、5%の有意水準を示しています（この水平線は、Y軸における参照線として設定されていて、任意の位置に変更することができます）。

赤い右上がりの直線は、多重性の調整を行っていないp値がFDR法で有意となる閾値を示します。FDR調整p値が青い線を下回る場合、未調整のp値は赤い線を下回ります。このため、調整前と調整後のいずれのp値を見ても、FDR法で有意となっている検定を読み取ることができます。

図17.5は、「Probe.jmp」サンプルデータテーブルの「FDR PValue Plot」です。一部の検定は、通常のp値だと「有意差あり」ですが、FDR法では「有意差なし」と判断されます。

図17.5 FDR PValue Plot



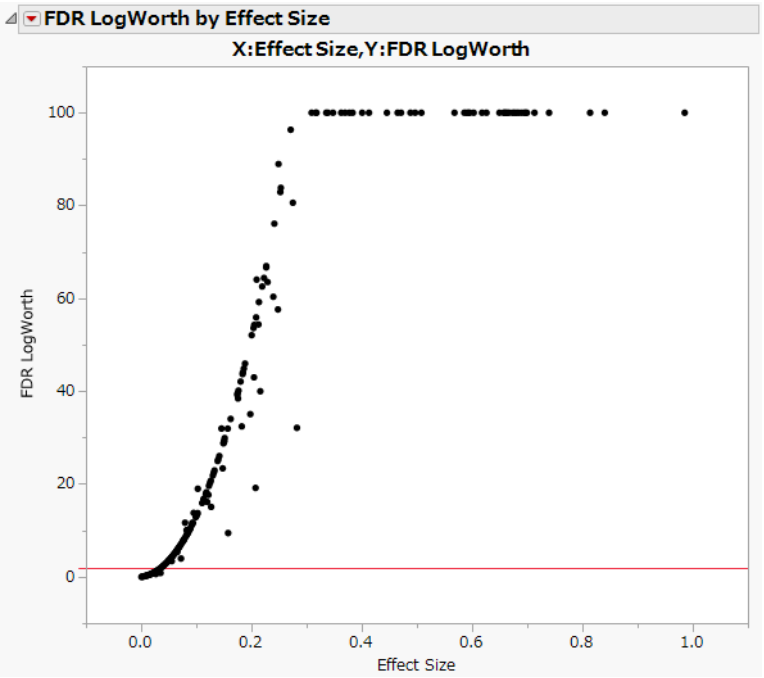
FDR LogWorth by Effect Size

効果量が非常に大きいと、*p*値はとても小さくなる場合があります。とても小さいp値をグラフにプロットすると、どれぐらいの大きさなのかがわかりづらくなります。対数価値(-$\log_{10}$(*p*値))というスケールに変換すると、とても小さい*p*値も見やすくなります。*p*値が小さい場合には対数価値は大きくなり、*p*値が大きい場合には対数価値は小さくなります。対数価値がゼロの場合、*p*値は1です。対数価値が2以上の場合、*p*値は0.01以下です。

「FDR LogWorth by Effect Size」プロットでは、縦軸が「FDR LogWorth」(FDR対数価値)、横軸が「Effect Size」(効果の大きさ、効果量)です。一般に、効果の大きさが大きいほど、*p*値が小さくなり、対数価値が大きくなります。しかし、*p*値の小ささは誤差分散にも依存するため、必ずしもそのような関係になるわけではありません。実際、誤差分散によっては、効果の大きさが大きくても対数価値が小さくなるときもあれば、逆に、効果の大きさが小さくても対数価値が大きくなることもあります。「FDR LogWorth by Effect Size」プロットでは、そのような関係を検討できます。

図17.6は、「Probe.jmp」データテーブルで「最大対数価値」を100に設定したときの「FDR LogWorth by Effect Size」プロットです。「FDR LogWorth」の大半は2を上回っていて、つまり、ほとんどの効果が有意水準1%で有意となっています。「FDR LogWorth」が100の位置にプロットされている点は、FDR調整した*p*値が非常に小さいものです。

図17.6 FDR LogWorth by Effect Size



FDR LogWorth by RSquare

「FDR LogWorth by RSquare」プロットでは、縦軸が「FDR LogWorth」（FDR 対数価値）で、横軸が「RSquare」（R2 乗値）です。R2 乗値が大きいほど、対数価値は大きくなります。ただし、対数価値の大きさは、標本サイズにも依存します。

「PValues」データテーブル

「PValues」データテーブルには、Y 変数と X 変数の組み合わせごとに、検定結果が 1 行ずつ含まれています。［グループ変数］列を指定した場合、「PValues」データテーブルの第 1 列は、［グループ変数］に指定した列になります。［グループ変数］列の水準、Y 変数、X 変数の組み合わせごとに、検定結果が出力されます。「PValues」データテーブルには、分析に使用したデータテーブルの名前を示す「Original Data」というテーブル変数も含まれます。By 変数を指定した場合は、By 変数の水準ごとに「PValues」テーブルが作成され、「Original Data」変数に By 変数とその水準が表示されます。

図 17.7 は、「「応答スクリーニング」の例」（283 ページ）で作成された「PValues」データテーブルです。

図 17.7「PValues」データテーブル（一部）

PValues 2
Original Data　Probe
FDR LogWorth by Effect Size
FDR LogWorth by RSquare
FDR PValue Plot
Fit Selected Items

列(15/0)
Y
X
Count
PValue
LogWorth
FDR PValue
FDR LogWorth
Effect Size
Rank Fraction
YMean
SSE
DFE
MSE
F Ratio
RSquare

行
すべての行 387
選択されている行 0
除外されている行 0
表示しない行 0
ラベルのついた行 0

	Y	X	Count	PValue	LogWorth	FDR PValue	FDR LogWorth
1	DELL_RPNBR	工程	5794	8.044596e-13	12.094495751	2.39467e-12	11.620754237
2	DELL_RPPBR	工程	5794	0.4782555718	0.3203399617	0.5449559037	0.2636386382
3	DELW_M1	工程	5789	0.1806001982	0.7432817773	0.2459236742	0.6091996613
4	DELW_M2	工程	5728	4.19956e-322	100	0	100
5	DELW_NBASE	工程	5778	0.388034756	0.4111293732	0.4570716144	0.3400157489
6	DELW_NEMIT	工程	5788	7.306034e-56	55.136318307	4.065967e-55	54.390836173
7	DELW_NENBNI	工程	5792	4.19956e-322	100	0	100
8	DELW_NSINK	工程	5782	4.19956e-322	100	0	100
9	DELW_PBASE	工程	5786	2.908704e-31	30.536300511	1.227409e-30	29.911010679
10	DELW_PCOLL	工程	5785	4.19956e-322	100	0	100
11	DELW_PEMIT	工程	5789	0.6709186367	0.1733301442	0.7257260746	0.1392272729
12	DELW_PSINK	工程	5789	1.666427e-40	39.778213644	7.709736e-40	39.112960512
13	DELW_RPNBR	工程	5789	9.008231e-24	23.045360484	3.391334e-23	22.469629432
14	DELW_RPPBR	工程	5793	1.169725e-26	25.931916295	4.583412e-26	25.338811147
15	DELW_SICR	工程	5787	0.1180928699	0.9277763229	0.1667193458	0.7780140026
16	M1_COMB_VG…	工程	5800	0.3778451214	0.4226861809	0.4548355067	0.3421456395
17	M1_TRENCH_V…	工程	5800	0.0002786205	3.5549868893	0.0005193703	3.2845228853
18	M2/M1_CAP_V…	工程	5800	0.6294372314	0.201047572	0.6866587979	0.1632590111
19	M2_COMB_BB…	工程	5800	0.9988767052	0.0004881149	0.999486039	0.0002232678
20	M2_COMB_VG…	工程	5800	0.2990592705	0.5242427304	0.3827958663	0.4170327607
21	NISO_TUB-TR…	工程	5800	1.58784e-164	100	1.38575e-163	100
22	NISO_TUB-TU…	工程	5800	0.4892691678	0.3104521513	0.5509658664	0.2588753059
23	PS_RPNBR	工程	5790	0.0000237641	4.6240777354	0.0000493267	4.3069182395
24	RCON_NEM	工程	5795	0.3685513548	0.4335019879	0.4464470671	0.3502300258
25	RCON_PEM	工程	5794	1.095985e-24	23.96019535	4.208583e-24	23.375864125

「PValues」データテーブルの列

「PValues」データテーブルには、選択した統計手法や、Y 変数と X 変数の尺度の組み合わせに応じて、いくつかの統計量が出力されます。このデータテーブルには、以下のような列があります。

Y　指定した応答列。

X 指定した因子列。

度数 検定に使用された行の数、または、［度数］変数または［重み］変数の和。

PValue X変数とY変数の関係に対する有意性検定の*p*値。「二変量の関係」の統計量の詳細については、『基本的な統計分析』の「「二変量の関係」プラットフォームの概要」章を参照してください。

LogWorth $-\log_{10}(p\text{値})$。*p*値のグラフのスケールが対数価値に変換されると、解釈がしやすくなります。2を上回る値は、有意水準0.01で有意となります（$-\log_{10}(0.01) = 2$）。

FDR PValue FDR（False Discovery Rate; 偽発見率）を制御するように調整された*p*値。Benjamini-Hochberg法で計算されています。FDRは、検定の多重性を考慮して、生の*p*値を調整したものです。［グループ変数］が指定されていない場合は、データテーブルに表示されるすべての検定を考慮して、多重性の調整が行われます。［グループ変数］が指定されている場合は、グループ変数の水準ごとに多重性の調整が行われます。FDRについては、Benjamini and Hochberg（1995）を参照してください。また、「FDR（False Discovery Rate; 偽発見率）」（311ページ）を参照してください。

FDR LogWorth $-\log_{10}(\text{FDR調整p値})$。これは、検定の有意性をグラフに表すのに適している統計量です。*p*値が小さいと、この値は大きくなります。FDR LogWorthの値が2より大きい（*p*値が0.01より小さい）セルは、背景色に色が付けられます。

Effect Size Xの水準や値によって応答の値がどの程度異なるかを示します。Effect Size（効果の大きさ、効果量）は、次のように計算される、尺度不変な指標です。

- Yが連続尺度の場合、効果の大きさは、モデル平均平方和の平方根を、応答の標準偏差の推定値で割った値です。応答の標準偏差の推定値は、四分位範囲（IQR）がゼロでない場合は、$IQR/1.3489795$で求められます。この方法は、外れ値に対してロバストです。IQRがゼロの場合、標本の標準偏差が使用されます。
- YがカテゴリカルでXが連続尺度の場合、効果の大きさは、モデル全体のカイ2乗検定統計量を標本サイズで割ったものの平方根です。
- YとXが両方ともカテゴリカルな場合、効果の大きさは、Pearsonのカイ2乗値を標本サイズで割ったものの平方根です。

Rank Fraction 対数価値の順位を、検定の総回数で割ったもの。検定の総回数を*m*とした場合、対数価値が最大のときに、「Rank Fraction」は$1/m$となります。また、対数価値が最小のときに、「Rank Fraction」は1となります。この「RankFraction」は、対数価値では大きい順ですが、*p*値では小さい順に対応しています。「Rank Fraction」は、「FDR PValue Plot」の横軸に使われます。

YMean Yの平均。

SSE Yが連続変数の場合に表示されます。誤差の平方和。

DFE Yが連続変数の場合に表示されます。誤差の自由度。

MSE Yが連続変数の場合に表示されます。誤差の平均平方。

F値 Yが連続変数の場合に表示されます。分散分析または回帰分析のF値。

R2乗　Yが連続変数の場合に表示されます。決定係数。モデルで説明される変動の割合を表します。

自由度　YとXが両方ともカテゴリカルな場合に表示されます。カイ2乗検定の自由度。

LR Chisq　YとXが両方ともカテゴリカルな場合に表示されます。尤度比カイ2乗統計量。

［ロバスト］オプションを選択した場合に追加される列

連続尺度の応答に外れ値があることが疑われる場合、起動ウィンドウで［ロバスト］オプションを選択すると、外れ値による影響を減らすことができます。このオプションを選択した場合は、HuberのM推定（Huber and Ronchetti, 2009）によって回帰モデルや分散分析モデルが推定されます。HuberのM推定による推定結果は、外れ値がない場合は最小2乗法によるものとかなり近い値になります。外れ値がある場合は、外れ値に小さい重みを与えることにより、その影響を減らします。

起動ウィンドウで［ロバスト］オプションを選択した場合は、次の列が「PValues」データテーブルに追加されます。［ロバスト］オプションはYが連続変数の場合にのみ適用されるため、Yがカテゴリカルな場合、「Robust」列のセルは空になります。HuberのM推定に関するその他の詳細については、『基本的な統計分析』の「二変量分析」章を参照してください。例として、「ロバストなあてはめの例」（306ページ）を参照してください。

Robust PValue　X変数とY変数の関係に対する有意性検定の*p*値。ただし、ロバストな推定で求められている。

Robust LogWorth　$-\log_{10}$(ロバスト推定によるp値)。

Robust FDR PValue　ロバスト推定によるp値を、Benjamini-Hochberg法によって調整したもの。［グループ変数］が指定されていない場合は、データテーブルに表示されるすべての検定を考慮して、多重性の調整が行われます。［グループ変数］が指定されている場合は、グループ変数の水準ごとに多重性の調整が行われます。

Robust FDR LogWorth　$-\log_{10}$(ロバスト推定によるFDR調整p値)。

Robust Rank Fraction　「Robust FDR LogWorth」の順位を、検定の総回数で割ったもの。

Robust Chisq　ロバスト推定に基づく検定のカイ2乗値。

Robust Sigma　誤差の標準偏差に対するロバストな推定値。

Robust Outlier Portion　ロバスト平均からの距離が「Robust Sigma」の3倍を上回る値の割合。

Robust CpuTime　ロバスト推定の処理にかかった時間（秒数）。

「PValues」データテーブルのスクリプト

「PValues」データテーブルには、スクリプトが保存されています。これらのスクリプトは、1つを除いて、レポートに表示されているグラフを再現するものです。「Fit Selected Items」スクリプトを、「PValues」データテーブルで行を選択した後に実行すると、選択された項目に対して「二変量の関係」の分析が行われます。

「応答のスクリーニング」プラットフォームのオプション

「応答のスクリーニング」の赤い三角ボタンのメニューには、表示形式をカスタマイズするオプションや、結果をデータテーブルに保存するオプションがあります。

選択した項目の二変量関係　選択した項目の関係について、適切な「二変量の関係」のレポートを、「応答のスクリーニング」レポートに追加します。「PValues」データテーブルで行を選択するか、グラフで点を選択すれば、対象とする項目を選択できます。

列の選択　「PValues」データテーブルで選択した行、または「応答のスクリーニング」レポートウィンドウのプロットで選択した点に対応する、元のデータテーブルの列を選択します。行または点を選択してから、［列の選択］を選択してください。データテーブル内の対応する列が選択されます。選択列を追加するには、再び「PValues」データテーブルの行またはプロット上の点を選択し、［列の選択］を選択します。別の行や点に対応する列を選択するには、まず、元のデータテーブルで列の選択を解除します。

平均の保存　連続尺度のYとカテゴリカルなXについて、カテゴリカルな変数の水準ごとに度数、平均、標準偏差を記録したデータテーブルを作成します。［ロバスト］オプションを選択した場合は、ロバスト推定による平均が含まれます。

平均の比較を保存　連続尺度のYとカテゴリカルなXについて、カテゴリカル変数の水準のすべてのペアごとに比較検定を行います。通常のt検定、実質的な差に対する検定、実質的な同等性に対する検定の結果を、比較のペアごとに含んだデータテーブルを新たに作成します。これらの列は、検定結果に基づいて色分けされます。このデータテーブルには、実質的な差に対する観測された差の比と、その対数価値との関係をプロットする「Practical LogWorth by Relative Practical Difference」というスクリプトも含まれています。「平均の比較のデータテーブル」（296ページ）を参照してください。例として、「実質的な差や実質的な同等性に対する検定の例」（302ページ）を参照してください。

標準化残差の保存　連続尺度のYとカテゴリカルなXに対して、元のデータテーブルに標準化残差を含んだ列を新たに作成します。作成された列は、「Residual Group」という名前の列グループにまとめられます。標準化残差とは、残差を、その標準偏差の推定値で割ったものです。これらの列は、計算式により定義されます。

［ロバスト］オプションが選択されている場合、標準化残差は、ロバストな推定値によって計算されます。

外れ値を示す指示変数の保存　連続尺度のYとカテゴリカルなXに対して、元のデータテーブルに、外れ値を示すフラグを含んだ列を新たに作成します。作成された列は、「Outlier Group」という名前の列グループにまとめられます。この［外れ値を示す指示変数の保存］は、［ロバスト］オプションが選択されている場合に、より役立ちます。

連続尺度のYとカテゴリカルなXごとに、外れ値を示す列が作成されます。平均から3シグマ以上離れた点が外れ値と判断されます。これは、標準化残差の絶対値が3を超える点です。これらの列は、計算式により定義されます。

［ロバスト］オプションが選択されている場合、ロバストな推定値が使用されます。つまり、平均や標準偏差がロバストな方法で計算されます。平均から3シグマ以上離れた点が外れ値と判断されます。

「Cluster Outliers」というスクリプトが元のデータテーブルに追加されます。このスクリプトは、外れ値を示す指示変数に対して階層型クラスター分析を行います。

以下のオプションについて詳しくは、『JMPの使用法』の「JMPのレポート」章を参照してください。

ローカルデータフィルタ　現在のレポートにおいてデータをフィルタリングするローカルデータフィルタを表示するか、非表示にします。

やり直し　分析を繰り返すオプションや、やり直すオプションを含みます。また、［自動再計算］オプションを選択すると、このオプションに対応しているプラットフォームにおいて、データテーブルに加えた変更が、該当するレポートに即座に反映されるようになります。

スクリプトの保存　レポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。

Byグループのスクリプトを保存　By変数の全水準に対するレポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。起動ウィンドウでBy変数を指定した場合のみ使用可能です。

平均のデータテーブル

平均のデータテーブルには、X変数の水準ごとに、各応答の平均と標準偏差が出力されます。「Probe.jmp」サンプルデータの場合、387変数の応答が、「工程」という2水準因子に対して検定されます。この場合、平均のデータテーブルには387×2＝774の行が含まれます（図17.8）。

図17.8 平均のデータテーブル

無題9

列(6/0)
- Y
- X
- Level
- Count
- Mean
- StdDev

行

すべての行	774
選択されている行	0
除外されている行	0
表示しない行	0
ラベルのついた行	0

	Y	X	Level	Count	Mean	StdDev
1	DELL_RPNBR	Process	New	3044	0.2035840106	0.1916031302
2	DELL_RPNBR	Process	Old	2750	0.2346329436	0.1278299414
3	DELL_RPPBR	Process	New	3044	-0.068072506	0.1784161107
4	DELL_RPPBR	Process	Old	2750	-0.043321135	1.9165048167
5	DELW_M1	Process	New	3039	-0.04387197	1.0774346893
6	DELW_M1	Process	Old	2750	-0.071400861	0.0389764703
7	DELW_M2	Process	New	3039	0.8014275806	0.0698264259
8	DELW_M2	Process	Old	2689	0.7274958056	0.0649383327
9	DELW_NBASE	Process	New	3028	0.1957655669	77.801881218
10	DELW_NBASE	Process	Old	2750	1.4765185729	0.1326787074
11	DELW_NEMIT	Process	New	3038	0.3411755212	0.0940762694
12	DELW_NEMIT	Process	Old	2750	0.3813085954	0.0976561441
13	DELW_NENBNI	Process	New	3043	3.7172025841	1.1419433505
14	DELW_NENBNI	Process	Old	2749	4.6556238745	0.1893626325
15	DELW_NSINK	Process	New	3032	9.2876152816	1.2840573691
16	DELW_NSINK	Process	Old	2750	7.8902297629	0.3781489659
17	DELW_PBASE	Process	New	3036	1.1651478779	0.1351193171
18	DELW_PBASE	Process	Old	2750	1.2087969753	0.1486923744

