THE POWER TO KNOW.

SAS® Visual Statistics 7.1

ユーザーガイド

SAS® ドキュメント

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2014. *SAS® Visual Statistics 7.1:ユーザーガイド*. Cary, NC: SAS Institute Inc.

SAS® Visual Statistics 7.1:ユーザーガイド

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

October 2014

SAS provides a complete selection of books and electronic products to help customers use SAS® software to its fullest potential. For more information about our offerings, visit **support.sas.com/bookstore** or call 1-800-727-3228.

目次

本書の利用について

利用者

SAS Visual Statistics は、多種多様なデータを大量に分析し、対話的に予測モデルを構築、評価して正確な知見を迅速に得る必要があるデータマイナー、統計学者、データサイエンティスト、データベースマーケティング担当者、ビジネスアナリストによる使用を想定して設計されています。

x

推奨資料

本書に関連する推奨参考資料のリストを次に示します。

- *SAS Visual Analytics:ユーザーガイド*
- *SAS Statistics by Example*
- *Elementary Statistics Using SAS*
- *Data Quality for Analytics Using SAS*
- *Data Preparation for Analytics Using SAS*
- *Logistic Regression Using SAS:Theory and Application*
- *Generalized Linear and Nonlinear Models for Correlated Data:Theory and Applications Using SAS*

SAS 刊行物の総一覧については、support.sas.com/bookstore にてご確認ください。必要な書籍についてのご質問は、下記までお寄せください。

SAS Books
SAS Campus Drive
Cary, NC 27513-2414
電話: 1-800-727-3228
ファクシミリ: 1-919-677-8166
メール: sasbook@sas.com
Web アドレス: support.sas.com/bookstore

1 部

SAS Visual Statistics の概要

2

1

SAS Visual Statistics について

SAS Visual Statistics とは

SAS Visual Statistics は、SAS LASR Analytic Server のインメモリ機能を利用してモデルの開発やテストを可能にする SAS Visual Analytics のアドオンです。SAS Visual Analytics Explorer(以下、エクスプローラ)では、データソースを探索、調査、可視化して、関連性のあるパターンを検出します。SAS Visual Statistics を使用すると、この機能をさらに拡張し、エクスプローラで検出されたパターンに基づいてモデルを作成、テストおよび比較できます。作成したモデルを他の SAS 製品で使用したり、本番環境に移行したりするには、SAS Visual Statistics でモデルの比較を実行する前または後にそのスコアコードをエクスポートします。

SAS Visual Statistics を使用する利点

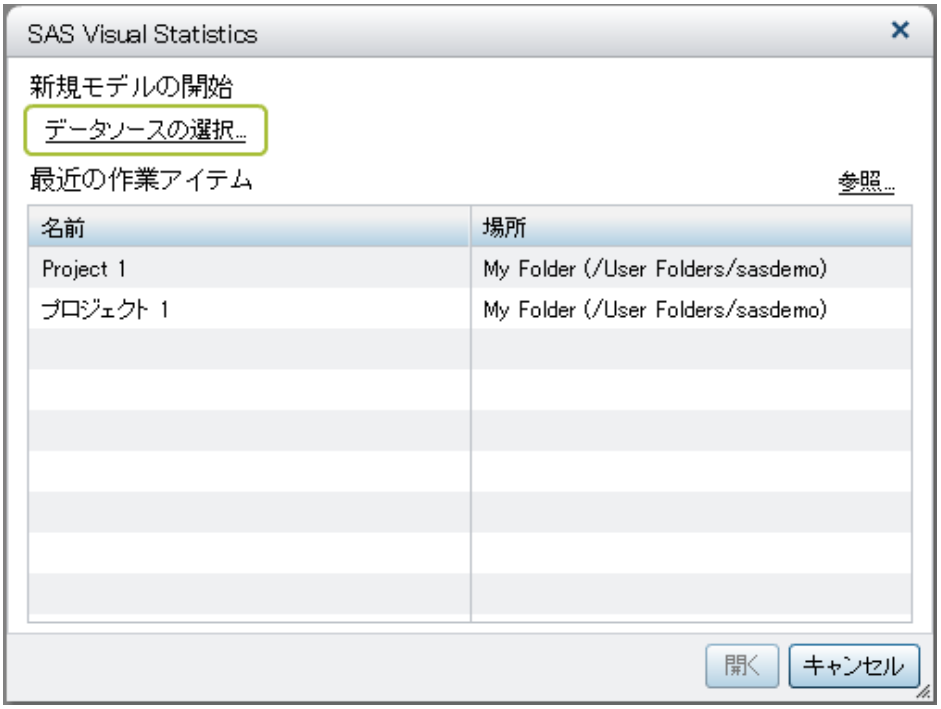
SAS Visual Statistics を使用すると、使いやすい Web ベースのインターフェイスで効果的な統計モデルを迅速に作成できます。データに対して複数の競合モデルを作成したら、SAS Visual Statistics のモデル比較ツールを使用します。モデル比較ツールでは、複数のモデルの相対パフォーマンスを相互に比較して評価し、チャンピオンモデルを選択できます。広範なモデルの選択基準を利用できます。モデルの比較を実行するかどうかにかかわらず、作成し

たモデルのスコアコードをエクスポートできます。エクスポートしたモデルスコアコードを使用すると、作成したモデルを新しいデータに簡単に適用できます。

SAS Visual Statistics へのアクセス

SAS Visual Statistics では、SAS アプリケーションの標準的なサインインウィンドウを使用します。サインインウィンドウを表示するには、システム管理者から指定された URL を使用します。たとえば、次のように入力します。http://host/SASVisualStatistics

適切な URL を入力して SAS Visual Statistics にアクセスしたら、システム管理者から指定されたユーザー ID とパスワードを使用してサインインしてください。SAS Visual Statistics にサインインすると、**Welcome** ウィンドウが表示されます。**Welcome** ウィンドウでは、新規プロジェクトを作成するか、最近使用したプロジェクトを開くかを選択できます。



Welcome ウィンドウでは、次のタスクを実行できます。

- **データソースの選択**をクリックして新規モデルを作成します。**Data sources** ウィンドウが表示されます。

- 既存のモデルを開きます。最近使用したモデルから選択するか、**Browse** をクリックして任意のモデルを選択します。

SAS Visual Statistics を終了するには、SAS Visual Statistics ユーザーインターフェイスの右上の隅にある**サインアウト**リンクをクリックします。

デフォルトでは、何も操作しない状態で一定時間が経過するか、サーバーとの接続が失われると、自動的にサインアウトされます。自動的にサインアウトされると、保存されていないデータは失われます。この場合、最後に保存した状態から、作業し直す必要が生じます。非作業時間、およびセッションのタイムアウト後にアプリケーションに戻るか、それともサインインウィンドウを表示するかは、システム管理者が指定します。

2

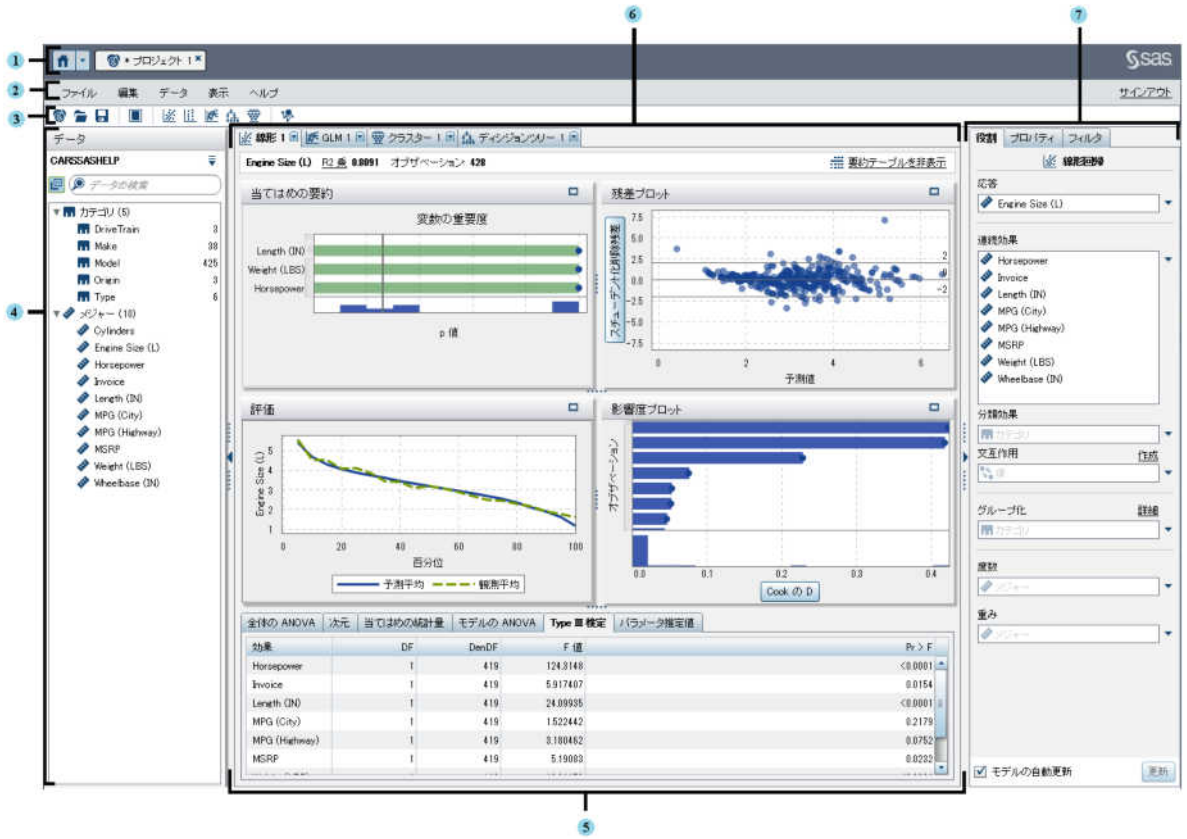
SAS Visual Statistics ユーザーインターフェイス

SAS Visual Statistics ユーザーインターフェイスの外観

概要

このセクションでは、SAS Visual Statistics のユーザーインターフェイスのコンポーネントと、一般的なナビゲーションタスクについて説明します。SAS Visual Statistics ユーザーインターフェイスの主なコンポーネントは次のとおりです。