平均のデータテーブルには次の列があります。

Y 連続尺度の応答変数。

X カテゴリカルな説明変数。

水準 カテゴリカルなX変数の水準。

度数 該当する水準における値の個数。

平均 該当する水準に対するY変数の平均。

StdDev 該当する水準に対するY変数の標準偏差。

ロバスト 平均 HuberのM推定による、平均のロバストな推定値。起動ウィンドウで［ロバスト］オプションを選択した場合に表示されます。

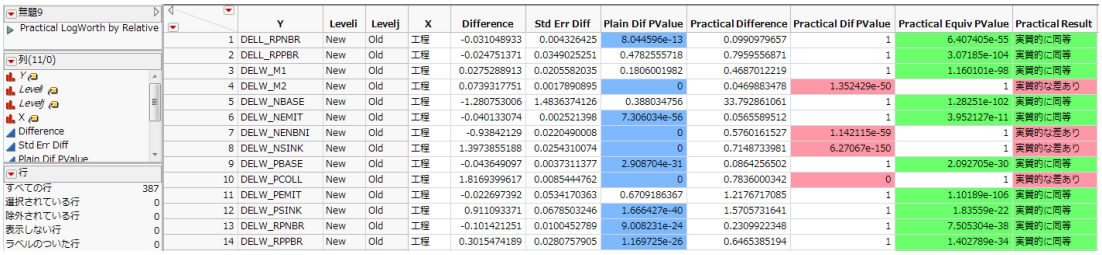
平均の比較のデータテーブル

データの行が増えると（*n*が大きくなると）、検定の計算で使われる標準誤差は小さくなっていきます。その結果、統計的には有意であっても、実際に観測された差がとても小さくて、実質的には意味がない場合があります。そこで、「応答のスクリーニング」では、「差がある」と判断する差の大きさを指定して、それに基づく検定を行えます。「実質的に意味がある差」は、**実質的な差**（practical difference）と呼ばれます。この検定では、差がゼロかどうかではなく、実質的な大きさを差が上回っているかどうかが検定されます。そのため、実質的に意味がある結論だけを取り出し、無意味な結果を調べる手間が省けます。

同等性の検定は、応答平均が実質的に同等かどうかを判断します。この検定では、「実質的な差がある」ということを帰無仮説とします。

平均の比較のデータテーブルには、実質的な差に対する検定と、実質的な同等性に対する検定の結果が含まれています。各行の結果は、カテゴリカルな因子の水準の各ペアにおいて応答を比較したものです。結果は色分けされており、有意性が分かりやすくなっています。実質的な差の指定方法については、「Practical Difference」（297ページ）を参照してください。例については、「実質的な差や実質的な同等性に対する検定の例」（302ページ）を参照してください。

図17.9 平均の比較のデータテーブル



無題9
Practical LogWorth by Relative
列(11/0)
Y
Level
Level
X
Difference
Std Err Diff
Plain Dif PValue
行
すべての行 387
選択されている行 0
除外されている行 0
表示しない行 0
ラベルのついた行 0

	Y	Leveli	Levelj	X	Difference	Std Err Diff	Plain Dif PValue	Practical Difference	Practical Dif PValue	Practical Equiv PValue	Practical Result
1	DELL_RPNBR	New	Old	工程	-0.031048933	0.004326425	8.044596e-13	0.0990979657	1	6.407405e-55	実質的に同等
2	DELL_RPPBR	New	Old	工程	-0.024751371	0.0349025251	0.4782555718	0.7959556871	1	3.07185e-104	実質的に同等
3	DELW_M1	New	Old	工程	0.0275288913	0.0205582035	0.1806001982	0.4687012219	1	1.160101e-98	実質的に同等
4	DELW_M2	New	Old	工程	0.0739317751	0.0017890895	0	0.0469883478	1.352429e-50	1	実質的な差あり
5	DELW_NBASE	New	Old	工程	-1.280753006	1.4836374126	0.388034756	33.792861061	1	1.28251e-102	実質的に同等
6	DELW_NEMIT	New	Old	工程	-0.040133074	0.002521398	7.306034e-56	0.0565589512	1	3.952127e-11	実質的に同等
7	DELW_NENBNI	New	Old	工程	-0.93842129	0.0220490008	0	0.5760161527	1.142115e-59	1	実質的な差あり
8	DELW_NSINK	New	Old	工程	1.3973855188	0.0254310074	0	0.7148733981	6.27067e-150	1	実質的な差あり
9	DELW_PBASE	New	Old	工程	-0.043649097	0.0037311377	2.908704e-31	0.0864256502	1	2.092705e-30	実質的に同等
10	DELW_PCOLL	New	Old	工程	1.8169399617	0.0085444762	0	0.7836000342	0	1	実質的な差あり
11	DELW_PEMIT	New	Old	工程	-0.022697392	0.0534170363	0.6709186367	1.2176717085	1	1.10189e-106	実質的に同等
12	DELW_PSINK	New	Old	工程	0.911093371	0.0678503246	1.666427e-40	1.5705731641	1	1.83559e-22	実質的に同等
13	DELW_RPNBR	New	Old	工程	-0.101421251	0.0100452789	9.008231e-24	0.2309922348	1	7.505304e-38	実質的に同等
14	DELW_RPPBR	New	Old	工程	0.3015474189	0.0280757905	1.169725e-26	0.6465385194	1	1.402789e-34	実質的に同等

平均の比較のデータテーブルには、相対的な実質的差と、その対数価値との関係をプロットする「PracticalLogWorth by Relative Practical Difference」というスクリプトもあります。この相対的な実質的差（relative practical difference）は、観測された差を、実質的な差で割った値として定義されています。

Y 連続尺度の応答変数。

X カテゴリカルな説明変数。

Leveli カテゴリカルなX変数の水準。

Levelj カテゴリカルなX変数のLeveliと比較されるもう1つの水準。

Difference 平均の差に対する推定値。［ロバスト］オプションが選択されている場合、平均のロバスト推定値が使用されます。

Std Err Diff 平均の差に対する標準誤差。［ロバスト］オプションが選択されている場合は、ロバストな推定値が使用されます。

Plain Dif PValue ペアごとの比較を行う、Studentのt検定における通常のp値。［ロバスト］オプションが選択されている場合は、ロバスト推定に基づくt検定が行われます。有意水準5%で有意なものが強調表示されます。

Practical Difference 実質的な意味があると判断される、平均の差。Y変数の列に［仕様限界］列プロパティが指定されている場合、実質的な差（Practical Difference）は、その指定されている仕様限界の範囲に「実質的な差の割合」を掛けたものとされます。「実質的な差の割合」が指定されていない場合、仕様限界の範囲に0.10を掛けたものとされます。

Y変数の列に［仕様限界］列プロパティが指定されていない場合、まず、標準偏差の推定値が四分位範囲（IQR）から計算されます。この推定値は$\hat{\sigma} = (IQR)/(1.3489795)$で求められます。そして、実質的な差は、$6\hat{\sigma}$に「実質的な差の割合」を掛けたものとされます。「実質的な差の割合」が指定されていない場合、$6\hat{\sigma}$に0.10を掛けたものとされます。

Practical Dif PValue LeveliとLeveljとの間における平均の差が、「Practical Difference」（実質的な差）を超えているかどうかの検定のp値。p値が小さい場合は、差の絶対値が、実質的な差を上回っていることを示唆します。この場合、LeveliとLeveljとの間には、実質的に意味のある差以上の違いがあると結論できます。

Practical Equiv PValue TOST法（Two One-Sided Tests; 片側検定を2回行う方法）を使って、平均の実質的な同等性が検定されます（Schuirmann, 1987）。「Practical Difference」（実質的な差）が、実質的に同等とみなす差の最大値を意味します。「真の差は、実質的な差を上回る」と、「真の差は、実質的な差の符号を逆にしたものを下回る」という帰無仮説に対して、それぞれ片側t検定が実行されます。これら2つの片側検定が棄却されたら、平均の差の絶対値が、実質的な差の範囲内に収まっていることを意味します。したがって、その場合、それら2群は実質的に同等とみなされます。

「Practical Equiv PValue」（実質的同等性のp値）は、2つの片側t検定のp値の大きい方の値です。「Practical Equiv PValue」が小さい場合、LeveliとLeveljの平均は実質的に同等であることを示唆します。

Practical Result 実質的な差に対する検定と、実質的な同等性の検定に対する結論。色分けして表示されるため、有意差の結果を簡単に見分けることができます。

– 「実質的な差あり」（ピンク）：差の絶対値が、実質的な差より有意に大きいことを示します。
– 「実質的に同等」（緑）：差の絶対値が、実質的な差の範囲に有意に収まっていることを示します。
– 「結論できない」（グレー）：実質的な差に関しても、実質的な同等性に関しても、いずれも有意でないことを示します。

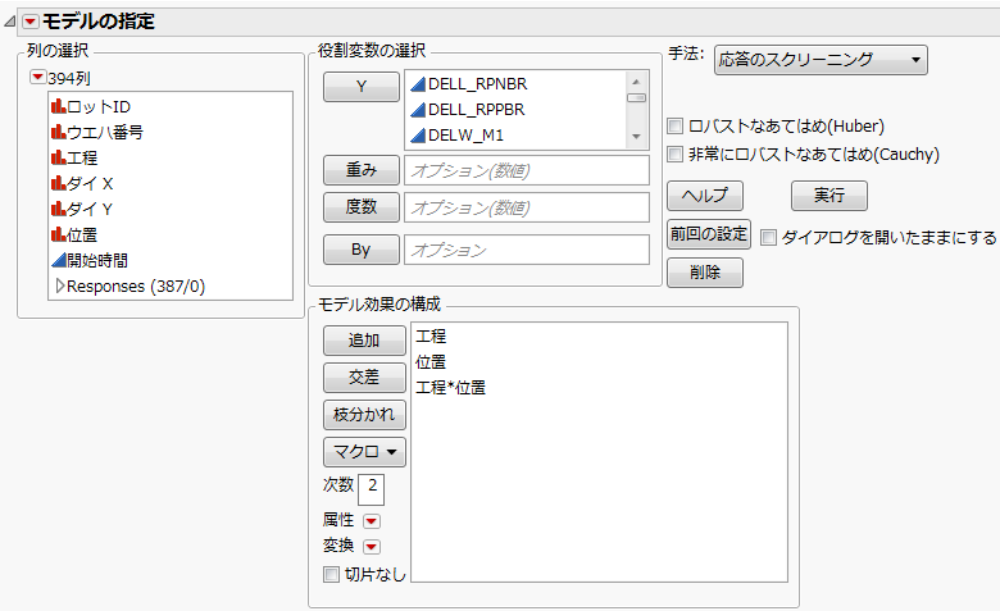
「モデルのあてはめ」の［応答のスクリーニング］手法

「モデルのあてはめ」の［応答のスクリーニング］手法を用いることにより、線形モデルにおける効果に対する検定に注目することができます。作成されるレポートやデータテーブルでは、各応答が、各因子に対して個別に検定されます。

「モデルのあてはめ」での「応答のスクリーニング」の起動

［分析］＞［モデルのあてはめ］を選択します。Y変数とモデル効果を指定します。「手法」のリストから**［応答のスクリーニング］**を選択します（図17.10）。

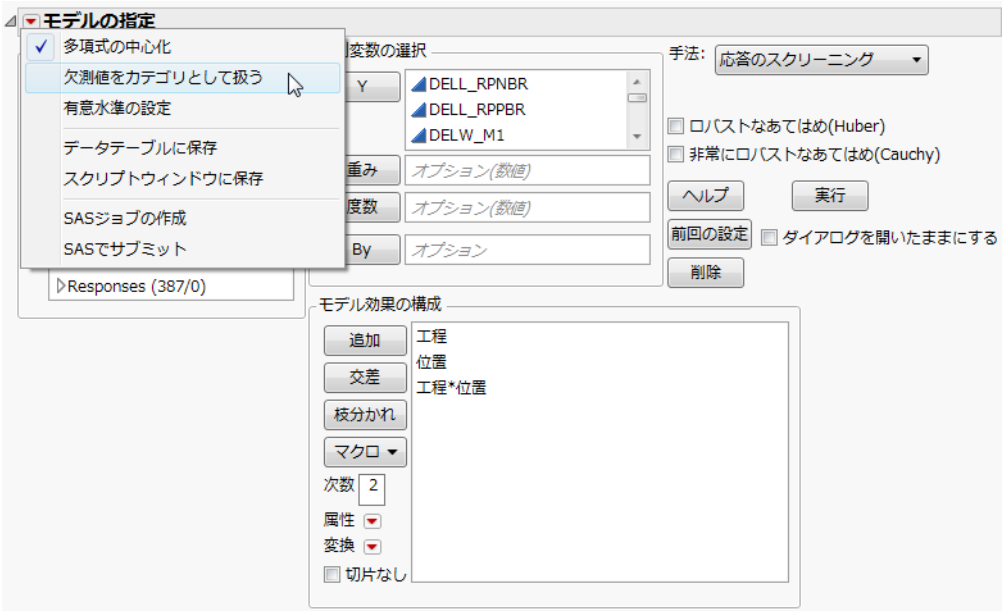
図17.10「モデルのあてはめ」ウィンドウの［応答のスクリーニング］



［ロバストなあてはめ］チェックボックスが表示されます。応答が連続尺度の場合にこのチェックボックスをオンにすると、HuberのM推定という、ロバストな推定が行われます。この推定では、外れ値に対して小さい重みが与えられます。外れ値がない場合、HuberのM推定の結果は、最小2乗推定のものと近くなります。このオプションを選択した場合は、計算時間がかかります。

［欠測値をカテゴリとして扱う］オプションを選択すると、欠測値が1つのカテゴリとして扱われます（図17.11）。このオプションを使うと、欠測値がある行も、計算に含まれます。欠測値にも意味がある場合に便利です。このオプションは、「モデルの指定」の赤い三角ボタンのメニューから選択します。

図17.11［欠測値をカテゴリとして扱う］オプション



「モデルのあてはめ」ウィンドウの詳細については、『基本的な回帰モデル』の「モデルの指定」章を参照してください。

「応答スクリーニングのあてはめ」レポート

「応答スクリーニングのあてはめ」レポートには、2つのプロットが表示されます。

- 「FDR PValue Plot」
- 「FDR LogWorth by Rank Fraction」プロット

「FDR PValue Plot」は、「応答のスクリーニング」プラットフォーム本体と同様に解釈できます。「「応答のスクリーニング」レポート」（288ページ）を参照してください。

「FDR LogWorth by Rank Fraction」には、「FDR LogWorth」（FDR対数価値）が有意性の高い順にプロットされます。プロットの点は、右下がりになるか、もしくは水平です。このプロットは、有意な検定を見るのに役立ちます。［応答のスクリーニング］手法を使用した例については、「［応答のスクリーニング］手法」（310ページ）を参照してください。

モデルダイアログ　レポートの作成時に実行した「モデルの指定」ダイアログが開きます。

推定値の保存 パラメータ推定値を含んだデータテーブルが作成されます。1つの応答変数につき1行あり、各列がモデル項に対応しています。このデータテーブルには、分析に使用したデータテーブルの名前を示す「Original Data」というテーブル変数も含まれます。By変数を指定した場合は、By変数の水準ごとに推定値のテーブルが作成され、「Original Data」変数にBy変数とその水準が表示されます。

予測式の保存 元のデータテーブルに、応答変数の予測式を含む列を追加します。

最小2乗平均の保存 各行が応答と効果の設定の組み合わせに対応しているデータテーブルが開きます。行には、その設定の組み合わせに対する最小2乗平均と標準誤差が含まれます。

以下のオプションについて詳しくは、『JMPの使用法』の「JMPのレポート」章を参照してください。

ローカルデータフィルタ 現在のレポートにおいてデータをフィルタリングするローカルデータフィルタを表示するか、非表示にします。

やり直し 分析を繰り返すオプションや、やり直すオプションを含みます。また、［自動再計算］オプションを選択すると、このオプションに対応しているプラットフォームにおいて、データテーブルに加えた変更が、該当するレポートに即座に反映されるようになります。

スクリプトの保存 レポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。

Byグループのスクリプトを保存 By変数の全水準に対するレポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。起動ウィンドウでBy変数を指定した場合のみ使用可能です。

「PValues」データテーブル

「PValues」データテーブルには、Y変数とモデル効果のペアごとに、以下の結果が1行ずつ出力されます。なお、起動ウィンドウで［ロバストなあてはめ］オプションを選択した場合、モデルはHuberのM推定法によって推定されます。

Y 指定した応答列。

Effect 指定したモデル効果。

FRatio 効果に対する検定の検定統計量。これは、「最小2乗法によるあてはめ」の「効果の検定」レポートに表示される値です。

PValue 「FRatio」（F値）に対応する*p*値。「効果の検定」の詳細については、『基本的な回帰モデル』の標準最小2乗に関する章を参照してください。

LogWorth $-\log_{10}(p\text{値})$。*p*値のグラフのスケールが対数価値に変換されると、解釈がしやすくなります。2を上回る値は、有意水準0.01で有意となります（$-\log_{10}(0.01) = 2$）。

FDR PValue FDR（False Discovery Rate; 偽発見率）を制御するように調整された*p*値。Benjamini-Hochberg法で計算されています。FDRは、検定の多重性を考慮して、生の*p*値を調整したものです。FDRについては、Benjamini and Hochberg（1995）を参照してください。偽発見率の詳細については、「FDR（False Discovery Rate; 偽発見率）」（311ページ）またはWestfall et al.（2011）を参照してください。

FDR LogWorth $-\log_{10}$(FDR調整p値)。これは、検定の有意性をグラフに表すのに適している統計量です。*p*値が小さいと、この値は大きくなります。

Rank Fraction 対数価値の順位を、検定の総回数で割ったもの。検定の総回数を*m*とした場合、対数価値が最大のときに、「Rank Fraction」は1/*m*となります。また、対数価値が最小のときに、「Rank Fraction」は1となります。この「RankFraction」は、対数価値では大きい順ですが、*p*値では小さい順に対応しています。「Rank Fraction」は、「FDR PValue Plot」の横軸に使われます。

Test DF 効果に対する検定の自由度。

「PValues」データテーブルには、分析に使用したデータテーブルの名前を示す「Original Data」というテーブル変数も含まれます。By変数を指定した場合は、By変数の水準ごとに「PValues」テーブルが作成され、「Original Data」変数にBy変数とその水準が表示されます。

「Y Fits」データテーブル

「Y Fits」データテーブルには、要約統計量が出力されます。各Y変数について1行ずつ、データテーブルの各列に、以下のような要約統計量が出力されます。なお、起動ウィンドウで［ロバストなあてはめ］オプションを選択した場合、モデルはHuberのM推定法によって推定されます。

Y 指定した応答列。

R2乗 R2乗値（重相関係数、寄与率、決定係数）

RMSE 誤差の標準偏差。

度数 オブザベーションの個数（または［重み］変数の和）。

Overall FRatio 「最小2乗法によるあてはめ」の「分散分析」レポートに表示される、モデル全体に対する有意性検定の検定統計量。

Overall PValue モデル全体に対する有意性検定の*p*値。

Overall LogWorth モデル全体に対する有意性検定の*p*値の対数価値。

Overall FDR PValue モデル全体に対する*p*値をFDR調整したもの。（「「応答のスクリーニング」レポート」（288ページ）を参照）。

Overall FDR LogWorth 「Overall FDR PValue」の対数価値。

Overall Rank Fraction 対数価値の順位を、検定の総回数で割ったもの。検定の総回数を*m*とした場合、対数価値が最大のときに、「Rank Fraction」は1/*m*となります。また、対数価値が最小のときに、「Rank Fraction」は1となります。

<効果> PValue これらの列には、各モデル効果に対する検定の p 値が表示されます。「列」パネルでは「PValue」というグループにまとめて表示されます。

<効果> LogWorth 各モデル効果に対する検定の p 値の対数価値が表示されます。「列」パネルでは「LogWorth」というグループにまとめて表示されます。

<効果> FDR LogWorth これらの列には、各モデル効果に対するFDR調整p値の対数価値が表示されます。「列」パネルでは「LogWorth」というグループにまとめて表示されます。

「Y Fits」データテーブルには、分析に使用したデータテーブルの名前を示す「Original Data」というテーブル変数も含まれます。By変数を指定した場合は、By変数の水準ごとに「Y Fits」テーブルが作成され、「Original Data」変数にBy変数とその水準が表示されます。