図 2.1 SAS Visual Statistics ユーザーインターフェイス



1 アプリケーションバーでは、SAS Visual Analytics のホームページや最近使用したプロジェクトにアクセスできます。

2 メニューバーでは、SAS Visual Statistics のすべての機能にアクセスできます。

3 ツールバーでは、SAS Visual Statistics のよく使用する機能にすばやくアクセスできます。

4 **データ**ペインには、分析に利用可能な変数が表示されます。

5 要約テーブルには、現在のモデルの詳細な統計量が表示されます。

6 モデルペインでは、作成したモデルにアクセスしたり、現在のモデルに対する結果のプロットを表示したりすることが可能です。

7 右ペインでは、**役割**、**プロパティ**および**フィルタ**タブにアクセスできます。

メニューとツールバー

SAS Visual Statistics メインメニューから、アプリケーションのすべての機能にアクセスできます。

SAS Visual Statistics ツールバーからは、よく使用するタスクにすばやくアクセスできます。

SAS Visual Statistics ツールバーでは、次のアイコンを使用できます。

アイコン	説明
	プロジェクトを新規作成します。
	保存済みのプロジェクトを開きます。
	現在のプロジェクトを保存します。
	モデリングワークスペースを最大化します。
	モデリングワークスペースをデフォルトの表示に戻します。
	線形回帰分析モデルを作成します。
	ロジスティック回帰分析モデルを作成します。
	一般化線形モデル(GLM)を作成します。
	決定木を作成します。
	クラスタモデルを作成します。
	2 つ以上のモデルを比較します。

データペイン

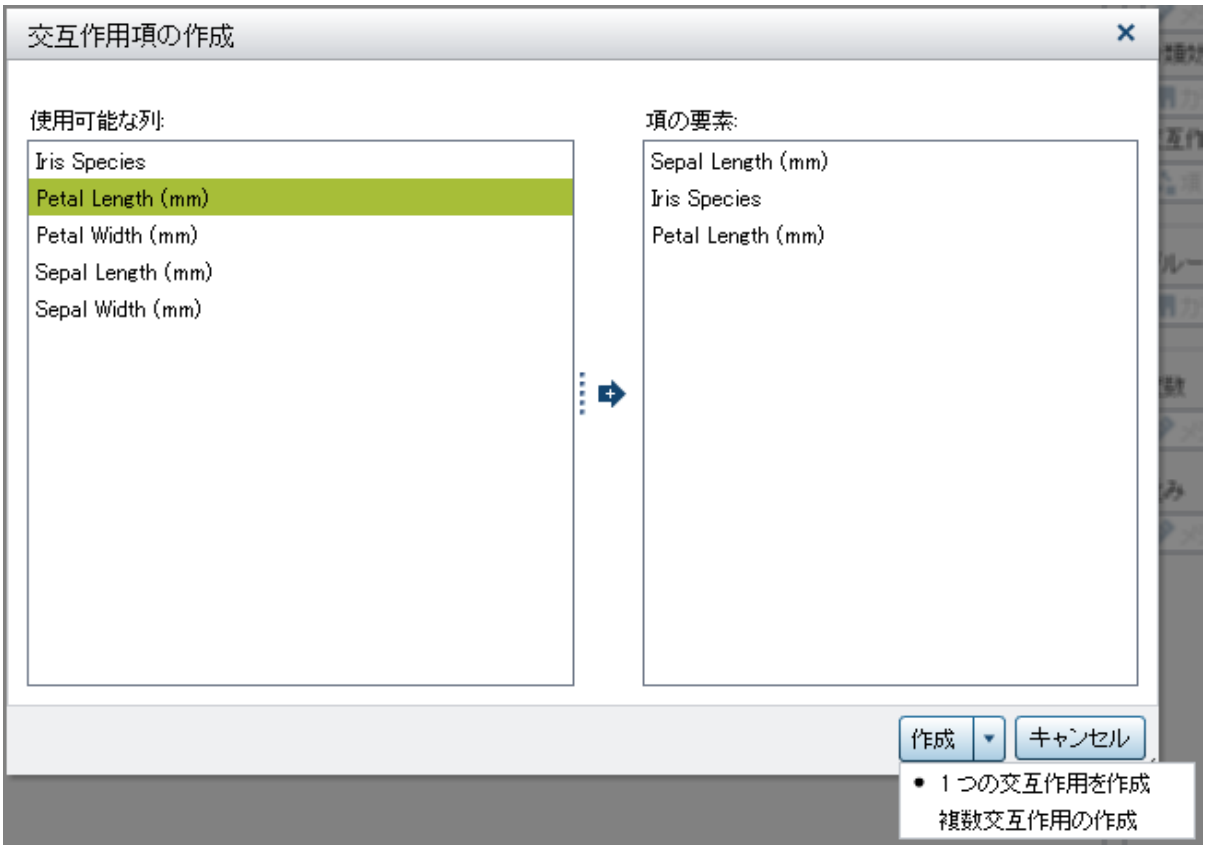
データペインでは、データセットのすべての変数にアクセスできます。変数は、**カテゴリ**変数グループと**メジャー**変数グループに分類されています。カテゴリ変数には、離散レベルがあります。尺度変数は、連続しています。変数の交互作用を作成できます。交互作用項は、**項**グループで利用できます。