「応答のスクリーニング」の別例

以降では、「応答のスクリーニング」のさまざまな使用例を紹介していきます。

実質的な差や実質的な同等性に対する検定の例

この例では、「Probe.jmp」サンプルデータテーブルを使って、実質的な差の検定を行います。

1. ［ヘルプ］>［サンプルデータライブラリ］を選択し、「Probe.jmp」を開きます。
2. ［分析］>［スクリーニング］>［応答のスクリーニング］を選択します。
   「応答のスクリーニング」起動ウィンドウが表示されます。
3. 「Responses」列グループを選択し、［Y, 応答変数］をクリックします。
4. 「工程」を選択し、［X］をクリックします。
5. 「実質的な差の割合」ボックスに「0.15」と入力します。
6. ［OK］をクリックします。
7. 「応答のスクリーニング」レポートの赤い三角ボタンのメニューから、［平均の比較を保存］を選択します。
   図17.12は、作成されるデータテーブルの一部を示しています。Yに指定した応答ごとに、「工程」の水準である「New」と「Old」を比較した検定の結果が出力されています。

図17.12 平均の比較のテーブル（一部）

	Y	Leveli	Levelj	Std Err Diff	Plain Dif PValue	Practical Difference	Practical Dif PValue	Practical Equiv PValue	Practical Result
1	DELL_RPNBR	New	Old	0.004326425	8.044596e-13	0.1486469486	1	1.57079e-153	実質的に同等
2	DELL_RPPBR	New	Old	0.0349025251	0.4782555718	1.1939335307	1	2.40983e-225	実質的に同等
3	DELW_M1	New	Old	0.0205582035	0.1806001982	0.7030518329	1	1.40119e-217	実質的に同等
4	DELW_M2	New	Old	0.0017890895	0	0.0704825217	0.0269561256	0.9730438744	実質的な差あり
5	DELW_NBASE	New	Old	1.4836374126	0.388034756	50.689291592	1	6.62236e-223	実質的に同等
6	DELW_NEMIT	New	Old	0.002521398	7.306034e-56	0.0848384267	1	7.733811e-69	実質的に同等
7	DELW_NENBNI	New	Old	0.0220490008	0	0.864024229	0.0003726256	0.9996273744	実質的な差あり
8	DELW_NSINK	New	Old	0.0254310074	0	1.0723100972	3.228438e-37	1	実質的な差あり
9	DELW_PBASE	New	Old	0.0037311377	2.908704e-31	0.1296384753	1	8.19663e-113	実質的に同等
10	DELW_PCOLL	New	Old	0.0085444762	0	1.1754000513	0	1	実質的な差あり
11	DELW_PEMIT	New	Old	0.0534170363	0.6709186367	1.8265075627	1	1.2499e-228	実質的に同等
12	DELW_PSINK	New	Old	0.0678503246	1.666427e-40	2.3558597461	1	3.147655e-97	実質的に同等
13	DELW_RPNBR	New	Old	0.0100452789	9.008231e-24	0.3464883523	1	1.6441e-125	実質的に同等
14	DELW_RPPBR	New	Old	0.0280757905	1.169725e-26	0.9698077792	1	7.48088e-120	実質的に同等
15	DELW_SICR	New	Old	0.0176173979	0.1180928699	0.6024110643	1	7.77189e-215	実質的に同等
16	M1_COMB_VG…	New	Old	0.6037284633	0.3778451214	20.662273741	1	1.83138e-223	実質的に同等
17	M1_TRENCH_V…	New	Old	0.1695447206	0.0002786205	5.8087989555	1	2.39165e-191	実質的に同等
18	M2/M1_CAP_V…	New	Old	0.0978790012	0.6294372314	3.3496974301	1	2.54456e-228	実質的に同等
19	M2_COMB_BB…	New	Old	0.1270241159	0.9988767052	4.3470387459	1	3.07472e-234	実質的に同等
20	M2_COMB_VG…	New	Old	0.2860872543	0.2990592705	9.7914324315	1	1.42412e-221	実質的に同等
21	NISO_TUB-TR…	New	Old	0.2482751793	1.58784e-164	9.0621071628	1	9.077234e-17	実質的に同等
22	NISO_TUB-TU…	New	Old	0.1081535234	0.4892691678	3.7013990674	1	8.97914e-226	実質的に同等
23	PS_RPNBR	New	Old	15.2943887	0.0000237641	530.27598935	1	3.11434e-189	実質的に同等

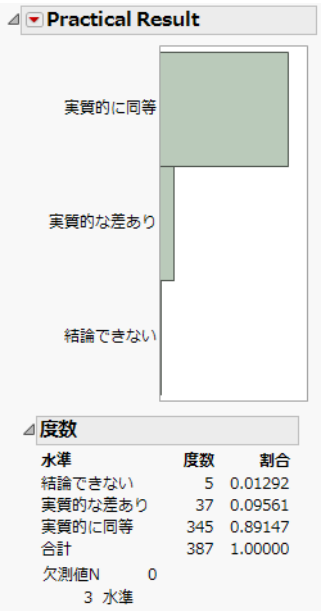
「Probe.jmp」には仕様限界が列プロパティとして保存されていないため、応答ごとに実質的な差が自動的に設定されます。実質的な差として指定した0.15に、応答の6σ範囲を掛け合わせて求めた値が、実質的な差と同等性の検定に使用されます。この値は、**「Practical Difference」**列に表示されます。

「Plain Dif PValue」列では、「差がゼロである」という帰無仮説に対して有意である応答を確認できます。「Practical Diff PValue」列と「Practical Equiv PValue」列には、実質的な差と実質的な同等性に対する検定のp値が表示されます。多数の比較において、「差がゼロである」という帰無仮説に対する検定には**統計的**有意差を示していますが、**実質的**有意差を示しているものはありません。

8. 平均の比較のテーブルを表示し、**［分析］＞［一変量の分布］**を選択します。
9. **「Practical Result」**を選択して、**［Y, 列］**をクリックします。
10. **［OK］**をクリックします。

図17.13は、実質的有意差の結果の分布を示しています。指定した実質的な差の検定により有意差ありと判定されているのは、わずか37の検定だけです。応答変数のうち5つについては、検定の結論が出ていません。これらの応答については、**「Process」**の水準間で実質的有意差があるかどうか判定不能です。

図17.13 実質的有意差の結果の分布



37の応答に対応する棒をプロット上でクリックして選択し、これらの応答についてさらに検討できます。

「最大対数価値」オプションの例

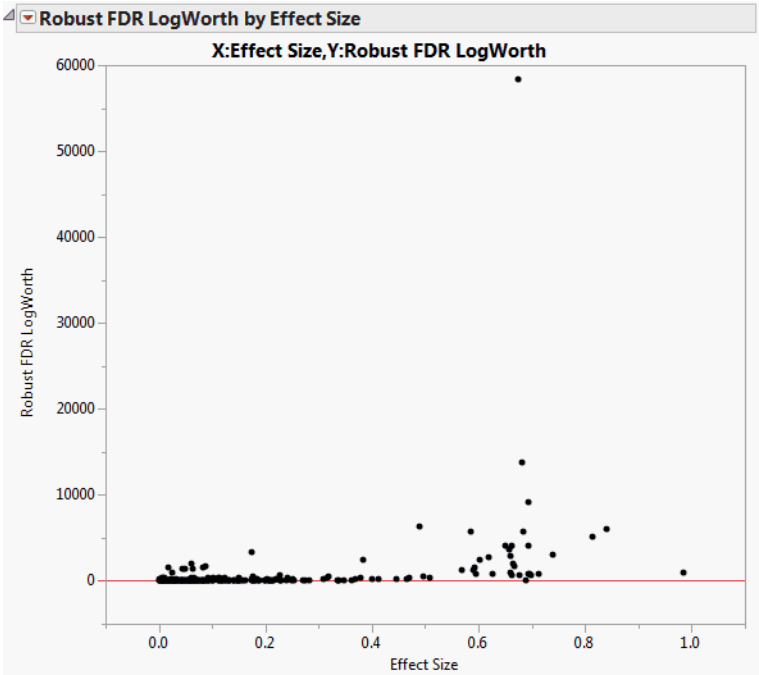
データの標本サイズが大きくなると、*p* 値が極端に小さくなる場合があります。このような場合も、対数価値を利用すれば、*p* 値をグラフにわかりやすく表示できます。ただし、場合によっては、*p* 値が小さすぎるせいで対数価値が大きくなりすぎ、スケールに影響が出ることがあります。

1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Probe.jmp」を開きます。
2. ［**分析**］>［**スクリーニング**］>［**応答のスクリーニング**］を選択します。
3. 「応答のスクリーニング」起動ウィンドウで、「**Responses**」列グループを選択して［**Y, 応答変数**］をクリックします。
4. 「**工程**」を選択し、［**X**］をクリックします。
5. ［**ロバスト**］チェックボックスをオンにします。
6. ［**OK**］をクリックします。

   この分析は数的負荷がかかるため、終了するまでに時間がかかることがあります。
7. 「応答のスクリーニング」レポートで、「Robust FDR LogWorth by Effect Size」レポートを開きます。

   このプロットは、「Robust FDR LogWorth」の値が大きすぎる（約58,000）点があるせいで、肝心な部分が見づらくなっています（図17.14）。グラフの細部を十分確認できるように、対数価値の最大値を設定できます。

図17.14「Robust FDR LogWorth」と「Effect Size」のプロット（「最大対数価値」は未指定）

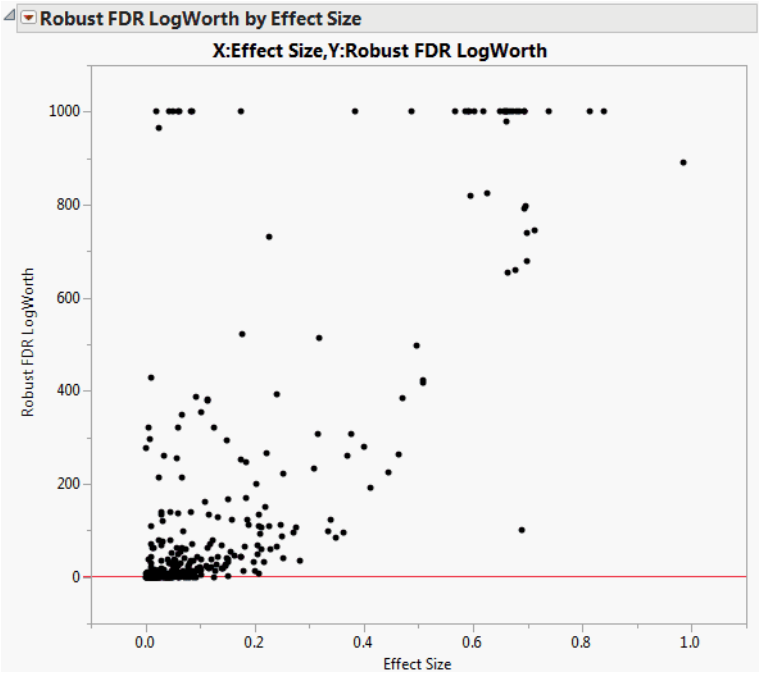


8. 手順1から手順5の手順を繰り返します。
9. 起動ウィンドウ下部の「最大対数価値」ボックスに「1000」と入力します。
10. [**OK**] をクリックします。

    分析が終了するまで、時間がかかることがあります。
11. 「応答のスクリーニング」レポートで、「Robust FDR LogWorth by Effect Size」レポートを開きます。

    今回は、プロットの細部がわかりやすく表示されます（図17.15）。

図17.15「Robust FDR LogWorth」と「Effect Size」のプロット（「最大対数価値」= 1000）

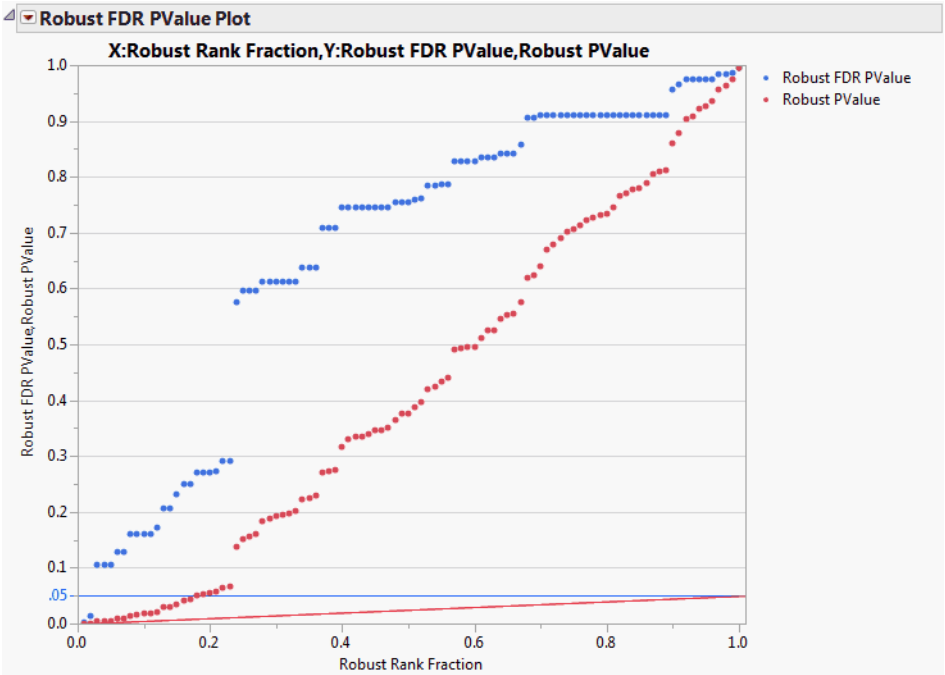


ロバストなあてはめの例

1. 「Drosophila Aging.jmp」テーブルを開きます。
2. ［**分析**］>［**スクリーニング**］>［**応答のスクリーニング**］を選択します。
3. 連続尺度の列をすべて選択して、［**Y, 応答変数**］をクリックします。
4. **「遺伝子型」**を選択し、［**X**］をクリックします。
5. ［**ロバスト**］チェックボックスをオンにします。
6. ［**OK**］をクリックします。

   図17.16は、「Robust FDR PValue Plot」を示しています。ロバスト推定の未調整p値を用いた場合、いくつかの検定で有意差が認められることが、0.05を下回る赤い点の存在で示されています。ただし、ロバスト推定のFDR調整p値を見ると、有意差が認められる検定は2つだけです。

図17.16「Drosophila Aging」データの「Robust FDR PValue Plot」

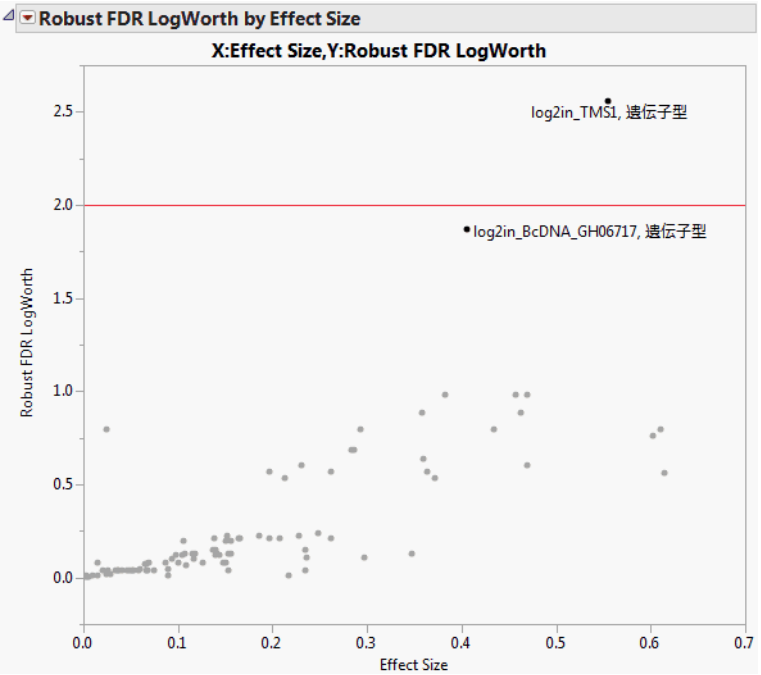


これらの2点は、「FDR LogWorth」（FDR補正の対数価値）を表示するプロットで、もっと簡単に見分けることができます。

7. 「Robust FDR LogWorth by Effect Size」の開閉アイコンをクリックします。
8. 「Robust FDR LogWorth」の値が1.5を上回っている2点を四角く囲むようにドラッグします。
9. 「PValues」データテーブルで、［**行**］>［**ラベルあり / ラベルなし**］を選択します。

プロットが図17.17のように表示されます。2の位置で引かれた赤い線より上にある点は、有意水準が0.01を下回っています。だいたい1.3のあたりで横に引いた線が、有意水準0.05に対応します。

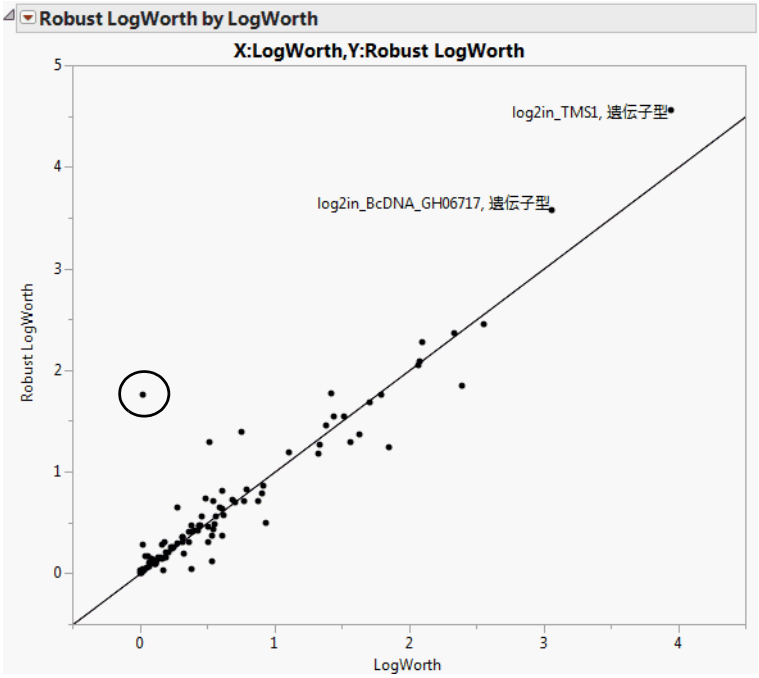
図17.17「Drosophila Aging」データの「Robust LogWorth by Effect Size」



10.「Robust LogWorth by LogWorth」の開閉アイコンをクリックします。

図17.18のようなプロットが作成されます。ロバストな検定が通常の検定とまったく同じであれば、図17.18の点は、対角線上に沿ってプロットされます。図で丸く囲まれている点は、「Robust LogWorth」値が「LogWorth」を上回っており、対角線付近に位置していません。

図17.18「Drosophila Aging」データの「Robust LogWorth by LogWorth」

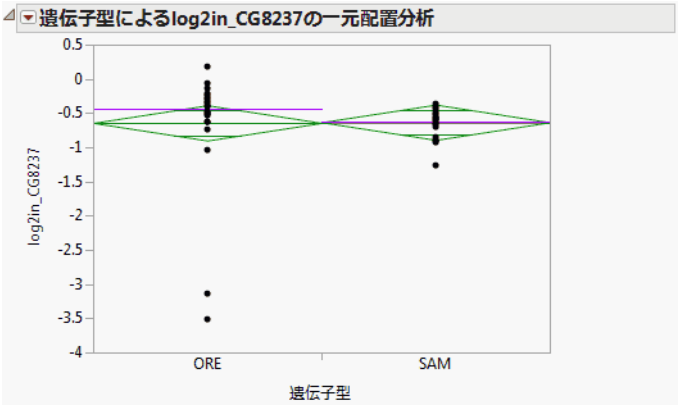


11. プロット上でこの点の周囲を四角くドラッグします。
12. 「PValues」データテーブルでこの点に該当する行を見つけます。

    応答**「log2in_CG8237」**の「PValue」が0.9568で、「Robust PValue」が0.0176であることがわかります。
13. 「応答のスクリーニング」レポートの赤い三角ボタンのメニューから、**［選択した項目の二変量関係］**を選択します。