データの検索フィールドに検索語を入力すると、その検索語を含む変数のみが表示されます。検索語の大文字と小文字は区別されません。

変数グループを折りたたむには、**データの検索**フィールドの左側にある、をクリックします。変数グループを展開するには、**データの検索**フィールドの左側にあるをクリックします。

データペインのドロップダウンリストアイコン は、**データ**ペインの右上隅にあります。次の項目を利用できます。

- **Create Interaction** を選択すると、**Create Interaction 項**ウィンドウが開きます。



交互作用項の作成ウィンドウでは、利用可能な変数は、**使用可能な列**領域にあります。使用する変数をドラッグアンドドロップするか、ダブルクリックするか、またはウィンドウの中心にある矢印を使用して、**項の要素**領域に移動します。**項の要素**領域に変数を移動した後、単一の交互作用を作成するには、**Create** をクリックします。あるいは、▼をクリックして、**1 つの交互作用を作成**または**複数交互作用の作成**を指定します。

複数交互作用の作成を選択した場合は、選択した変数に対して想定し得るすべての交互作用の組が作成されます。ただし、2 乗項は除外されます。たとえば、前ページの画像では、2 因子の交互作用の組は、Sepal Length (mm)と Iris Species、Sepal Length (mm)と Petal Length (mm)、Iris Species と Petal Length (mm)になります。2 乗項を作成するには、**データ**ペインでその変数を選択し、変数を右クリックして **1 つの交互作用の作成**を選択します。

- **データプロパティ**を選択すると、**データプロパティ**ウィンドウが開きます。**データプロパティ**ウィンドウには、データセットの変数ごとに、名前、分類、データの種類、モデルの種類、形式が表示されます。

- **メジャー詳細**を選択すると、**メジャー詳細**ウィンドウが開きます。**メジャー詳細**ウィンドウには、要約統計量と各尺度変数のヒストグラムが表示されます。
- **アイテムの表示/非表示**を選択すると、**アイテムの表示/非表示**ウィンドウが開きます。**表示アイテム**領域の変数は、**データ**ペインに表示されます。**非表示アイテム**領域の変数は、表示されません。

  変数を一方の領域から他方に移動するには、その変数をドラッグアンドドロップするか、ダブルクリックするか、ウィンドウの中心にある矢印を使用します。複数の変数を移動するには、まずそれらの変数を選択してから、ドラッグアンドドロップするか、矢印を使用します。

  行った変更を保存して**アイテムの表示/非表示**ウィンドウを閉じるには、**OK** をクリックします。
- **Sort Items** を選択すると、変数を昇順で並べ替えるか、降順で並べ替えるかを指定できます。

データペインで、変数または交互作用を右クリックすると、ポップアップメニューが表示されます。このポップアップメニューでは、次の項目を使用できます。カテゴリ変数、尺度変数、項のうちどれを右クリックしたかにより、使用できる項目は一部異なります。

割り当て

次に示す、1 つ以上の役割に変数を割り当てます。

- **応答**は、変数の種類がモデルの応答変数の種類に一致している場合のみ使用できます。
- **連続効果**は、尺度変数のみに使用できます。
- **分類効果**は、カテゴリ変数のみに使用できます。
- **Interactions** は、項のみに使用できます。
- **重み**は、尺度変数のみに使用できます。
- **Frequency** は、尺度変数のみに使用できます。
- **Group By** は、カテゴリ変数のみに使用できます。
- **Filter**