    「選択した項目の二変量関係」レポートに、応答**「log2in_CG8237」**の一元配置分析の結果が表示されます。プロットには、**「遺伝子型」**OREについて、2つの外れ値が表示されます（図17.19）。これらの外れ値が、ロバストな検定と通常の検定の結果が大きく異なっていた理由です。通常の検定では、外れ値により誤差分散が過大に推定され、有意な効果を検出しにくくなっています。一方、ロバストな検定では、これらの外れ値に小さな重みを与えるので、誤差分散の推定に外れ値の与える影響が小さくなっています。

図17.19 log2in_CG8237の一元配置分析



［応答のスクリーニング］手法

「モデルのあてはめ」の［応答のスクリーニング］手法では、線形モデルの効果に対する検定を、複数の応答に対して一度に行えます。この例では、主効果が2つ、交互作用が1つの線形モデルを分析してみます。

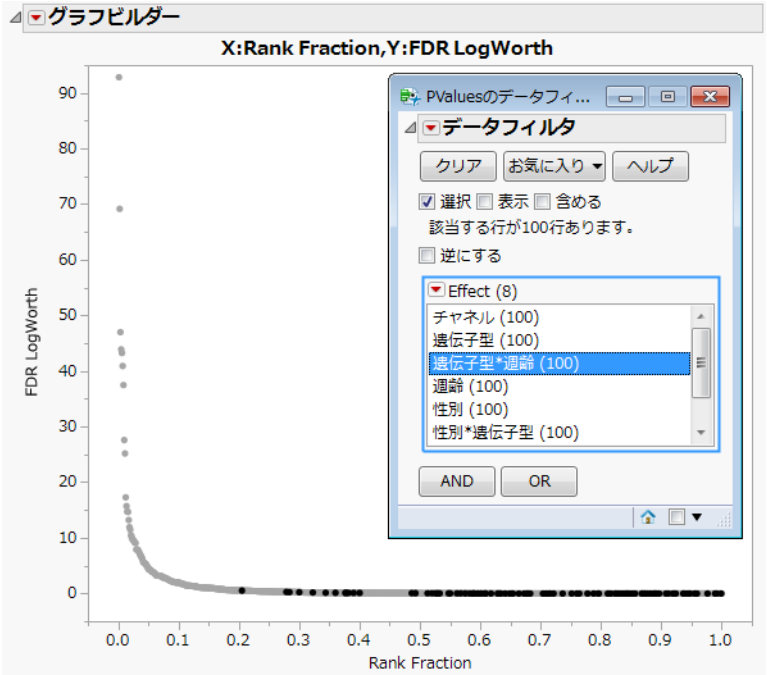
1. 「Drosophila Aging.jmp」テーブルを開きます。
2. ［**分析**］>［**モデルのあてはめ**］を選択します。
3. 連続尺度の列をすべて選択して、［**Y**］をクリックします。
4. **「チャネル」**を選択し、［**追加**］をクリックします。
5. **「性別」**、**「遺伝子型」**、および**「週齢」**を選択して、［**マクロ**］>［**完全実施要因**］を選択します。
6. 「手法」のリストから［**応答のスクリーニング**］を選択します。
7. ［**実行**］をクリックします。

   「応答スクリーニングのあてはめ」レポートが表示されます。「Y Fits」と「PValues」という2つのデータテーブルも作成されます。前者にはモデル全体の検定の結果が、後者にはYごとにモデルの各効果の検定結果が表示されます。

   重要な効果を見極めるために、次の操作を行います。
8. 「PValues」データテーブルの「FDR LogWorth by Rank Fraction」スクリプトを実行します。
9. ［**行**］>［**データフィルタ**］を選択します。
10. 「データフィルタ」ウィンドウで、**「Effect」**を選択して［**追加**］をクリックします。
11. 「データフィルタ」のリストボックスで効果を順次クリックし、それに応じて「Rank Fraction」と「FDR LogWorth」のグラフで選択される点を確認します。

    対数価値（LogWorth）の値が2を上回る点は、0.01水準で有意となります。データフィルタの適用により、**「性別」**と**「チャネル」**を除けば、他のモデル効果が0.01水準で有意となることは稀であることがわかります。図17.20では、2の位置に参照線が表示されています。**「遺伝子型*週齢」**交互作用項の検定の点が選択されています。いずれの点も、0.01水準で有意となっていません。

図17.20「FDR LogWorth」と「Rank Fraction」のプロット（「遺伝子型*週齢」の検定を選択）



統計的詳細

FDR（False Discovery Rate; 偽発見率）

「応答のスクリーニング」では、Benjamini and Hochberg（1995）法によって調整したp値を算出します。Westfall et al.(2011)も参照してください。この手法では、p値が一様分布に従い、独立していると仮定します。

この方法で計算すると、FDR（False Discovery Rate; 偽発見率）を水準α以内に抑えることができます。この方法は、以下の手順で計算されます。

1. 通常の仮説検定を行い、複数のp値（$p_1, p_2, \ldots, p_m$）を得ます。
2. p値を小さいものから順に並べます。小さい順に並べたp値を$p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$と表記します。
3. $p_{(i)} \leq (i/m)\alpha$となる最大のp値を探します。この条件を見たす最大のp値が、k番目に大きい（$p_{(k)}$）だったとします。
4. p値が$p_{(k)}$以下である、k個の仮説を棄却します。

こうすれば、FDRの期待値がαを上回りません。

FDR調整したp値（$p_{(i),\,FDR}$）は、次式で計算されます。

$$p_{(i),\,FDR} = \begin{cases} p_{(m)} & \text{for } i = m \\ min\left[p_{(i+1),\,FDR}, \left(\frac{m}{i}\right)p_{(i)}\right] & \text{for } i = m-1, \ldots, 1 \end{cases}$$

このFDR調整したp値がαを下回る場合、その仮説は棄却されます。

第18章

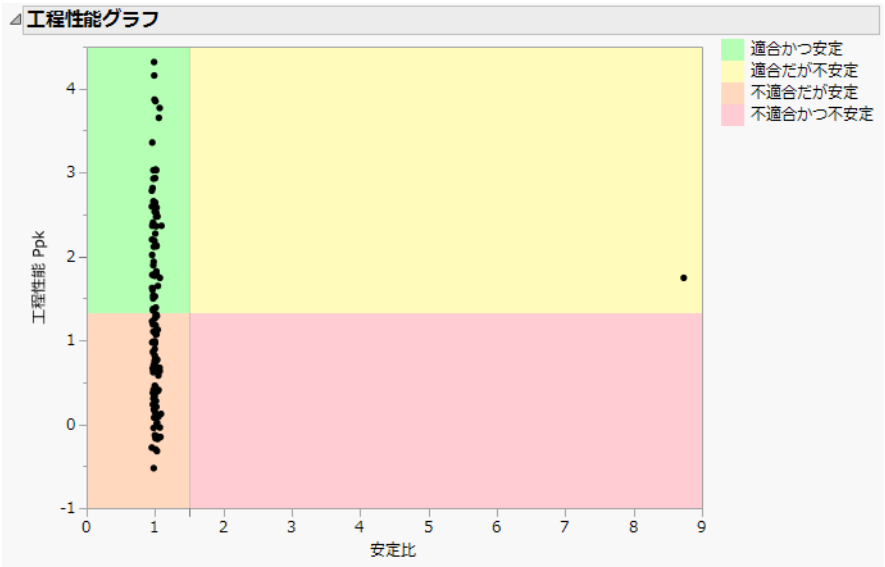
工程のスクリーニング

安定性や工程能力の高い工程を探し出す

「工程スクリーニング」プラットフォームでは、一定の期間における多数の工程を検討できます。このプラットフォームは、工程の安定性と工程能力に関する指標を算出します。また、管理図を描いたり、工程に生じたシフトを検出したりできます。このプラットフォームは、不安定になっている工程、仕様限界に適合していない工程、シフトが生じている工程をすばやく特定でき、工程が多数ある場合でも評価がしやすいです。

最初の結果から、気になる工程をグラフで検討したり、詳しく分析したりできます。「管理図ビルダー」や「工程能力」プラットフォームを簡単に呼び出せます。また、すべての工程、または特定の工程に関して、それらの詳細な結果を保存できます。

図18.1「工程性能グラフ」の例



「工程のスクリーニング」プラットフォームの概要

「工程のスクリーニング」プラットフォームは、多数の工程における安定性（stability）や工程能力（capability）を評価するためのプラットフォームです。また、管理図で管理外となっているかどうか、なども求められます。「工程のスクリーニング」では、以下のことを行えます。

- 管理図で用いるサブループを構成するために、一定のサブグループサイズを指定するか、サブグループIDを含む変数を指定する。
- 異なる工程を識別するための変数を、グループ変数に指定する。グループ変数がもつ値の各組み合わせに対し、分析が行われます。
- 中央線やシグマの計算にメディアンを使い、外れ値の影響を受けにくくする。
- 工程平均のどこに大きなシフトが生じたかを調べる。

工程の平均や散らばりにおける変化を検出するために、管理図のテストのいくつかが要約表に表示されますが、これらの要約表に表示するテストは選択できます。仕様限界を指定した場合は、工程能力も求められます。また、「工程性能グラフ」は、安定性と、工程能力（仕様限界への適合性）をプロットしたグラフです。「シフトグラフ」は、上昇シフトや下降シフトが生じた位置を示すグラフです。

「工程のスクリーニング」では、詳しく分析したい工程を簡単に選べます。また、「簡易グラフ」は、サイズが小さいので、かなりの数の工程を一度に見ることができます。「管理図ビルダー」と「工程能力」の各プラットフォームを簡単に呼び出すこともでき、選択した工程を続けて詳しく分析できます。

すべての工程、または特定の工程の結果を含むデータテーブルを、さまざまな形式で保存できます。

「工程スクリーニング」の例

「Semiconductor Capability.jmp」データテーブルには、工程の測定値が128列にわたってまとめられています。どの列にも 1,455 個の測定値が記録されています。ここでは、どの工程が不安定であるかを調べてみましょう。各列には、それぞれ「仕様限界」列プロパティも含まれています。安定している工程は、工程能力の計算に適しています。最後に、このデータテーブルで安定性と工程能力を評価します。

1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Semiconductor Capability.jmp」を開きます。
2. ［**分析**］>［**スクリーニング**］>［**工程のスクリーニング**］を選択します。
3. 「**Processes**」列グループを選択し、［**工程変数**］をクリックします。

   「管理図の種類」が［I-MR管理図］になっていることを確認します。
4. ［**OK**］をクリックします。

図18.2 最初のレポート（一部）

工程のスクリーニング

	計量値			要約			管理図の警告		工程能力				
列	安定比	群内シグマ	全体シグマ	平均	度数	警告率	テスト 1	最後の警告	仕様限界外の度数	仕様限界外の割合	最後の仕様限界外	Cpk	Ppk
PMS1	8.73	6.4e-15	1.9e-14	-50	1455	0.03643	53	3	0	0	.	5.159	1.746
IVP8	1.09	0.45226	0.47293	-21.69	1455	0.00550	8	49	0	0	.	2.474	2.366
B1	1.09	5.70202	5.94318	125.972	1455	0.00275	4	43	604	0.4151	1	0.130	0.124
IVP7	1.08	5.92093	6.14904	-79.865	1455	0.00275	4	114	1032	0.7093	1	-0.157	-0.151
SIT2	1.07	3.59488	3.71963	41.1722	1455	0.00550	8	21	0	0	.	1.806	1.745
CAP	1.07	0.92619	0.95805	38.1075	1455	0.00412	6	12	1152	0.7918	1	-0.039	-0.038
VPM7	1.07	3.20356	3.3107	85.321	1455	0.00550	8	107	54	0.0371	13	0.656	0.634
NPN10	1.07	1.76238	1.82043	73.5168	1455	0.00481	7	38	0	0	.	3.895	3.770
INV2	1.07	0.6307	0.65112	17.3705	1455	0.00481	7	11	34	0.0234	25	0.697	0.675
NPN9	1.06	2.40507	2.47045	47.8068	1455	0.00344	5	9	0	0	.	3.752	3.653
IVP9	1.05	1.61702	1.65762	-30.721	1455	0.00344	5	137	145	0.0997	1	0.420	0.410
VTN210	1.05	2.30557	2.35977	0.09116	1455	0.00344	5	9	55	0.0378	6	0.596	0.582
SIT1	1.05	15.3977	15.7506	149.659	1455	0.00481	7	16	583	0.4007	2	0.090	0.088
INM1	1.04	3.28224	3.34868	82.4373	1455	0.00412	6	3	0	0	.	1.682	1.649
IVP3	1.04	3.08312	3.14263	-49.493	1455	0.00206	3	3	168	0.1155	4	0.403	0.395
VTP210	1.04	3.31312	3.3769	1.29622	1455	0.00344	5	25	1	0.0007	1052	1.150	1.128
PBL1	1.04	0.26689	0.27188	2.71456	1455	0.00481	7	65	31	0.0213	53	0.681	0.669
NPN6	1.04	1.12479	1.14543	43.2968	1455	0.00206	3	69	1115	0.7663	2	-0.177	-0.174
E2A1	1.04	0.00132	0.00134	0.62966	1455	0.00344	5	21	0	0	.	2.521	2.478
VDP1	1.03	0.23752	0.24127	28.8111	1455	0.00137	2	17	15	0.0103	27	0.783	0.771
M1_M1	1.03	0.20527	0.20846	0.23703	1455	0.00412	6	26	1002	0.6887	1	-0.148	-0.146
A2N	1.03	8.53741	8.66111	56.0799	1455	0.00344	5	25	925	0.6357	1	0.019	0.019

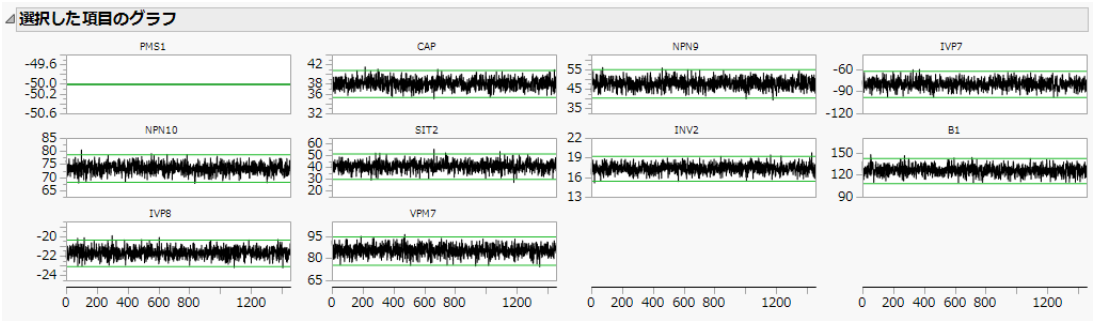
「工程のスクリーニング」ウィンドウが開き、各工程の結果をまとめた要約表が表示されます。各工程は「安定比」の降順に並んでいます。「安定比」は工程の安定性を示す指標で、その数値が1に近いほど工程は安定しています。安定比が大きいものほど、工程が安定していないことを示します。（この表において、「安定比」の隣にある「^」というマークがありますが、これは同列が並び替えの基準であることを示しています。）ここでは、安定比が1.05を超えている工程を、詳しく調べるとしましょう。

5. レポートウィンドウで**「PMS1」**から**「NPN9」**までの工程を選択します。

   最初の10個の工程は、いずれも「安定比」の値が1.05を上回っています。

6. 「工程のスクリーニング」の赤い三角ボタンをクリックし、**［選択した項目の簡易グラフ］**を選択します。

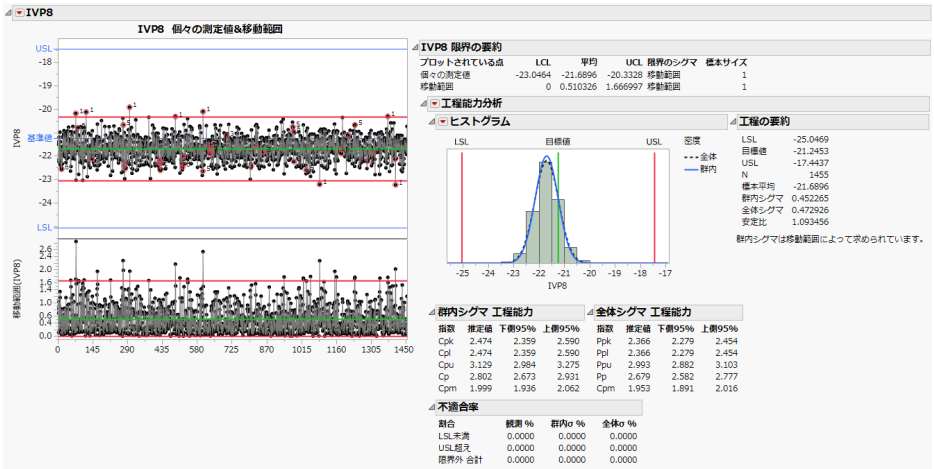
図18.3 安定比が最も大きい工程の簡易グラフ



「IVP8」（「選択した項目のグラフ」の上から3段目、一番左）を詳しく調べてみましょう。

7. 要約テーブルで2番目にある**「IVP8」**を選択します。
8. 「工程スクリーニング」の赤い三角ボタンをクリックし、**［選択した項目の管理図］**を選択します。

図18.4「IVP8」の管理図ビルダー

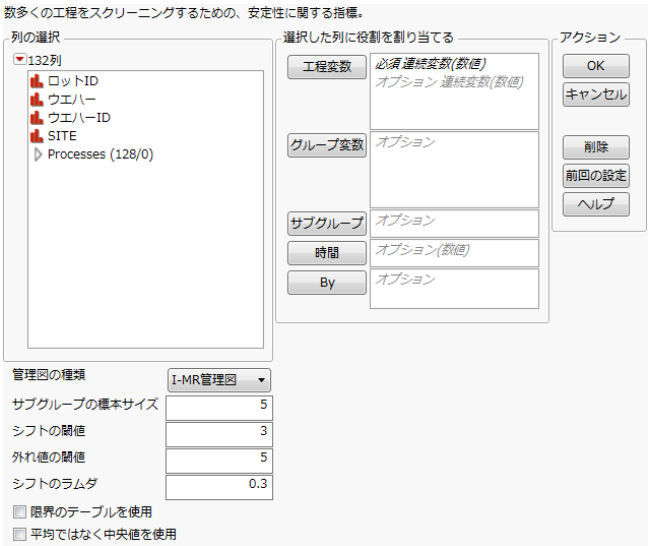


「管理図ビルダー」レポートが表示されます。**「IVP8」**には「仕様限界」列プロパティがあるため、工程能力分析も表示されます。

「工程のスクリーニング」プラットフォームの起動

「工程のスクリーニング」プラットフォームを起動するには、[**分析**] > [**スクリーニング**] > [**工程のスクリーニング**] を選択します。

図18.5「工程のスクリーニング」起動ウィンドウ



起動ウィンドウの役割

工程変数 分析したい測定値を含む工程データの列。データタイプは「数値」でなければなりません。

グループ変数 工程変数として指定した各列と、グループ変数の水準の各組み合わせごとに、該当する工程が個別に分析されます。結果は1つのレポートにまとめられます。

サブグループ この役割を割り当てた列の値は、各行のサブグループIDとして使用されます。なお、計算を行う前に、指定したサブグループ変数に従ってデータは並べ替えられます。

時間 数値列で、この列の値によってデータの時間的順序が決まります。この［時間］には、タイムスタンプ（時刻、時点）のデータを指定します。ここで指定された時間は、簡易グラフとシフトグラフの時間軸に使用されます。なお、計算を行う前に、指定したサブグループ変数に従ってデータは並べ替えられます。

By 別々に分析を行いたいときに、そのグループ分けをする変数を指定します。指定された列の各水準ごとに、別々に分析が行われます。各水準の結果は別々のレポートに表示されます。複数のBy変数を割り当てた場合、それらのBy変数の水準の組み合わせごとに別々のレポートが作成されます。

起動ウィンドウのオプション

管理図の種類 ［I-MR管理図］（個々の測定値と移動範囲の管理図）、［XBar-R管理図］、［XBar-S管理図］の3つの中から1つを選択します。統計的詳細については、『品質と工程』の「管理図ビルダー」章を参照してください。

サブグループの標本サイズ サブグループの標本サイズを指定します。2以上の値を使用してください。デフォルトでは5に設定されています。「サブグループの標本サイズ」は、管理図の種類が［I-MR管理図］である場合、またはサブグループ変数を指定した場合には無視されます。

シフトの閾値 「シフトグラフ」の感度を決める係数を指定します。デフォルトでは3に設定されています。「シフトグラフ」には、外れ値は除外した後、「シフトの閾値」に群内シグマを掛けたものを超えるシフトの発生時点がすべてプロットされます。「シフトグラフ」（324ページ）を参照してください。

外れ値の閾値 シフト検出と「シフトグラフ」において外れ値とみなすもの決めるときに用いる係数を指定します。「外れ値の閾値」は、デフォルトで5に設定されています。「外れ値の閾値」に群内シグマを掛けたものよりも両隣の測定値が離れている場合、該当の測定値を、最も近い測定値から1群内シグマだけ離れた値に置き換えます。「シフトの大きさと位置」（322ページ）を参照してください。

シフトのラムダ 「シフトグラフ」で使用する指数加重移動平均（EWMA）の重みを変更できます。「シフトの大きさと位置」（322ページ）を参照してください。

限界のテーブルを使用 管理限界と仕様限界の履歴データをデータテーブルから読み込むことができます。このオプションをオンにして起動ウィンドウの［OK］をクリックし、「限界のテーブルを選択」ウィンドウでデータテーブルを選択すると、「限界の指定」ウィンドウが開きます。そこで、限界を含むデータテーブルの列に適切な役割を割り当て、［OK］をクリックしてください。詳細については、「限界のテーブル」（318ページ）を参照してください。

平均ではなく中央値を使用　このオプションを選択すると、中央線の推定に中央値が使用されます。また、モンテカルロシミュレーションによって得られた係数に基づいて、シグマが推定されます。「統計的詳細」（327ページ）に、その係数の表を記載しています。その計算方法は、選択した管理図の種類によって異なります。

- ［XBar-R管理図］および［I-MR管理図］では、範囲の中央値からシグマが推定されます。
- ［XBar-S管理図］では、標準偏差の中央値からシグマが計算されます。
- なお、サブグループの標本サイズが等しくない場合には、標本サイズに対応する係数を見るのに、サブグループサイズの平均を丸めた整数が使われます。

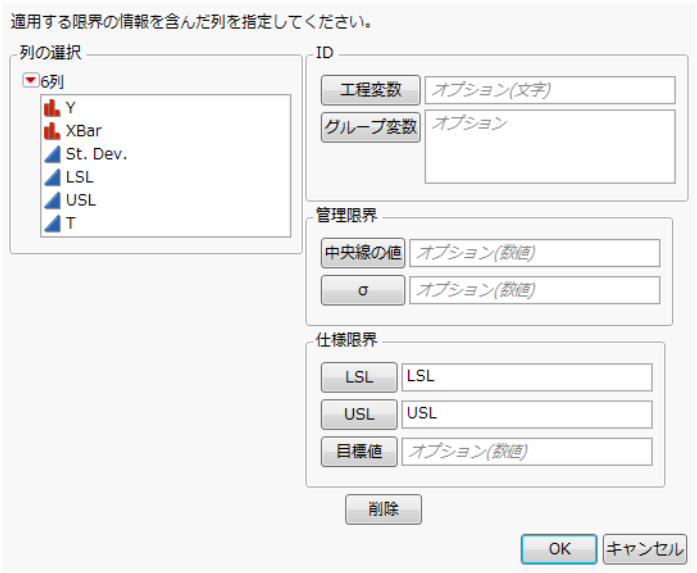
1つまたは複数の外れ値がある場合、平均値を中央線に用いると、多数のサブグループが管理外であるかのように見えるようになるかもしれません。その問題は、中央値を中央線に使うことで回避できます。

メモ：［平均ではなく中央値を使用］を選択すると、赤い三角ボタンのメニューにある［選択した項目の管理図］で得られる結果が「工程のスクリーニング」の結果と一致しなくなります。「管理図ビルダー」には、平均の代わりに中央値を使用する手法がないためです。

限界のテーブル

限界のテーブルには、分析対象のデータで工程変数とグループ変数で定義された工程ごとに、1つの行を含めてください。限界のテーブルを使用する場合、「限界の指定」ウィンドウで変数に以下のような役割を割り当ててください。すべての役割に対して変数を指定する必要はありません。どの役割もオプションで、必須ではありません。

図18.6「限界の指定」ウィンドウ



管理限界と仕様限界の列は、列名が、「Center」、「Sigma」、「LSL」、「USL」、「Target」となっているものは自動的に入力されます。また、列名が「Process Variable」、「Process」、「Column」、「Parameter」のいずれかになっている列は、［工程変数］のリストに自動入力されます。

管理限界だけがあり、［中央線］と［σ］の列がない場合には、以下のような計算式を使って限界のテーブルに**「中央線」**と**「σ」**の列を作成することができます。計算式の例を紹介します。

$$\text{中央線} = (UCL + LCL)/2$$
$$\sigma = d(UCL - LCL)/6$$

上の式で、*d* はサブグループの標本サイズの平方根です。

工程変数　工程変数の列名を含む列。

グループ変数　グループ変数の列名を含む列。グループ変数は複数あっても構いません。

中央線　各工程の中央線として使用する値を含む列。通常は、工程の履歴平均を使用します。

σ　各工程の群内標準偏差の値を含む列。通常は、履歴標準偏差を使用します。

LSL　各工程の下側仕様限界を含む列。

USL　各工程の上側仕様限界を含む列。

目標値　各工程の目標値を含む列。

「工程のスクリーニング」レポート

「工程のスクリーニング」レポートには、工程の安定性などに関する結果をまとめた要約表が表示されます。仕様限界を指定した場合は、工程能力に関する統計量も表示されます。工程やグループは、「安定比」の降順で並び替えられています。

ヒント：他の列で並べ替えるには、列名をクリックします。並び替えに使われた列の隣には、「^」というマークが表示されます。「^」は、昇順なら上向き、降順なら下向きとなります。昇順と降順を切り替えるには、もう一度列名をクリックします。

要約表にある管理図の情報には、特殊原因に対するテストと、［範囲の限界外］に関するテストが含まれます。管理図に関するこれらのテストには、次のような中央線や管理限界が使われます。

- XBar管理図やX管理図の中央線には、すべての測定値の平均が使われる。ただし、［平均ではなく中央値を使用］オプションを選択した場合は、中央値が使われる。
- 管理限界は、中央線から3シグマ離れた位置に配置される。
- シグマの推定値は、指定した管理図の種類に応じた方法で計算される。また、［平均ではなく中央値を使用］オプションを選択した場合は、「平均ではなく中央値を使用」（318ページ）で説明されているように計算される。

ヒント：「工程のスクリーニング」プラットフォームで行われる特殊原因テストは、「管理図ビルダー」プラットフォームの環境設定に従います。テストをカスタマイズするには、［ファイル］>［環境設定］>［プラットフォーム］>［管理図ビルダー］を選択します。

要約表には、指定によっては、以下の情報も表示されます。

列 ［工程変数］として指定した列。［工程変数］と［グループ変数］の列の一意の組み合わせごとに、1つの行があります。工程変数が1つしかない場合、この列は表示されません。

グループ列 ［グループ変数］に指定した列ごとに、1つのレポート列があります。グループ変数列の水準がリストされ、工程の名前とグループ変数列の値の一意の組み合わせに対し、1つの行が表示されます。

ばらつき 次の列で構成されます。

安定比 工程の安定性に関する指標。安定比は、次のように定義されます。

$$(\text{全体シグマ} / \text{群内シグマ})^2$$

安定した工程の安定比は、1 に近い値を取ります。大きい値ほど、工程が安定していないことを示唆しています。

群内シグマ 群内変動に基づいて計算した標準偏差の推定値。この推定値は、短期における工程の変動（ばらつき）を見ています。その計算方法は、指定した管理図の種類によって異なります。統計的詳細については、『品質と工程』の「工程能力」章を参照してください。［平均ではなく中央値を使用］オプションを選択した場合、群内シグマは「平均ではなく中央値を使用」（318 ページ）で説明されているように計算されます。

全体シグマ すべての測定値から通常の方法で算出された、標準偏差の推定値。

要約 以下の列で構成されます。

中央線 （起動ウィンドウで［平均ではなく中央値を使用］を選択した場合、または限界のテーブルを使って中央線の値を読み込んだ場合のみ表示されます。）「中央線」にリストされた値は、管理図の計算で中央線として使用されます。

– 起動ウィンドウで［平均ではなく中央値を使用］を選択した場合は、中央値が表示されます。
– 限界のテーブルから中央値の値を読み込んだ場合は、その読み込んだ値が表示されます。

平均 すべての測定値から計算された平均。

度数 測定値の個数。

サブグループ サブグループの個数。

管理図の警告 8つのWestern Electricルールを含むさまざまなテストで警告が生じたサブグループの情報。以下の警告に関する説明において、I-MR管理図では1つの測定値が標本サイズが1のサブグループとみなされます。また、標準偏差の推定値には、「群内シグマ」が使われます。デフォルトの要約表には、「警告率」、「テスト 1」、「最後の警告」の列だけが表示されます。

警告率 ［テストの選択］オプションで選択したテストのいずれかで警告が生じたサブグループの個数（「すべての警告」に表示されている個数）を、すべてのサブグループの個数で割ったもの。

すべての警告 （テストの列が複数表示されている場合にのみ表示されます。）［テストの選択］オプションで選択したテストのいずれかで警告が生じたサブグループの個数。カウントの対象になりうるテストには、8 つの特殊原因テストと［範囲の限界外］テストがあります。

ヒント：「工程のスクリーニング」プラットフォームで行われる特殊原因テストは、「管理図ビルダー」プラットフォームの環境設定に従います。テストをカスタマイズするには、［ファイル］>［環境設定］>［プラットフォーム］>［管理図ビルダー］を選択します。

テスト 1 該当する 1 個の点が中央線から ±3 標準偏差より外にある場合。その 1 点のサブグループに警告を出します。

テスト 2 9 個以上の連続する点が中央線に対して同じ側にある場合。9 番目の点のサブグループに対して警告を出します。

テスト 3 6 個以上の連続する点が単調増加または単調減少している場合。6 番目の点のサブグループに対して警告を出します。

テスト 4 14 個の連続する点が増加と減少を交互に繰り返している場合。14 番目の点のサブグループに対して警告を出します。

テスト 5 連続する 3 個の点のうち 2 個が、中央線に対して同じ側にあり、かつ、中央線から ±2 標準偏差より外にある場合。その区間の外にある 2 番目の点のサブグループに対して警告を出します。

テスト 6 連続する 5 個の点のうち 4 個が、中央線に対して同じ側にあり、かつ、中央線から ±1 標準偏差より外にある場合。その区間の外にある 4 番目の点のサブグループに対して警告を出します。

テスト 7 連続した 15 個の点が中央線から ±1 標準偏差の区間内にある場合。15 番目の点のサブグループに警告を出します。

テスト 8 8 個の連続した点が中央線から ±1 標準偏差より外にある場合。8 番目の点のサブグループに警告を出します。

範囲の限界外 R、S、MR の各管理図の計算において上側管理限界を超えるサブグループの数。

最後の警告 特殊原因テストまたは［範囲の限界外］テストのいずれかで最後に警告が生じたサブグループの位置。値は、最後のサブグループから数えて何番目かを示します。

工程能力 （一部の工程に対して［仕様限界］を指定した場合にのみ表示されます。）以下のオプションがあります。

仕様限界外の度数 仕様限界の外にある測定値の個数。

仕様限界外の割合 仕様限界の外にある測定値の割合。

最後の仕様限界外 仕様限界外になっている最後の測定値の位置。値は、最後から数えて何番目かを示します。

Cpk 正規分布に従うことを仮定し、群内シグマに基づいて求めた工程能力指数。統計的詳細については、『品質と工程』の「工程能力」章を参照してください。

Ppk 正規分布に従うことを仮定し、全体シグマに基づいて求めた工程能力指数。統計的詳細については、『品質と工程』の「工程能力」章を参照してください。

シフトの大きさと位置 （「工程のスクリーニング」の赤い三角ボタンのメニューから［シフトの検出］を選択した場合にのみ表示されます。）1群内シグマを超えるシフトを検出します。そのアルゴリズムでは、外れ値を除外し、また、個々の測定値に対する指数加重移動平均（EWMA; Exponentially-Weighted Moving Average）で平滑化する方法を採用しています。

- 外れ値を除外することで、外れ値がシフトとみなされてしまうことがなくなります。起動ウィンドウにある「外れ値の閾値」（デフォルトでは5）にて、外れ値を除外する感度を変更できます。「外れ値の閾値」に群内シグマを掛けたものよりも両隣の測定値が離れている場合、該当の測定値を、最も近い測定値から1群内シグマだけ離れた値に置き換えられます。
- 指数加重移動平均（EWMA）が、時間順に並べたサブグループ平均に対して、および、時間の逆順に並べたサブグループ平均に対して計算されます。デフォルトでは、このEWMAの計算におけるラムダが0.3に設定されています。
- 連続したEWMAの値のなかで、1群内シグマを超える最大の正および負のシフトが特定されます。
- これらのシフトの絶対値を群内シグマで割った値が、「最大の上昇シフト」および「最大の下降シフト」として表示されます。
- 「上昇シフトの位置」と「下降シフトの位置」は、これらのシフトが生じている最初のサブグループの位置です。

最大の上昇シフト 1群内シグマを超えるもののなかで最大の上昇シフトが、群内シグマの何倍になっているかが表示されます。

上昇シフトの位置、または上昇シフト < 時間変数 > 最大の上昇シフトを持つサブグループの位置。時間変数を指定した場合、要約表における列の名前は「上昇シフト < 時間変数 >」となり、シフトの位置が時間変数を使って表されます。

最大の下降シフト 1群内シグマを超えるもののなかで最大の下降シフトが、群内シグマの何倍になっているかが表示されます。

下降シフトの位置、または下降シフト < 時間変数 > 最大の下降シフトを持つサブグループの位置。時間変数を指定した場合、要約表における列の名前は「下降シフト < 時間変数 >」となり、シフトの位置が時間変数を使って表されます。

「工程のスクリーニング」プラットフォームのオプション

「工程のスクリーニング」の赤い三角ボタンのメニューには、表示形式をカスタマイズするオプションや、計算された統計量を保存するオプションがあります。

要約 要約表を表示します。「「工程のスクリーニング」レポート」（319ページ）を参照してください。

検索と選択 起動ウィンドウで［工程変数］または［グループ変数］に指定した列を対象に、任意の文字列を検索できます。それらの変数ごとにテキスト入力ボックスが表示されます。実行すると、要約表で検索文字列を含む工程が選択されます。

選択した項目の簡易グラフ 要約表で選択した各工程に対して、「選択した項目のグラフ」レポートとして小さなグラフを描きます。このグラフは、多くの工程を一度に比較できます。各グラフは、起動ウィンドウで指定された順序で並びます。

選択した項目の管理図 要約表で選択した各工程を描いた管理図を含む「管理図ビルダー」ウィンドウが開きます。各管理図は、「工程のスクリーニング」の起動ウィンドウで指定された順序で並びます。

選択した項目の工程能力分析 「工程能力」ウィンドウが開き、要約表で選択した各工程に関する「各列の詳細レポート」がそこに表示されます。仕様限界が指定されていない工程を選択した場合は、「仕様限界」ウィンドウが開きます。このウィンドウでは、データテーブルを選択するか、値を直接入力することで仕様限界を指定できます。

「工程能力分析」では、「測定値は正規分布に従う」と仮定したときの工程能力指数が算出されます。これらの工程能力指数では、「工程のスクリーニング」起動ウィンドウで選択した管理図の種類に応じて群内シグマが計算されます。

- I-MR管理図の場合は移動範囲
- XBar-R管理図の場合は範囲の平均
- XBar-S管理図の場合は不偏標準偏差の平均

選択した項目を色付け 要約表で選択した行に、指定の色をつけます。

テストの表示 「工程のスクリーニング」レポートの表に、［テストの選択］オプションで選択した特殊原因テストの結果を表示します。

テストの選択 「警告率」と「すべての警告」の計算に含めたいテストを選択します。

ヒント：複数のテストを選択するには、Alt キーを押しながら「工程のスクリーニング」の赤い三角ボタンをクリックし、メニューを開きます。

シフトの検出 外れ値を除外した後、シフトを検出します。「シフトの大きさと位置」（322ページ）を参照してください。

最大の上昇シフト 要約表に、「最大の上昇シフト」と「上昇シフトの位置」という列が追加されます。「最大の上昇シフト」は、1群内シグマを超える上昇シフトのなかで最大のものです。「シフトの大きさと位置」（322 ページ）を参照してください。

最大の下降シフト 要約表に、「最大の下降シフト」と「下降シフトの位置」という列が追加されます。「最大の下降シフト」とは、1群内シグマを超える下降シフトのなかで最大のものです。「シフトの大きさと位置」（322 ページ）を参照してください。

シフトグラフ （「シフトの閾値」を超えるシフトがある場合にのみ使用可能。）「シフトの閾値」で指定した群内シグマの倍数（デフォルトは3倍）を超えるすべてのシフトの時点をプロットしたもの。緑色のマーカーは上昇シフト、赤いマーカーは下降シフトを意味します。マーカーは、シフトのうちでも局所的なピークになっている時点に配置されます。

どの工程にシフトが生じているかを特定するには、点を選択し、[工程の選択]をクリックします。そうすると、要約表において、該当する工程が選択されます。なお、「シフトの閾値」で指定した倍数を超えるシフトが1つもない工程は、「シフトグラフ」が描かれません。

メモ：要約表にある「最大の上昇シフト」と「最大の下降シフト」が、「シフトの閾値」で指定した群内シグマのユニット数を下回る場合、そのシフトは「シフトグラフ」にプロットされません。「「工程のスクリーニング」の別例」（325ページ）を参照してください。

簡易グラフでシフトを表示 （簡易グラフがレポートウィンドウに追加されている場合のみ。）簡易グラフに、シフトの位置を示す緑色と赤の縦線を表示します。

工程性能グラフ 安定性（stability）と工程能力（capability; 仕様限界への適合性）を示した4象限のグラフが描かれます。このグラフにプロットされる各点は、仕様限界が指定されている各工程を表します。横軸はその工程の安定比、縦軸はPpkで表した工程能力指数です。デフォルトのグラフは、以下の境界に従って4つの象限に分割されています。

- 安定比が1.5を超える工程は、「不安定」。
- Ppkが1.33より小さい工程は、「不適合」。

グラフ上の点を選択すると、要約表でも該当する工程が選択されます。

性能グラフの境界 ウィンドウが開き、工程性能グラフの境界を定義している安定比とPpkの値を変更できます。

ヒント：安定比とPpkのデフォルトの値を変更したい場合は、[ファイル]>[環境設定]で値を指定します。

要約テーブルの保存 要約表に表示されるすべての情報が、「工程の要約」というデータテーブルに保存されます。仕様限界が指定されている工程が1つでもある場合は、「工程の要約」データテーブルに仕様限界の詳細も含まれます。

詳細テーブルの保存 管理図の計算に関する詳細情報が、「工程の詳細」というデータテーブルに保存されます。このデータテーブルには、工程変数とグループ変数の各組み合わせごとに1つの行が含まれ、各行に次のような値が表示されます。

- サブグループの標本統計量の値。
- 管理限界。
- サブグループの標本サイズ。
- 警告が生じた場合はそれを示す値。特殊原因テストで警告が生じた場合は、テストの番号が表示されます。[範囲の限界外]テストで警告が生じた場合は、「R」の文字が表示されます。

選択した詳細の保存 「工程の詳細」というデータテーブルに、要約表で選択した行の情報が保存されます。情報の種類は、［詳細テーブルの保存］で保存されるものと同じです。