1 つの交互作用の作成
選択した変数組に対して単一の交互作用を作成します。変数を 1 つのみ選択した場合は、2 乗項の交互作用を作成します。

名前の変更
選択した変数の新しい表示名を指定します。

非表示
選択した変数を非表示にします。

削除
選択した項を削除します。これは、SAS Visual Statistics で作成された項目に対してのみ使用できます。

カテゴリ
選択した変数をカテゴリ変数にするかどうかを指定します。

メジャー
選択した変数を尺度変数にするかどうかを指定します。数値変数は、カテゴリ変数、尺度変数のどちらにでも割り当て可能です。

プロパティ
選択した変数または項に関する情報を表示します。

右ペイン

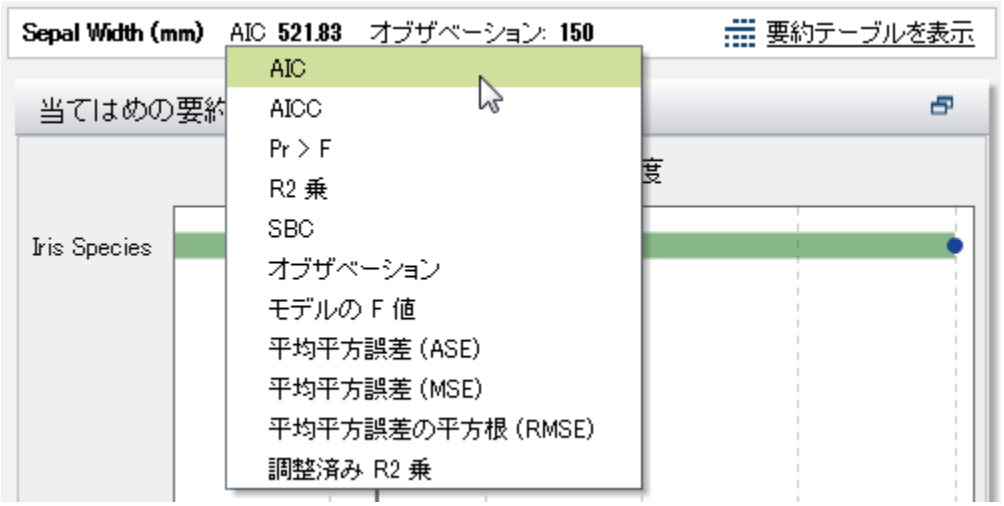
右ペインには、**役割**、**プロパティ**、および**フィルタ**タブが含まれています。これらの 3 つのタブはすべて、モデリングパラメータの定義に使用します。**役割**タブでは、モデルで使用する変数とモデルにおけるその変数の目的を指定します。**プロパティ**タブでは、各モデルに固有の特性を指定します。**フィルタ**タブでは、モデル化対象のデータをサブセット化できます。

タブ	説明
役割	このタブを使用すると、モデルに変数を追加できます。**データ**ペインから、使用する変数を**役割**タブの任意の役割にドラッグアンドドロップします。あるいは、複数の変数を選択して、モデルペインにドラッグアンドドロップすることもできます。この場合、各変数は、最初の有効で利用可能な役割に割り当てられます。応答変数がない場合は、最初の有効な変数が**応答**役割に割り当てられます。この方法では、**Group By**、**Frequency**、**Filter**、**重み**変数は割り当てられません。個々のフィールドから変数を追加または削除するには、▼アイコンを使用します。
プロパティ	モデルの特性を指定します。利用可能なオプションは、選択したモデルによって異なります。
フィルタ	データセットのフィルタリングに使用する変数を指定します。カテゴリ変数、尺度変数またはその両方をフィルタリングできます。フィルタ変数を追加するには、その変数を**データ**ペインから**フィルタ**タブにドラッグアンドドロップするか、▼アイコンを使用します。フィルタ変数を削除するには、その変数名の横にある✕をクリックします。

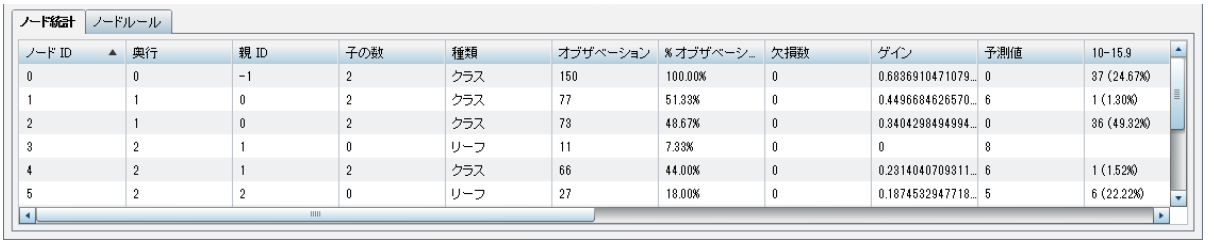
モデルペイン

モデルペインには、モデリングの結果およびプロットが含まれています。利用可能なウィンドウは、選択するモデルによって異なるため、このセクションでは、すべてのモデルに共通の要素に重点を置いて説明します。各モデルの詳細な情報については、そのモデルの章を参照してください。

要約バーには、応答変数、モデルの評価基準(利用可能な場合)、そのモデルで使用されるオブザベーションの数が表示されます。利用可能なすべてのモデル評価基準を表示するには、要約バーにある現在のモデルの評価基準の名前をクリックして、ポップアップメニューを開きます。



要約バーの右側には、**要約テーブルの表示**があります。モデルペインの下部に要約テーブルを開くには、**要約テーブルの表示**をクリックします。決定木モデルの要約テーブルの一例を次に示します。各要約テーブルに表示される詳細情報は、モデルによって異なります。