以下のオプションについて詳しくは、『JMPの使用法』の「JMPのレポート」章を参照してください。

ローカルデータフィルタ 現在のレポートにおいてデータをフィルタリングするローカルデータフィルタを表示するか、非表示にします。

やり直し 分析を繰り返すオプションや、やり直すオプションを含みます。また、［自動再計算］オプションを選択すると、このオプションに対応しているプラットフォームにおいて、データテーブルに加えた変更が、該当するレポートに即座に反映されるようになります。

スクリプトの保存 レポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。

Byグループのスクリプトを保存 By変数の全水準に対するレポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。起動ウィンドウでBy変数を指定した場合のみ使用可能です。

「工程のスクリーニング」の別例

以下の例では、**「Consumer Prices.jmp」**データテーブルを例に、グループ変数列の使い方とシフトグラフについて説明します。

消費者物価指数をまとめたこのデータテーブルには、17種類の商品に関する月次データが含まれています。期間は商品によって異なります。17種類の製品名は、**「系列」**という列に含まれています。この**「系列」**という列をグループ変数に指定すれば、製品別に集計できます。

1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Consumer Prices.jmp」を開きます。
2. ［**分析**］>［**スクリーニング**］>［**工程のスクリーニング**］を選択します。
3. **「価格」**を選択し、［**工程変数**］をクリックします。
4. **「系列」**を選択し、［**グループ変数**］をクリックします。

   これで、**「系列」**の各水準が個別の工程として扱われるようになります。
5. **「日付」**を選択し、［**時間**］をクリックします。
6. **「管理図の種類」**として［XBar-R管理図］を指定します。
7. **「サブグループの標本サイズ」**を「3」に設定します。

   月次のデータなので、3という標本サイズは四半期に相当します。
8. ［**OK**］をクリックします。
9. Altキーを押しながら「工程のスクリーニング」の赤い三角ボタンをクリックします。

   メニューのオプションをすべて含んだウィンドウが開きます。このウィンドウは、複数のオプションを一度に選択したいときに便利です。

10. [**最大の上昇シフト**]、[**最大の下降シフト**]、[**シフトグラフ**] を選択します。
11. [**OK**] をクリックします。

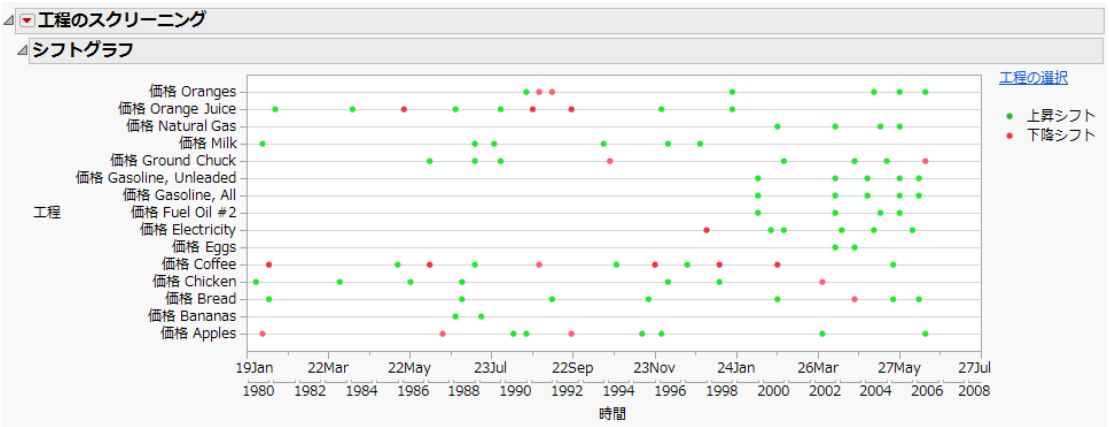
    「最大の上昇シフト」、「上昇シフト 日付」、「最大の下降シフト」、「下降シフト 日付」が要約表に加わります。これらのシフトは、**1** 群内シグマを超えるシフトのうちで最大のものです。シフトの位置は、時間変数として指定された**「日付」**で表されています。「シフトの大きさと位置」(322 ページ)を参照してください。

    「シフトグラフ」も表示されます。「シフトグラフ」には、**「シフトの閾値」**（デフォルトは 3）に群内シグマをかけた値を超えるシフトがすべてプロットされています。「シフトグラフ」（324 ページ）を参照してください。緑色は上昇シフト、赤色は下降シフトを示します。

    要約表の**「Gasoline, All」**を見ると、「最大の上昇シフト」と「最大の下降シフト」の両方に数値があります。ただし、「最大の下降シフト」は 1.8296 となっており、3 を下回っています。「シフトグラフ」には、群内シグマの 3 倍以上となっているシフトしか描かれないため、**「Gasoline, All」**の「最大の下降シフト」は「シフトグラフ」にはプロットされません。

    また、**「Tomatoes」**は「シフトグラフ」に含まれていません。なぜなら、**「Tomatoes」**には、群内シグマの 3 倍以上のシフトが 1 つも検出されなかったからです。
12. 「シフトグラフ」の横軸をダブルクリックすると、「X 軸の指定」ウィンドウが開きます。
13. 「目盛り / 棒の間隔」パネルで、**「補助目盛りの数」**を「1」に設定します。
14. **「ラベルの階層」**を「2」に設定します。
15. [**OK**] をクリックします。

図 18.7 シフトグラフ

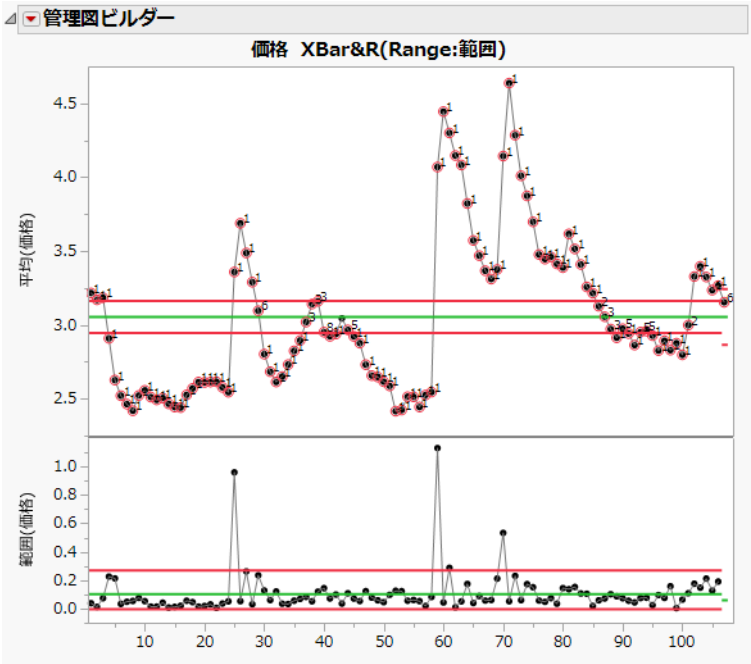


ほとんどの系列で上昇シフトが検出されていますが、「価格 Coffee」は上昇シフトと下降シフトを順番に繰り返しています。この系列について調べるため、管理図を作成してみましょう。

16. 「シフトグラフ」で「価格 Coffee」の右側にあるいずれかの点を選択し、[**工程の選択**] をクリックします。

    これで、要約表でも**「Coffee」**の行が選択されます。
17. 「工程のスクリーニング」の赤い三角ボタンをクリックし、[**選択した項目の管理図**] を選択します。

図18.8「Coffee」の管理図



管理図から、「シフトグラフ」で検出された上下のシフトがどうなっているかが分かります。

要約表を見ると、1994年9月で最大の上昇シフト（群内シグマの25.399倍）が検出されています。これは、図18.8の管理図で59の位置にあるサブグループです。さらに要約表からは、1981年3月に最大の下降シフト（群内シグマの9.1674倍）が生じていることがわかります。これは、管理図で5の位置にあるサブグループです。

なお、「工程のスクリーニング」におけるシフトの検出では、外れ値が除外され、また、指数加重移動平均（EWMA）法で平滑化された系列が使われています。そのため、検出されるシフトは、Shewhart管理図の見た目で判断されるシフトとは必ずしも一致しません。

統計的詳細

中央値を使ってシグマの推定値を求めるときの係数

［平均ではなく中央値を使用］を選択した場合、シグマの推定値は、標準正規分布から生成された確率変数の、範囲の中央値、または、標準偏差の中央値にもとづいて計算されます。下の表は、モンテカルロシミュレーションで得た、それらの係数です。

正規分布から抽出したサイズが n のサブグループの場合は、次のことがあてはまります。

- 範囲の中央値は、*d2_Median* σにほぼ等しくなります。ここで、*d2_Median* は、n に対応する係数です。

- 標準偏差の中央値は、*c4_Median* σにほぼ等しくなります。ここで、*c4_Median*は、*n*に対応する係数です。

表18.1 範囲の中央値および標準偏差の中央値対する係数

n	d2_Median	c4_Median
2	0.953	0.675
3	1.588	0.833
4	1.978	0.888
5	2.257	0.917
6	2.471	0.933
7	2.646	0.944
8	2.792	0.952
9	2.915	0.959
10	3.024	0.963
11	3.118	0.967
12	3.208	0.969
13	3.286	0.972
14	3.357	0.975
15	3.422	0.976
16	3.483	0.978
17	3.539	0.979
18	3.590	0.980
19	3.640	0.981
20	3.685	0.982
21	3.731	0.983
22	3.770	0.984
23	3.811	0.984
24	3.846	0.985
25	3.883	0.986

第19章

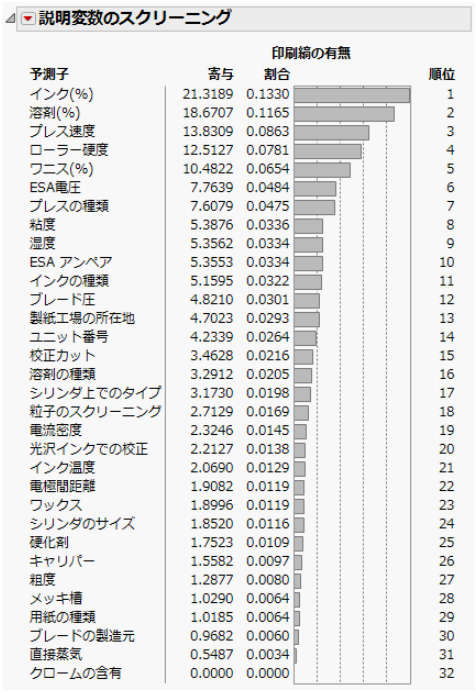
説明変数のスクリーニング

多数の説明変数の中から有意な効果を探し出す

部品や工程、標本などの測定値を数百あるいは数千も含むような大規模なデータを分析するには、革新的な統計手法が必要です。「説明変数のスクリーニング」プラットフォームは、多くの説明変数をスクリーニングし、応答の予測能力が高いものを選び出します。「説明変数のスクリーニング」を使うと、たとえば、ある症状を持つ患者と持たない患者から標本を採り、試験した数千というバイオマーカーの中から発症の予測に有効なものを特定できます。

「説明変数のスクリーニング」は、「応答のスクリーニング」とは異なります。「応答のスクリーニング」は、説明変数を1変数ずつ検定します。「説明変数のスクリーニング」では、ブートストラップ森を使い、応答変数に対する各説明変数の寄与度を評価します。このブートストラップ森は、モデルのなかに複数の説明変数を含んでいます。「説明変数のスクリーニング」では、たとえ1変数だけでは弱かったとしても、他の説明変数と組み合わされたときに強い予測能力を持つような説明変数を特定します。「応答のスクリーニング」についての詳細は、「応答のスクリーニング」章（281ページ）を参照してください。

図19.1「説明変数のスクリーニング」レポートの例



説明変数のスクリーニング

印刷縞の有無

予測子	寄与	割合	順位
インク(%)	21.3189	0.1330	1
溶剤(%)	18.6707	0.1165	2
プレス速度	13.8309	0.0863	3
ローラー硬度	12.5127	0.0781	4
ワニス(%)	10.4822	0.0654	5
ESA電圧	7.7639	0.0484	6
プレスの種類	7.6079	0.0475	7
粘度	5.3876	0.0336	8
湿度	5.3562	0.0334	9
ESA アンペア	5.3553	0.0334	10
インクの種類	5.1595	0.0322	11
ブレード圧	4.8210	0.0301	12
製紙工場の所在地	4.7023	0.0293	13
ユニット番号	4.2339	0.0264	14
校正カット	3.4628	0.0216	15
溶剤の種類	3.2912	0.0205	16
シリンダ上でのタイプ	3.1730	0.0198	17
粒子のスクリーニング	2.7129	0.0169	18
電流密度	2.3246	0.0145	19
光沢インクでの校正	2.2127	0.0138	20
インク温度	2.0690	0.0129	21
電極間距離	1.9082	0.0119	22
ワックス	1.8996	0.0119	23
シリンダのサイズ	1.8520	0.0116	24
硬化剤	1.7523	0.0109	25
キャリパー	1.5582	0.0097	26
粗度	1.2877	0.0080	27
メッキ槽	1.0290	0.0064	28
用紙の種類	1.0185	0.0064	29
ブレードの製造元	0.9682	0.0060	30
直接蒸気	0.5487	0.0034	31
クロームの含有	0.0000	0.0000	32

「説明変数のスクリーニング」プラットフォームの概要

「説明変数のスクリーニング」は、多数の候補の中から有意な説明変数を探したいときに便利です。数百個の説明変数があり、そこから応答変数を予測するのに役立つという点で有意なものだけを選択したい場合に、「説明変数のスクリーニング」が役立ちます。

「説明変数のスクリーニング」プラットフォームは、ブートストラップ森を使い、応答変数に対する潜在的な説明変数を選択します。各応答変数ごとに、100のディシジョンツリー（決定木）から構成されるブートストラップ森モデルがあてはめられます。ブートストラップ森モデルから、各説明変数の寄与度が算出されます。なお、ブートストラップ森では乱数が利用されているため、分析をを再実行すると、説明変数の寄与度が多少異なる値になります。ディシジョンツリーの詳細については、『予測モデルおよび発展的なモデル』の「パーティション」章を参照してください。

「説明変数のスクリーニング」の例

「Bands Data.jmp」データテーブルには、印刷会社の輪転グラビア印刷機から得られた測定データが含まれています。データセットには539個のレコードと38個の変数が含まれています。応答Yは**「印刷縞の有無」**の列で、「band」と「noband」の値を取ります。ここでの関心は、どの印刷設定が印刷結果に最も寄与しているかを理解することです。

1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Bands Data.jmp」を開きます。
2. ［**分析**］>［**スクリーニング**］>［**説明変数のスクリーニング**］を選択します。
3. **「印刷縞の有無」**を選択し、［**Y, 応答変数**］をクリックします。
4. グループ化されている**「粒子のスクリーニング」**から**「クロームの含有」**までの列を選択し、［**X**］をクリックします。
5. ［**OK**］をクリックします。

図19.2 列の寄与の順位付け

説明変数のスクリーニング

印刷縞の有無

予測子	寄与	割合	順位
インク(%)	21.3189	0.1330	1
溶剤(%)	18.6707	0.1165	2
プレス速度	13.8309	0.0863	3
ローラー硬度	12.5127	0.0781	4
ワニス(%)	10.4822	0.0654	5
ESA電圧	7.7639	0.0484	6
プレスの種類	7.6079	0.0475	7
粘度	5.3876	0.0336	8
湿度	5.3562	0.0334	9
ESA アンペア	5.3553	0.0334	10
インクの種類	5.1595	0.0322	11
ブレード圧	4.8210	0.0301	12
製紙工場の所在地	4.7023	0.0293	13
ユニット番号	4.2339	0.0264	14
校正カット	3.4628	0.0216	15
溶剤の種類	3.2912	0.0205	16
シリンダ上でのタイプ	3.1730	0.0198	17
粒子のスクリーニング	2.7129	0.0169	18
電流密度	2.3246	0.0145	19
光沢インクでの校正	2.2127	0.0138	20
インク温度	2.0690	0.0129	21
電極間距離	1.9082	0.0119	22
ワックス	1.8996	0.0119	23
シリンダのサイズ	1.8520	0.0116	24
硬化剤	1.7523	0.0109	25
キャリパー	1.5582	0.0097	26
粗度	1.2877	0.0080	27
メッキ槽	1.0290	0.0064	28
用紙の種類	1.0185	0.0064	29
ブレードの製造元	0.9682	0.0060	30
直接蒸気	0.5487	0.0034	31
クロームの含有	0.0000	0.0000	32

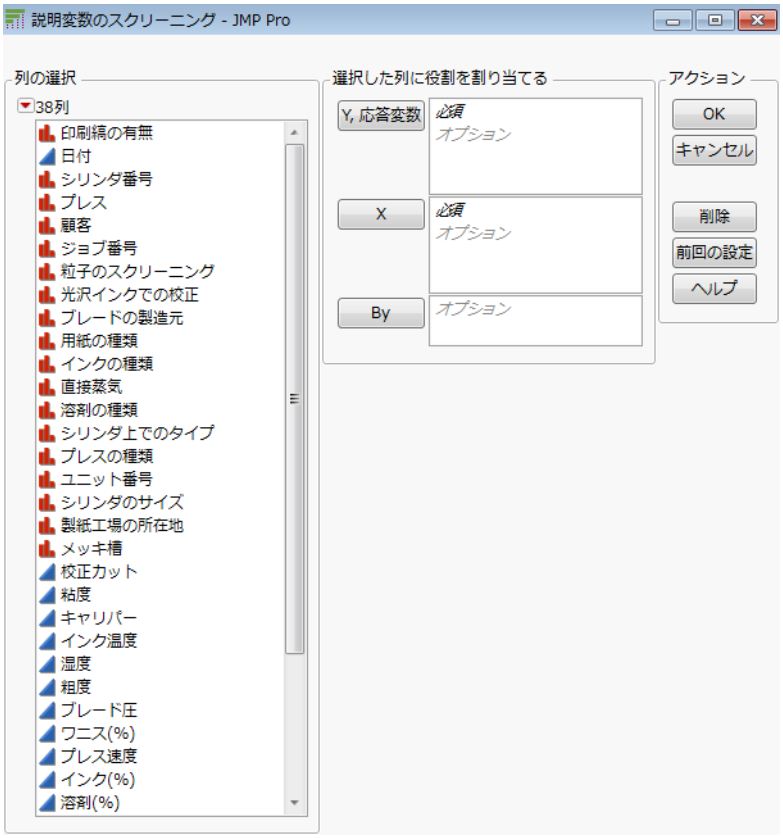
メモ: この分析で使用されるブートストラップ森では乱数が使われています。そのため、実際の結果は図19.2とは多少異なったものになります。「ブートストラップ森」（107ページ）を参照してください。

説明変数は、ブートストラップ森での寄与度の順に並び替えられています。寄与が大きな説明変数は、応答変数を予測する変数の有力な候補になります。

「説明変数のスクリーニング」プラットフォームの起動

「説明変数のスクリーニング」プラットフォームを起動するには、[分析] > [スクリーニング] > [説明変数のスクリーニング] を選択します。

図19.3「説明変数のスクリーニング」起動ウィンドウ



Y, 目的変数　応答の列。

X　説明変数の列。

By　別々に分析を行いたいときに、そのグループ分けをする変数を指定します。指定された列の各水準ごとに、別々に分析が行われます。各水準の結果は別々のレポートに表示されます。複数のBy変数を割り当てた場合、それらのBy変数の水準の組み合わせごとに別々のレポートが作成されます。

「説明変数のスクリーニング」レポート

「説明変数のスクリーニング」のレポート（図19.2）には、説明変数と、その寄与度（contribution）および順位が表示されます。寄与度が高い説明変数は、Yを予測するのに重要であると考えられます。

「寄与」列には、ブートストラップ森モデルへの各説明変数の寄与度が表示されます。レポートの「割合」列には、各説明変数における寄与度の割合（percent contribution）が表示されます。