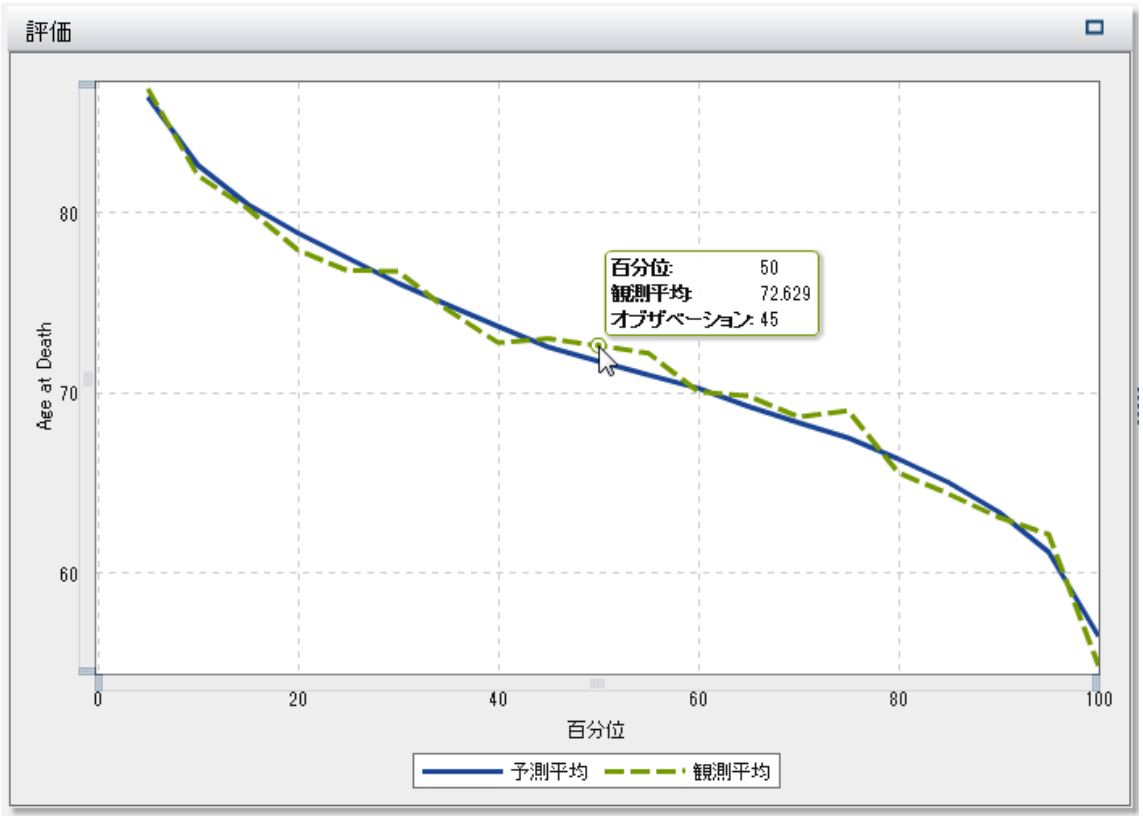
決定木モデル以外のモデルでは、利用可能なすべてのウィンドウは、デフォルトでモデルペインに表示されます。ウィンドウを最大化するには、ウィンドウの右上隅にある□をクリックします。この操作を行うと、モデルペインの他のすべてのウィンドウは、非表示になりますが、要約テーブルは非表示になりません。デフォルトの表示に戻すには、🗗をクリックします。

利用可能なウィンドウを次の表に示します。

ウィンドウ名	使用対象モデル	要約
当てはめの要約	線形回帰分析モデル、ロジスティック回帰分析モデル、一般化線形モデル(GLM)	各モデリング変数の *p* 値を対数目盛りで表示します。アルファ値は、-対数(アルファ)としてプロットされ、クリックやドラッグで調整できる垂直線で示されます。*p* 値のヒストグラムは、ウィンドウの下部に表示されます。 このウィンドウは、Group By 変数が使用される際に分割されます。左側にはグループがリストされ、右側には各グループの *p* 値が単一の線形散布図に凝縮されます。任意のグループの結果のみを表示するように残差プロット、影響プロットおよび評価プロットを変更するには、左側でそのグループをクリックします。
Residual Plot	線形、ロジスティック、GLM	そのモデルのさまざまな残差プロットを表示します。プロットのラベルがボタンの場合には、その軸にプロットする値を選択できます。各モデルには、固有のプロットの組み合わせを利用できます。
評価	線形、ロジスティック、GLM、決定木	尺度目標変数の場合は、ビン化(ビン分割)されたデータセットに対する平均予測値および平均観測値がプロットされます。カテゴリ目標変数の場合は、リフトプロット、ROC プロットおよび誤分類プロットが表示されます。
Influence Plot	線形、ロジスティック	計算された各種統計量に対する各オブザベーションをプロットします。X 軸のラベルは、ボタンでプロットする値を選択できます。各モデルには、固有のプロットの組み合わせを利用できます。

ウィンドウ名	使用対象モデル	要約
Tree	決定木	決定木および決定木マップを表示します。このウィンドウから、決定木の学習を対話的に実行できます。マウスポインタの位置を拡大または縮小するには、お使いのマウスのスクロールホイールを使用します。
リーフの統計量	決定木	決定木にある各リーフノードの応答変数の積み上げヒストグラムを表示します。
クラスターマトリックス	クラスタ	各組のモデリング変数のすべてのクラスタの 2 次元の投影を表示します。個々の投影のさらに大きなプロットを表示するには、そのセルを右クリックして、**開く**を選択します。
Parallel Coordinates	クラスタ	最初に所属クラスタに基づいて分類されたオブザベーションごとに色分けされた線分で表示します。特定のクラスタまたはモデリング変数の範囲に一致するオブザベーションに表示を制限することもできます。

すべてのウィンドウで、任意のオブジェクト上にマウスポインタを置くと、ツールチップによりそのオブジェクトに関する詳細情報が表示されます。情報は、表示されているプロットにより異なります。たとえば、次の画像では、ツールチップにより、(そのビンの)パーセンタイル値、そのビンの平均観測値およびそのビンのオブザベーションの数が表示されています。



値の範囲が表示または選択される場合は必ず、その区間は半開区間です。最小値は、その区間に含まれます。最大値は、その区間に含まれません。これは、ヒートマップ、並列座標プロットおよびその他の表示区間または選択区間に影響します。

プロジェクトとモデルの管理

プロジェクト

SAS Visual Statistics プロジェクトは、1 つ以上のモデルおよび関連するデータにより構成されます。各プロジェクトには、1 つのデータセットのみ含められます。アクティブなデータセットを変更する場合には、プロジェクトを新規作成する必要があります。

プロジェクトを作成するには、次のいずれかの方法を使用します。

- メインメニューで、**ファイル** ▶ **新規** ▶ **プロジェクト**の順に選択します。データソースを選択して、**開く**をクリックします。
- ツールバーで、アイコンをクリックします。

注: プロジェクトがアクティブになっている場合は、現在のプロジェクトを保存するかどうかを確認するウィンドウが表示されます。現在のプロジェクトに加えた変更を保存して、新規プロジェクトを開くには、**保存**をクリックします。現在のプロジェクトに加えた変更を破棄して新規プロジェクトを開くには、**保存しない**をクリックします。現在のプロジェクトに戻るには、**キャンセル**を選択します。

現在のプロジェクトを保存するには、**ファイル** ▶ **名前を付けて保存**の順に選択してから、保存場所と名前を指定します。あるいは、アイコンをクリックしてプロジェクトを保存します。

現在のプロジェクトを閉じるには、メインメニューで、**ファイル** ▶ **閉じる**を選択します。あるいは、アプリケーションバーのプロジェクト名の横にある×アイコンをクリックします。

モデル

モデルを作成するには、メインメニューで、**ファイル** ▶ **新規**の順に選択し、モデルの種類を選択します。ツールバーで、作成するモデルの種類のアイコンをクリックしてモデルを作成することもできます。

モデルの名前を変更するには、メインメニューで**編集** ▶ **名前の変更**を選択します。**名前の変更**ウィンドウが表示されます。**新しい名前**フィールドに新しい名前を入力し、**OK** をクリックします。これは、モデルペインの現在のモデルに影響します。

モデルを複製するには、メインメニューで**編集** ▶ **複製**を選択します。この操作を行うと、コピー元のモデルと同じ設定値を持つモデル名が付けられた*<Model Type>*のコピーが作成されます。モデルの複製は、良いモデルを所有しており、なんらかの強化を行うことでさらにそのモデルを改良できる可能性があるものの、現在のモデルを失うリスクを回避したい場合にお勧めします。この機能を使用することで、元のモデルをそのままの形で保持しながら、複製モデルを調整することができます。これは、モデルペインの現在のモデルに影響します。

モデルを削除するには、メインメニューで、**編集** ▶ **削除**を選択します。これは、モデルペインの現在のモデルに影響します。モデルを削除するたびに、その操作の確認を求めるプロンプトが表示されます。このプロンプトが今後表示されないようにするには、**Don’t show this**

message again を選択します。このプロンプトは、**プリファレンス**ウィンドウで再設定できます。

モデルの名前変更、複製、または削除は、メインメニューを使用するほかに、モデル名の横にある▼を使用しても実行できます。このアイコンからは、**要約テーブルの表示**オプションも使用できます。

環境設定の指定

プリファレンスウィンドウにアクセスするには、メインメニューで、**ファイル** ▶ **プリファレンス**の順に選択します。**プリファレンス**ウィンドウでは、グローバル環境設定とローカル環境設定を指定できます。グローバル環境設定は、SAS Web アプリケーションに一貫して適用されます。この設定には、ユーザーのロケール情報や表示テーマが含まれます。ローカル環境設定は、SAS Visual Statistics のみに適用されます。この設定には、デフォルトのモデルの種類、ステッパの遅延時間および *p* 値の精度が含まれます。

設定した環境設定はすべて、SAS Visual Statistics のセッションに一貫して適用されます。

SAS Visual Analytics Explorer との統合

SAS Visual Analytics Explorer(以下、エクスプローラ)は、SAS Virtual Statistics を迅速に起動するための方法を備えています。データをエクスプローラに読み込んだ後に、**ファイル** ▶ **拡張機能** ▶ **SAS Visual Statistics でデータを表示する**の順に選択します。この操作により、エクスプローラに読み込んだデータを表示した SAS Visual Statistics が起動します。

また、SAS Visual Statistics は、エクスプローラで箱ひげ図、散布図、相関行列からも起動できます。これらの画像のいずれかで、画像内を右クリックして、**拡張機能** ▶ **SAS Visual Statistics で応答をモデル化する**の順に選択します。

散布図の場合は、2 つの尺度を指定する必要があります。Y 軸にプロットされた変数が、SAS Visual Statistics で応答変数として指定されます。利用可能な場合には、指定されたデフォルトのモデルの種類が使用されます。デフォルトのモデルの種類が利用できない場合には、線形回帰分析モデルが使用されます。