「説明変数のスクリーニング」レポートで、重要な説明変数をクリックして選ぶことができます。重要な説明変数をクリックすると、データテーブルでも該当する列が選択状態になりますので、別のプラットフォームの起動ウィンドウにこれらの列を簡単に指定できます。「説明変数のスクリーニング」は、このような操作でモデル作成に役立ちます。

「説明変数のスクリーニング」プラットフォームのオプション

以下のオプションについて詳しくは、『JMPの使用法』の「JMPのレポート」章を参照してください。

やり直し 分析を繰り返すオプションや、やり直すオプションを含みます。また、［自動再計算］オプションを選択すると、このオプションに対応しているプラットフォームにおいて、データテーブルに加えた変更が、該当するレポートに即座に反映されるようになります。

スクリプトの保存 レポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。

Byグループのスクリプトを保存 By変数の全水準に対するレポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。起動ウィンドウでBy変数を指定した場合のみ使用可能です。

第 20 章

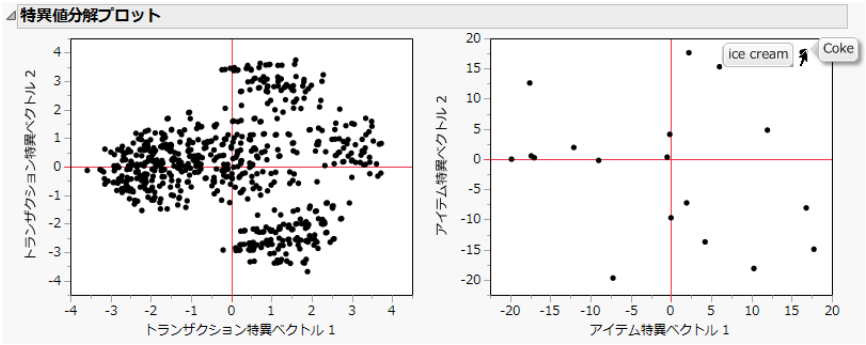
JMP PRO アソシエーション分析
マーケットバスケット分析の実行

「アソシエーション分析」プラットフォームは、JMP Prode のみ利用できます。

アソシエーション分析では、互いに関係のあるアイテムを特定できます。この手法は、トランザクションデータ（**マーケットバスケット**）の分析によく使われています。1つの取引で同時に生じることが多いアイテム（商品）を特定するのに役立ちます。たとえば、食料品店やオンライン通販業者は、アソシエーション分析に基づいて、一緒に購入される傾向にある製品を戦略的に分類し、推奨します。

アソシエーション分析は、依存関係にあるイベントや関連性のあるイベントの特定にも利用できます。たとえば、ほぼ同時期に故障する傾向にある自動車のパーツなどを特定できます。自動車の各点検を1つのマーケットバスケットとみなし、各点検で見つかった故障パーツの関連性を分析できます。

図 20.1 特異値分解プロットの例



JMP PRO「アソシエーション分析」プラットフォームの概要

「アソシエーション分析」プラットフォームは、独立したイベント（**トランザクション**）に現れるアイテムの関連性を調べます。アソシエーション分析で分析対象となる基本単位は、「**アイテム**」（item）です。アイテムは、たとえば製品、Webページ、サービスなどが相当します。また、1つまたは複数のアイテムを含む集合を、「**アイテム集合**」（item set）と言います。

2つのアイテム集合の関係は、**アソシエーションルール**（association rule）によって記述されます。アソシエーションルールは、**条件**（condition）のアイテム集合と**帰結**（consequent）のアイテム集合から構成されます。条件に含まれるアイテムは、「**先行**」（antecednt）とも呼ばれています。アソシエーション分析は、ある条件がトランザクションで生じている場合に、ある帰結がどれぐらいの頻度で含まれるかを予測します。関連が強いアソシエーションルールを探し出すことは有用でしょう。アソシエーションルールの強さを表すのに、次の3つの指標が使われています。

- **支持度**（support）は、該当のアイテム集合を含んでいるトランザクションの割合です。支持度が高いということは、そのアイテム集合が頻繁に現れることを意味します。
- **信頼度**（confidence）は、条件アイテム集合を含むトランザクションのうち、帰結アイテム集合を含むものの割合を指します。信頼度は、関係の強さや予測能力の強さを表します。
- **リフト値**（lift）は、条件アイテム集合と帰結アイテム集合がそれぞれ独立してトランザクションに現れると仮定した場合の信頼度の期待値に対する、観測された信頼度の比です。リフト値は、帰結アイテム集合がどれだけ条件アイテム集合の存在に依存しているかを表します。リフト値の最小値は0です。
  - リフト値が1より小さい場合、条件と帰結が一緒に生じる頻度が、偶然に期待される頻度よりも低いことから、条件と帰結が反発していると考えられます。
  - リフト値が1に近い場合、条件と帰結は、偶然に期待される頻度と同じ頻度で生じていると考えられます。
  - リフト値が1より大きい場合、条件と帰結との間には関連があると考えられます。条件が生じているときに、帰結が偶然に期待される頻度より高い頻度で一緒に生じるためです。

これらの指標の詳細については「アソシエーション分析のパフォーマンス指標」（348ページ）を参照してください。

「アソシエーション分析」プラットフォームでは、特異値分解（SVD; Singular Value Decomposition）も実行できます。特異値分解は、アソシエーション分析とは別の枠組みにもとづいた分析です。特異値分解により、類似のトランザクションやアイテムを見つけることができます。アソシエーション分析の結果を補完するかたちで、別の観点からデータを見ることができます。

アソシエーション分析の詳細については、Hastie et al.（2009）およびShmueli et al.（2010）を参照してください。特異値分解の詳細については、Jolliffe（2002）を参照してください。

JMP PRO「アソシエーション分析」プラットフォームの例

この例では、食料品店におけるトランザクションデータをまとめた「Grocery Purchases.jmp」データテーブルを使用します。このデータテーブルには、1001 人の顧客が購入した商品がリストされています。顧客には、それぞれ一意の顧客 ID が割り当てられています。消費者行動のパターンを特定する目的で、商品間の関連性を探ってみましょう。

1. ［ヘルプ］>［サンプルデータライブラリ］を選択し、「Grocery Purchases.jmp」を開きます。
2. ［分析］>［スクリーニング］>［アソシエーション分析］を選択します。
3. 「商品」を選択し、［アイテム］をクリックします。
4. 「顧客 ID」を選択し、［ID］をクリックします。
5. ［OK］をクリックします。

図 20.2「アソシエーション分析」レポート

アソシエーション分析

高頻度のアイテム集合

ルール

ルール

条件	帰結	信頼度	リフト 値
peppers	apples	44%	1.4
sardines	apples	43%	1.357
steak	apples	52%	1.657
avocado, baguette	apples	52%	1.661
corned beef, herring	apples	46%	1.47
corned beef, olives	apples	49%	1.56
corned beef, steak	apples	77%	2.445
herring, olives	apples	45%	1.42
corned beef, herring, olives	apples	54%	1.72
steak	apples, corned beef	45%	2.979
herring, olives	apples, corned beef	43%	2.823
corned beef, olives	apples, herring	46%	2.932
corned beef, herring	apples, olives	44%	2.801
avocado	artichoke	58%	1.908
ham	artichoke	42%	1.377
Heineken	artichoke	42%	1.378
avocado, baguette	artichoke	52%	1.71
avocado, crackers	artichoke	69%	2.275

「ルール」という表における該当箇所を見ると、アボカド（avocado）を購入した顧客の58%がアーティチョーク（artichoke）も購入していることがわかります。そのリフト値は1.908であり、依存関係があるらしいことを示しています。次に、トランザクションのなかで、アボカドとアーティチョークがどれぐらい一緒に購入されているかを考えてみましょう。

6. 「高頻度のアイテム集合」の開閉アイコンをクリックします。

図20.3「高頻度のアイテム集合」レポート

高頻度のアイテム集合

アイテム集合	支持度	N個のアイテム
{Heineken}	60%	1
{crackers}	49%	1
{herring}	49%	1
{olives}	47%	1
{bourbon}	40%	1
{baguette}	39%	1
{corned beef}	39%	1
{crackers, Heineken}	37%	2
{avocado}	36%	1
{soda}	32%	1
{chicken}	31%	1
{apples}	31%	1
{ice cream}	31%	1
{artichoke}	30%	1
{ham}	30%	1
{Coke}	30%	1
{peppers}	30%	1
{sardines}	30%	1
{Heineken, herring}	29%	2
{turkey}	28%	1
{baguette, Heineken}	26%	2
{Heineken, soda}	26%	2
{herring, olives}	26%	2
{artichoke, Heineken}	25%	2

「高頻度のアイテム集合」レポートからは、顧客の36%がアボカドを購入していることが分かります。先ほどの図20.2の「ルール」からは、アボカドを購入した顧客のうち58%がアーティチョークも購入していることが分かります。この割合は大きいので、アボカドとアーティチョークを近い売り場に配置するのがよいかもしれません。

また、リフト値が最も高いアソシエーションルールについても見てみましょう。

7. 「ルール」表内を右クリックし、［**列の値で並べ替え**］を選択します。

   「列の選択」ウィンドウが開きます。

8. **「リフト値」**を選択し、［**OK**］をクリックします。

   「ルール」表がリフト値の降順で並べ替えられます。2番目のアソシエーションルールを見ると、リフト値は6.912、信頼度は97%です。この場合、条件集合（{Coke, Heineken, sardines}）と帰結集合（{chicken, ice cream}）の支持度が適切な値であることを確認する必要があります。

9. 「高頻度のアイテム集合」レポート内を右クリックし、［**列の値で並べ替え**］を選択します。

   「列の選択」ウィンドウが開きます。

10. **「アイテム集合」**を選択し、［昇順］チェックボックスをオンにします。

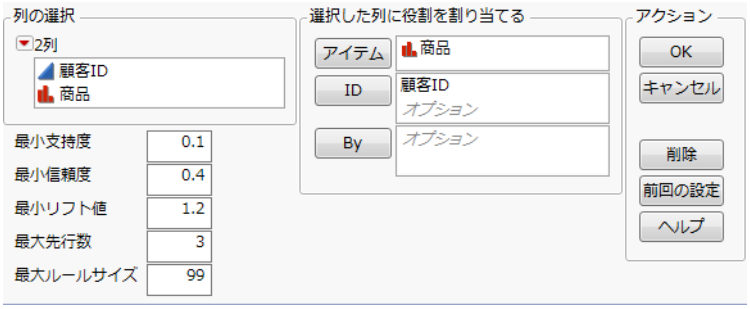
11. ［**OK**］をクリックします。

    「高頻度のアイテム集合」がアイテム集合のアルファベット順に並べ替えられます。リストをスクロールして条件アイテム集合の{Coke, Heineken, sardines}を見ると、支持度が12%であること、および帰結アイテム集合の{chicken, ice cream}の支持度が14%であることがわかります。このアソシエーションルールでは、リフト値が高いものの、トランザクションの数は最初に検討したアソシエーションルールより少なくなっています。

JMP PRO「アソシエーション分析」プラットフォームの起動

「アソシエーション分析」プラットフォームを起動するには、[分析] > [スクリーニング] > [アソシエーション分析] を選択します。

図20.4「アソシエーション分析」起動ウィンドウ



アイテム　分析対象とするアイテムデータを含むカテゴリカル列。

ID　アイテムの属するトランザクションのIDを含む列。

By　By変数の水準ごとに個別のレポートが作成されます。複数のBy変数を割り当てた場合、それらのBy変数の水準の組み合わせごとに個別のレポートが作成されます。

最小支持度　アイテム集合が出現する割合の最小値。0～1の間で指定します。支持度がこの値以上であるアイテム集合のみ、分析に含められます。

最小信頼度　条件アイテム集合を含むトランザクションのうち、帰結アイテム集合が生じるものの割合の最小値。0～1の間で指定します。信頼度がこの値以上であるアソシエーションルールのみ、レポートに含められます。

最小リフト値　依存度の最小値。0以上の値を指定します。リフトがこの値以上であるアソシエーションルールのみ、レポートに含められます。

最大先行数　条件アイテム集合を構成するアイテムの最大数。アイテム数がこの数を超えるアソシエーションルールは、分析から除外されます。

最大ルールサイズ　条件アイテム集合と帰結アイテム集合の和集合に現れるアイテムの最大数。アイテムの和がこの数を超えるアソシエーションルールは、分析から除外されます。

メモ: 起動ウィンドウで、最小支持度、最大先行数、最大ルールサイズの各オプションを使用すると、大規模なデータセットでの計算時間を短縮できます。これらの指標の詳細については「「アソシエーション分析」プラットフォームの統計的詳細」（347ページ）を参照してください。

JMP PRO「アソシエーション分析」レポート

デフォルトの「アソシエーション分析」レポートは、以下のセクションで構成されます。

- 「高頻度のアイテム集合」(340ページ)
- 「ルール」(340ページ)

ヒント: レポートの表の項目は、任意の列に従って並べ替えることができます。それには、表内を右クリックし、[**列の値で並べ替え**]を選択します。

JMP PRO 高頻度のアイテム集合

「高頻度のアイテム集合」レポートには、アイテム集合が支持度の降順で表示されます。これらのアイテム集合は、起動ウィンドウで指定した「最小支持度」の条件を満たしています。これらのアイテム集合は、アソシエーションルールの条件と帰結に使われます。この表には、以下の列があります。

アイテム集合 アソシエーションルールの条件と帰結に使われるアイテム集合。

支持度 該当のアイテム集合が登場しているトランザクションの割合。

N個のアイテム 該当のアイテム集合に含まれているアイテムの個数。

JMP PRO ルール

「ルール」レポートには、条件部分で使われているアイテム集合のアイテム数の昇順で、アソシエーションルールが表示されます。ルールはさらに、条件と帰結の和集合に含まれるアイテムに従い、アルファベット順に並べられます。起動ウィンドウで指定した「最小支持度」、「最小信頼度」、「最小リフト値」、「最大先行数」、「最大ルールサイズ」を満たすアソシエーションルールのみが、このレポートに表示されます。

「ルール」レポートの表には以下の列があります。

ルール 条件と帰結で構成されたアソシエーションルール。

条件 トランザクションにおける帰結部分の有無に影響を及ぼすと考えられるアイテム集合。

帰結 条件部分によってその有無が影響を受けると考えられるアイテム集合。

信頼度 条件アイテム集合を含むトランザクションのなかで、帰結アイテム集合を含むものの割合。信頼度は、アソシエーションルールの関係の強さや予測能力の強さを表します。

リフト値 •条件アイテム集合と帰結アイテム集合がそれぞれ独立してトランザクションに現れると仮定した場合の信頼度の期待値に対する、観測された信頼度の比を指します。リフト値は、帰結アイテム集合がどれだけ条件アイテム集合の存在に依存しているかを表します。リフトの最小値は0です。

– リフト値が1未満の場合は、条件アイテム集合と帰結アイテム集合が一緒に生じる頻度が、偶然に期待される頻度よりも低いことを意味します。条件と帰結が反発しあっていると考えられます。

– リフト値が1に近い場合、条件アイテム集合と帰結アイテム集合は、偶然に期待される頻度と同じ頻度で一緒に生じています。
– リフト値が1を超えている場合は、条件アイテム集合と帰結アイテム集合が一緒に生じる頻度が、偶然に期待される頻度よりも高いことを意味します。条件と帰結が関係しあっていると考えられます。

JMP PRO「アソシエーション分析」プラットフォームのオプション

「アソシエーション分析」の赤い三角ボタンのメニューには、以下のオプションがあります。

トランザクションの一覧 トランザクションIDの値とそのトランザクションに含まれるアイテムをリストした表が表示されます。表は、「トランザクションID」列に従って並べられます。

高頻度のアイテム集合 起動ウィンドウで指定した「最小支持度」を上回る支持度を持つアイテム集合のリストが表示されます。詳細については、「高頻度のアイテム集合」(340ページ)を参照してください。

ルール 起動ウィンドウで指定した「最小支持度」、「最小信頼度」、「最小リフト値」、「最大先行数」、「最大ルールサイズ」を満たすアソシエーションルールの表が表示されます。詳細については、「ルール」(340ページ)を参照してください。

特異値分解 トランザクションとアイテムの最初の2つの特異ベクトルをプロットした散布図が表示されます。特異ベクトルは、アイテムの指示行列を特異値分解することにより算出されます。降順に並べられた特異値の表もレポートされます。**「寄与率」**は、該当の次元だけで説明される変動を、**「累積寄与率」**の列はそこまでの次元の累積によって説明される変動を示します。棒グラフは、特異値によって説明される「寄与率」を示します。詳細は、「特異値分解」(342ページ)を参照してください。

回転後の特異値分解 ([特異値分解]が選択されている場合にのみ使用可能。)「トピックのアイテム」と「トピックのスコア」の各レポートが表示されます。このオプションは、トランザクション-アイテムの行列に対してVarimax回転による特異値分解を行い、類似のトランザクションをグループ(トピック)にまとめます。「回転後の特異値分解」(344ページ)を参照してください。

トランザクション特異値分解の保存 トランザクションごとに指定の数の特異ベクトルを含むデータテーブルが作成されます。これらは、トランザクション-アイテムの行列における左特異ベクトルです。「特異値分解」(342ページ)を参照してください。

アイテム特異値分解の保存 アイテムごとに指定の数の特異ベクトルを含むデータテーブルが作成されます。これらは、トランザクション-アイテムの行列における右特異ベクトルです。「特異値分解」(342ページ)を参照してください。

以下のオプションについて詳しくは、『JMPの使用法』の「JMPのレポート」章を参照してください。

ローカルデータフィルタ 現在のレポートにおいてデータをフィルタリングするローカルデータフィルタを表示するか、非表示にします。

やり直し 分析を繰り返すオプションや、やり直すオプションを含みます。また、［自動再計算］オプションを選択すると、このオプションに対応しているプラットフォームにおいて、データテーブルに加えた変更が、該当するレポートに即座に反映されるようになります。

スクリプトの保存 レポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。

By グループのスクリプトを保存 By 変数の全水準に対するレポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。起動ウィンドウで By 変数を指定した場合のみ使用可能です。

JMP PRO 特異値分解

特異値分解（**SVD; Singular Value Decomposition**）は、アソシエーション分析とは異なった枠組みで、互いに関連性のあるアイテムを特定します。この分析では、トランザクション - アイテムの行列が特異値分解され、元の行列が少数の次元に要約されます。それらの結果から、類似のトランザクションや類似のアイテムを見出すことができます。

JMP PRO トランザクション - アイテムの行列

トランザクション - アイテムの行列は、行がトランザクション、列がアイテムに対応する行列です。行列の要素は、0 と 1 です。あるアイテムがあるトランザクションに出現する場合は、それらに対応する行と列の要素が 1 になります。出現しない場合は、その要素は 0 になります。トランザクション - アイテムの行列における要素は、1 よりも 0 になっているものが多いです。そのような行列を、「**疎な行列（sparse matrix）**」と言います。

JMP PRO 特異値分解

特異値分解は、**U**、**S**、**V'** の 3 つの行列を使ってトランザクション - アイテムの行列を近似します。4 つの行列には、次のような関係があります。

トランザクション - アイテムの行列 ≈ **U** * **S** * **V'**

ここで、*nTransactions* を、トランザクション - アイテム行列の行数（トランザクションの個数）とします。また、*nItems* をトランザクション - アイテム行列の列数（アイテムの個数）、*nVec* を指定した特異ベクトルの次元数とします。このとき、*nVec* は、min(*nTransactions, nItems*) 以下でなければなりません。左特異ベクトルの行列 **U** は、*nVec*×*nTransactions* の行列です。**S** は、*nVec* 次の対角行列です。この **S** の対角成分は、特異値になっています。**V'** は、*nTransactions*×*nVec* の行列です。**V'** の行が右特異ベクトルです。

特異ベクトルは、特徴やトピックが似ているアイテム間の関係を反映しています。同じトランザクションに出現することの多いアイテムが 3 つあった場合、特異値分解によって得られる右特異ベクトル **V'** では、これらの 3 つのアイテムに対する値が大きくなります。また、左特異ベクトル **U** は、得られたアイテム空間に射影されたトランザクションを表します。

特異値分解は、間接的な関係も捉えます。同じトランザクションに一緒に出現することのない2つのアイテムがあり、ただし、それら2つは第3のアイテムと一緒には出現しているとします。特異値分解はその間接的な関係を部分的に捉えることができます。2つのトランザクションに直接的には共通のアイテムはなくとも、間接的に関係している共通のアイテムが含まれている場合、特異値分解プロットにおいてそれらのトランザクションは近い位置にプロットされます。

トランザクションデータを特異値分解によって低次元のベクトル空間に変換することで、クラスター分析、分類、回帰分析などが適用できるようになります。［保存］オプションによって特異ベクトルがデータテーブルに出力されるので、それらを他のJMPプラットフォームで分析できます。

JMP PRO 特異値分解レポート

JMP PRO 特異値分解プロット

特異値分解プロットは、トランザクションデータとアイテムデータの最初の2つの特異ベクトルを散布図としてプロットしたものです。

ヒント：ある点がどのトランザクションまたはアイテムを表すのかは、点の上にカーソルを置くことで確認できます。プロットにラベルを表示するには、点を選択し、プロット内を右クリックして、**［行ラベル］**を選択します。

トランザクション特異値分解プロットの点は、各トランザクションに対応しています。これらの点は、左特異ベクトル**U**における最初の2次元です。トランザクション特異値分解プロットで、点の塊が見られる場合、それらのトランザクションは構成が似ていると考えられます。

アイテム特異値分解プロットの点は、各アイテムに対応しています。これらの点は、右特異ベクトル**V**における最初の2次元です。アイテム特異値分解プロットで、点の塊が見られる場合、それらのアイテムは特徴やトピックが似ていると考えられます。

「「アソシエーション分析」の別例：特異値分解」（345ページ）を参照してください。

注意：最初の2つの特異ベクトルは、データの構造を適切に捉えていない可能性があります。「特異値」レポートを見ると、特異ベクトルによってどれだけの変動が説明されているかがわかります。

JMP PRO 特異値

「特異値」表のk番目の行は、k番目の特異ベクトルによって説明される変動、および、それらの累積の割合を示します。

JMP PRO 回転後の特異値分解

(「アソシエーション分析」の赤い三角ボタンのメニューで［特異値分解］が選択されている場合にのみ使用可能。)［回転後の特異値分解］オプションは、トランザクション-アイテムの行列を特異値分解した結果を、Varimax回転します。「トランザクション-アイテムの行列」(342ページ)を参照してください。回転後における特異ベクトルの数を指定する必要があります。指定の数だけの**トピック**が作成されます。

トピックとは、いくつかのアイテムで特徴付けられるトランザクションの集まりを指します。各トピックの各アイテムには、「重み」が計算されます。そして、あるトランザクションに含まれるすべてのアイテムの重みを合計したものを、「**トピックスコア**」といいます。このトピックスコアは、そのトピックにトランザクションが所属する強さを表します。

Varimax回転は、特異ベクトルが少数の座標軸により近づくように、特異ベクトルを回転させます。少数の軸における負荷量を大きくし、他の軸における負荷量を小さくなるように回転が行われるので、解釈がしやすくなります。負荷量は、「回転後のV行列」レポートと「回転後のU行列」レポートに表示されます。

「「アソシエーション分析」の別例: 特異値分解」(345ページ)を参照してください。

JMP PRO トピックのアイテム

(「アソシエーション分析」の赤い三角ボタンのメニューで［特異値分解］が選択されている場合にのみ使用可能。)「トピックのアイテム」レポートには、トランザクションを分類するトピックが表示されます。各トピックにおいて、各アイテムが、その重みの絶対値の降順で並べられます。重みの絶対値が最も大きいアイテムは、そのトピックの主題になっていると解釈できます。また、トピックのスコアは、各トピックに対する各トランザクションの関係の大きさをスコア化したものです。「トピックのスコア」(345ページ)を参照してください。レポートには、Varimax回転に関する以下の情報も表示されます。

変換　Vaimax回転の回転行列。

回転後のV行列　各トピックに対するアイテムの重みを表した行列。列がアイテムに対応しています。回転後のV行列は、特異値分解分析のV行列にVarimax回転を適用したものです。重みが大きいアイテムは、そのアイテムとトピックの間に関連性があることを示します。

回転後のU行列　各トピックにおけるトランザクションスコアを表した行列。列がトランザクションに対応しています。あるトピックでスコアが高いトランザクションは、そのトピックと関連性があると考えられます。大きな値は、トランザクションとトピックの間に関連性があることを示します。

トピック 割合　各トピックの割合の値を示します。

「「アソシエーション分析」の別例: 特異値分解」(345ページ)を参照してください。

JMP PRO トピックのスコア

（「アソシエーション分析」の赤い三角ボタンのメニューで［特異値分解］が選択されている場合にのみ使用可能。）「トピックのスコア」レポートには、すべてのトランザクションにおけるトピックのスコアが、1次元の散布図に描かれます。負のスコアは、そのトランザクションとトピックの間には負の関連性があることを示します。これらのプロットから、各トピックにおけるトランザクションの分布を調べることができます。「「アソシエーション分析」の別例：特異値分解」（345ページ）を参照してください。

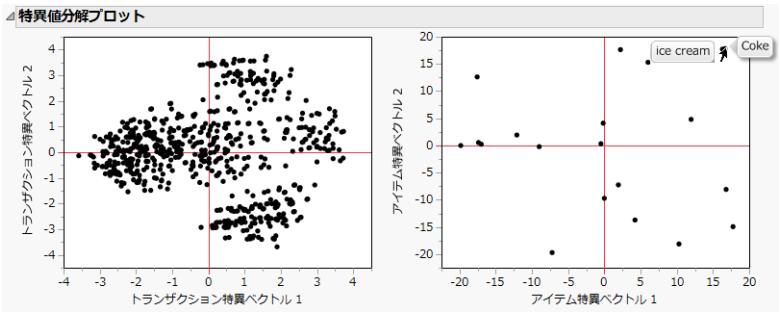
ヒント：「トピックのスコア」のプロットにおいて点を選択すると、データテーブルの対応する行も選択され、また、プロット内の他のトピックでも該当するトランザクションが選択されます。

JMP PRO「アソシエーション分析」の別例：特異値分解

この例では、**「Grocery Purchases.jmp」**サンプルデータからより深い洞察を得るため、トランザクション-アイテムの行列の特異値分解を行います。

1. ［**ヘルプ**］＞［**サンプルデータライブラリ**］を選択し、「Grocery Purchases.jmp」を開きます。
2. ［**分析**］＞［**スクリーニング**］＞［**アソシエーション分析**］を選択します。
3. **「商品」**を選択し、［**アイテム**］をクリックします。
4. **「顧客ID」**を選択し、［**ID**］をクリックします。
5. ［**OK**］をクリックします。
6. 「アソシエーション分析」の赤い三角ボタンをクリックし、［**特異値分析**］を選択します。

図20.5 特異値分解プロット



トランザクション特異値分解プロットを見ると、トランザクションのグループが2つか3つあるようです。アイテム特異値分解プロットの右上の方に、「Coke」と「ice cream」の点が重なり合っています。点の近さから、この2つのアイテムには強い関連性があると考えられます。

7. 「アソシエーション分析」の赤い三角ボタンをクリックし、［**回転後の特異値分析**］を選択します。
8. 「トピック(回転後の特異ベクトル)の個数」として「3」を入力し、［**OK**］をクリックします。

「トピックのアイテム」と「トピックのスコア」の各レポートが表示されます。

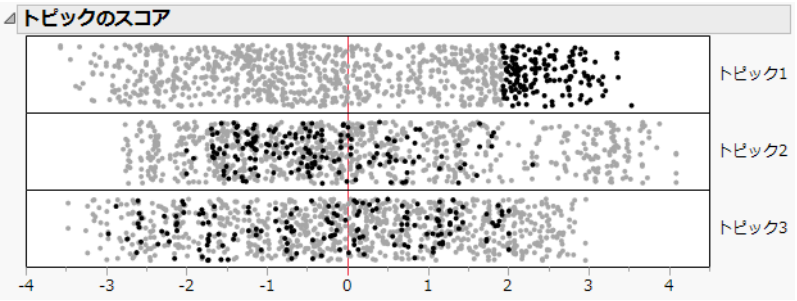
図20.6「トピックのアイテム」レポート

トピックのアイテム

トピック1		トピック2		トピック3	
アイテム	スコア	アイテム	スコア	アイテム	スコア
avocado	0.4272	Coke	0.4234	crackers	0.4545
olives	-0.4257	ice cream	0.4190	apples	-0.4214
baguette	0.3598	herring	-0.3948	soda	0.4156
turkey	-0.3330	sardines	0.3730	Heineken	0.4045
bourbon	-0.2943	chicken	0.3110	sardines	-0.2416
artichoke	0.2801	corned beef	-0.2561	bourbon	0.2297
Heineken	0.2371	steak	-0.2032	steak	-0.2217
corned beef	-0.2235	olives	-0.1853	corned beef	-0.2113
sardines	0.1873	ham	-0.1830		

3つのグループ（トピック）が作成され、「トピックのアイテム」レポートに表示されます。「トピックのアイテム」表に最初にリストされたアイテムは、そのグループの主要なアイテムです。たとえば、トピック1は「avocado」を含むが「olives」を含まないトランザクションのグループと言えます。

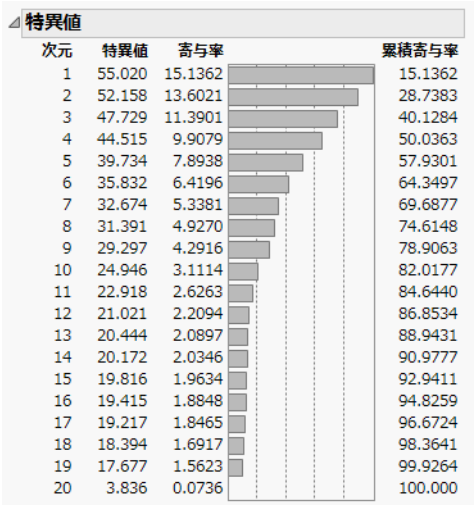
図20.7 トピックのスコア



「トピックのスコア」レポートには、1001のトランザクションそれぞれに割り当てられたトピックのスコアがプロットされます。あるトピックで点のグループを選択すると、これらのトランザクションが他のトピックとどのように関連しているかがわかります。たとえば、トピック1で高い値を持つトランザクションは、トピック2とトピック3では値が低い傾向にあります。

9. 「特異値」レポートを開いてみましょう。

図20.8「特異値」表



特異値

次元	特異値	寄与率	累積寄与率
1	55.020	15.1362	15.1362
2	52.158	13.6021	28.7383
3	47.729	11.3901	40.1284
4	44.515	9.9079	50.0363
5	39.734	7.8938	57.9301
6	35.832	6.4196	64.3497
7	32.674	5.3381	69.6877
8	31.391	4.9270	74.6148
9	29.297	4.2916	78.9063
10	24.946	3.1114	82.0177
11	22.918	2.6263	84.6440
12	21.021	2.2094	86.8534
13	20.444	2.0897	88.9431
14	20.172	2.0346	90.9777
15	19.816	1.9634	92.9411
16	19.415	1.8848	94.8259
17	19.217	1.8465	96.6724
18	18.394	1.6917	98.3641
19	17.677	1.5623	99.9264
20	3.836	0.0736	100.000

図20.8を見ると、最初の2つの特異値は、食料品店データの変動のおよそ30%しか説明していないことがわかります。十分に変動を説明するためには、3次元以上の次元が必要そうです。

JMP PRO「アソシエーション分析」プラットフォームの統計的詳細

この節では、「アソシエーション分析」プラットフォームの統計的詳細を紹介します。

JMP PRO「高頻度のアイテム集合」の生成

「アソシエーション分析」プラットフォームは、高頻度のアイテム集合を抽出するのに、計算時間を短くするため、**Aprioriアルゴリズム**を用いています。Aprioriアルゴリズムは、「あるアイテム集合における支持度は、そのアイテム集合の部分集合における支持度より大きくなることはない」という数学的性質を用いています。Aprioriアルゴリズムは、最小支持度の条件を満たしているアイテム集合だけを組み合わせることにより、より大きなアイテム集合を生成していきます。なお、「アソシエーション分析」プラットフォームでは、「最大先行数」や「最大ルールサイズ」を超えるアイテム集合も計算から落としていきます。これらのオプションは、大規模なデータセットを処理する際に役立ちます。アイテム数が増えるにつれて可能なルールの総数は、指数関数的に増加してしまうため、それを抑えるのにこれらのオプションは役立ちます。Aprioriアルゴリズムの詳細については、Agrawal and Srikant（1994）を参照してください。

JMP PRO アソシエーション分析のパフォーマンス指標

この節では、アソシエーション分析で計算される指標を説明します。条件アイテム集合X、帰結アイテム集合をYとします。条件アイテム集合Xと帰結アイテム集合Yとのアソシエーションルールは、$X \Rightarrow Y$と表されます。

JMP PRO 支持度

支持度は、あるアイテム集合が現れるトランザクションの割合を指します。

$$\text{支持度}(X) = \frac{X\text{を含むトランザクションの数}}{\text{トランザクションの合計数}}$$

JMP PRO 信頼度

信頼度は、条件アイテム集合を含むトランザクションのうち、帰結アイテム集合を含むものの割合を指します。

$$\text{信頼度}(X \Rightarrow Y) = \frac{\text{支持度}(X \cup Y)}{\text{支持度}(X)}$$

信頼度が0%の場合、そのアソシエーションルールの帰結アイテム集合は、条件アイテム集合を含むトランザクションのいずれにも出現しません。信頼度が100%の場合、そのアソシエーションルールの条件アイテム集合を含むトランザクションは、必ず帰結アイテム集合も含みます。

JMP PRO リフト値

リフト値は、XとYの間の依存関係を示す指標です。

$$\text{リフト値}(X \Rightarrow Y) = \frac{\text{支持度}(X \cup Y)}{\text{支持度}(X) \times \text{支持度}(Y)}$$

リフト値の分子は、XとYが一緒に生じるトランザクションの割合です。分母は、XとYが独立して生じると仮定した場合にXとYが一緒に生じる割合の期待値です。

リフト値が1の場合、XとYは、まったくの偶然だけで一緒に生じると期待される頻度と同じ頻度で、一緒に生じています。リフト値が1より大きい場合、Xが生じた場合に、まったくの偶然だけよりも高い頻度でYは生じます。

付録 A

参考文献

Agrawal, R. and Srikant, R. (1994), "Fast Algorithms for Mining Association Rules," *Proceedings of the 20th VLDB Conference*. Santiago, Chile: IBM Almaden Research Center. Retrieved July 5, 2016 from http://rakesh.agrawal-family.com/papers/vldb94apriori.pdf.

Bates, D.M. and Watts, D.G. (1988), *Nonlinear Regression Analysis & its Applications*. New York, John Wiley and Sons.

Benjamini, Yoav and Hochberg, Yosef (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society, Series B*, 57, 289–300.

Box, G., et al. (1994). *Time series analysis: forecasting and control*. New York, John Wiley and Sons.

Dwass, M. (1955), "A Note on Simultaneous Confidence Intervals," *Annals of Mathematical Statistics* 26: 146–147.

Farebrother, R.W. (1981), "Mechanical Representations of the L1 and L2 Estimation Problems," *Statistical Data Analysis*, Second Edition, Amsterdam, North Holland: edited by Y. Dodge.

Fieller, E.C. (1954), "Some Problems in Interval Estimation," *Journal of the Royal Statistical Society, Series B*, 16, 175-185.

Goodnight, J.H. (1978), "Tests of Hypotheses in Fixed Effects Linear Models," *SAS Technical Report R–101*, Cary: SAS Institute Inc, also in Communications in Statistics (1980), A9 167–180.

Goodnight, J.H. and W.R. Harvey (1978), "Least Square Means in the Fixed Effect General Linear Model," *SAS Technical Report R–103*, Cary NC: SAS Institute Inc.

Hand, D, Mannila, H, and Smyth, P. (2001), Principles of Data Mining, MIT Press.

Hastie, T., Tibshirani, R., and Friedman, J.H.(2009), *Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition, New York: Springer Science and Business Media.

Hawkins D.M., Kass G.V. (1982), "Automatic Interaction Detection," in: Hawkins D.M. ed. *Topics in Applied Multivariate Analysis*. Cambridge University Press, Cambridge

Hocking, R.R. (1985), *The Analysis of Linear Models*, Monterey: Brooks–Cole.

Huber, Peter J. and Ronchetti, Elvezio M. (2009), *Robust Statistics*, Second Edition, New York: John Wiley and Sons.

Jolliffe, I.T. (2002), *Principal Component Analysis*, Second Edition, New York, Springer-Verlag New York, Inc.

Kass GV (1980), "An Exploratory Technique for Investigating Large Quantities of Categorical Data," *Applied Statistics*, 29:119–127

McCullagh, P. and Nelder, J.A. (1983), *Generalized Linear Models*, London: Chapman and Hall Ltd.

Nagelkerke, N.J.D. (1991), "A Note on a General Definition of the Coefficient of Determination," *Biometrika*, 78:3 691-692.

Nelder, J.A. and Wedderburn, R.W.M. (1983), "Generalized Linear Models," *Journal of the Royal Statistical Society*, Series A, 135, 370–384.

Parker, R. J. (2015), *Efficient Computational Methods for Large Spatial Data* [PhD Dissertation], North Carolina State University, Raleigh, NC. Retrieved June 30, 2016 from http://repository.lib.ncsu.edu/ir/bitstream/1840.16/10572/1/etd.pdf

Qian, P.Z., Huaiquing, W., and Wu, C.F. (2012). "Gaussian process models for computer experiments with qualitative and quantitative factors." *Technometrics*, 50:3 383-396.

Ratkowsky, D.A. (1990), *Handbook of Nonlinear Regression Models*, New York, Marcel Dekker, Inc.

Sall, J. (2002), "Monte Carlo Calibration of Distributions of Partition Statistics," SAS Institute. Retrieved July 29, 2015 from http://www.jmp.com/content/dam/jmp/documents/en/white-papers/montecarlocal.pdf.

Santer, T., Williams, B., and Notz, W. (2003), *The Design and Analysis of Computer Experiments*, New York, Springer-Verlag New York, Inc.

SAS Institute Inc. (2013), *SAS/ETS User's Guide*, Version 13.1, Cary, NC: SAS Institute Inc.

SAS Institute Inc. (2010), *SAS/ETS User's Guide*, Version 9.22, Cary, NC: SAS Institute Inc.

SAS Institute Inc. (2004), *SAS/STAT User's Guide*, Version 9.1, Cary, NC: SAS Institute Inc.

Schuirmann, D. J. (1987), "A Comparison of the Two One-sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability," *J. Pharmacokin. Biopharm.*, 15, 657–680.

Shiskin, J., Young, A.H., and Musgrave, J.C. (1967), "The X-11 Variant of the Census Method II Seasonal Adjustment Program," *Technical Report 15*, U.S. Department of Commerce, Bureau of the Census.

Shmueli, G., Patel, N.R., Bruce, P.C (2010), *Data Mining For Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner*, Second Edition, Hoboken, New Jersey: John Wiley & Sons, Inc.

Shmueli, G., Bruce, P.C., Stephens M.L., and Patel, N.R., (2017), *Data Mining For Business Intelligence: Concepts, Techniques, and Applications with JMP Pro*, Hoboken, New Jersey: John Wiley & Sons, Inc.

Westfall, P.H., Tobias, R.D., and Wolfinger, R.D. (2011), *Multiple Comparisons and Multiple Tests Using SAS*, Second Edition, SAS Institute.

Wright, S.P. and R.G. O'Brien (1988), "Power Analysis in an Enhanced GLM Procedure: What it Might Look Like," *SUGI 1988, Proceedings of the Thirteenth Annual Conference*, 1097–1102, Cary NC: SAS Institute Inc.

索引
予測モデルおよび発展的なモデル

P

Q

R

S

T

W-Z

ア

イ

エ

オ

カ

キ

ク

ケ

コ

サ

シ

ス

セ

ソ

タ

チ

ツ

テ

ト

ナ

ニ

ハ

ヒ

フ

ヘ

ホ

マ

モ

ヤ

ユ

ヨ

ラ

リ

ル

レ

ロ

ワ