相関行列の場合には、単一のセル、行全体または列全体を転送できます。単一のセルを選択した場合には、2 つの変数が転送されます。Y 軸にプロットされた変数が、SAS Visual Statistics で応答変数として指定されます。行全体または列全体を選択した場合には、選択した行または列を定義する変数が応答変数として指定されます。隣接していないセルの選択およびそれらのデータの SAS Virtual Statistics への転送はできません。同様に、単一の行または列に含まれていない隣接しているセルを選択した場合も、そのデータを SAS Visual Statistics mに転送することはできません。利用可能な場合には、指定されたデフォルトのモデルの種類が使用されます。デフォルトのモデルの種類が利用できない場合には、線形回帰分析モデルが使用されます。

箱ひげ図の場合には、少なくとも 1 つのカテゴリ変数と少なくとも 1 つの尺度変数を指定する必要があります。SAS Visual Statistics では、**カテゴリ**フィールドの変数は、常に応答変数として指定されます。追加の格子変数は、Group BY 変数として割り当てられます。利用可能な場合には、指定されたデフォルトのモデルの種類が使用されます。デフォルトのモデルの種類が利用できない場合には、ロジスティック回帰分析モデルが使用されます。

注: エクスプローラから SAS Visual Statistics にデータを転送する場合には、未加工データが転送されます。名前変更した変数、分類の変更、非表示変数、その他の変更は、SAS Visual Statistics でデータが開かれるときには適用されません。

同様に、SAS Visual Statistics から、エクスプローラを起動できます。SAS Visual Statistics でプロジェクトを作成した後に、**ファイル** ▶ **拡張機能** ▶ **SAS Visual Analytics Explorer でデータ表示**の順に選択します。

2 部

モデルの構築

3

モデリング情報

利用可能なモデル

SAS Visual Statistics では、次のモデルを利用できます。

- 線形回帰分析 (33 ページ) では、区間応答の値を 1 つ以上の効果変数の線形関数として予測します。
- ロジスティック回帰分析 (45 ページ) では、2 項応答または順序応答によって、目的のイベントが 1 つ以上の効果の関数として取得される確率を予測します。
- 一般化線形モデル (59 ページ) は、従来の線形モデルを拡張したものです。GLM を使用すると、非線形リンク関数を介して母平均を線形予測子に依存させることができます。

- 決定木 (71 ページ) では、各オブザベーションに適用された一連のルールに基づいて入力データの階層状のセグメントが作成されます。
- クラスタ (85 ページ) では、入力データを同じ特性を共有するグループにセグメント化します。

変数と交互作用項

変数

カテゴリ変数

カテゴリ変数は、離散レベルを持つ数値変数または非数値変数です。カテゴリ変数のレベルは、SAS Visual Statistics では、非順序型とみなされます。カテゴリ変数の例には、ドリンクのサイズ(スモール、ミドル、ラージ)、エンジンの気筒数(2、4、6、8)または顧客による購買(有りまたは無し)などがあります。

カテゴリ変数は、応答変数から作成できます。作成するには、応答変数を右クリックして **Category** を選択します。この場合、尺度変数の各個別値がカテゴリ変数のレベルに変わります。

カテゴリ変数は、分類モデル、分類効果変数、決定木の予測子、フィルタ変数および Group BY 変数の応答変数として使用できます。

注: 最適なパフォーマンスと有効なモデリング結果を確保するには、カテゴリ変数に許容される個別階層の最大数をモデルの種類および変数の役割に基づいて制限します。

尺度変数

尺度変数は、2 つの数値の間に無限の値を想定できる連続数値変数です。カウント変数などのように、一部の数値変数が連続していない場合でも、モデリングにおいては、これらの変数を連続した値として取り扱うことができます。尺度変数の例には、ドリンクの温度、エンジンの排気量、または顧客の購買額の合計などがあります。

尺度変数ごとの要約統計量やヒストグラムは、**データ**ペインで変数を右クリックして、**プロパティ**を選択することにより取得できます。表示する変数を指定するには、**名前**ドロップダウンメニューを使用します。

尺度変数は、連続モデル、連続効果変数、決定木の予測子、オフセット変数、度数変数、重み変数、フィルタ変数の応答変数として使用できます。

交互作用項

2 つの変数、A および B は、モデルの一方の変数の効果が変化すると他方の変数の効果も変化する場合、*交互作用*の関係があります。つまり、モデルにおいて、変数 A と変数 B の効果は相加的ではありません。

SAS Visual Statistics を使用すると、2 つ以上の入力変数間に、2 乗項の交互作用を含む、交互作用を作成できます。2 乗項の交互作用とは、任意の変数とその変数自身との交互作用です。カテゴリ変数に対しては、2 乗項の交互作用は作成できません。

交互作用項が役立つ例として、複数の車の燃費(MPG: 1 ガロンあたりの自動車の走行距離)をモデル化する場合を考えてみましょう。2 つの入力変数は、エンジンの排気量(リットル単位)およびエンジンのサイズ(気筒数)です。いずれかの値が増加すれば、燃費は悪くなると予想されます。ただし、エンジンの排気量による燃費に対する効果が、エンジンサイズ全体で一定ではないと疑われる場合、これらの 2 つのの変数の間に交互作用項を作成することを検討する必要があります。

SAS Visual Statistics では、2 因子のみの交互作用項の作成に限定されません。任意の数(ただし、利用可能な入力変数を超えない数)の変数を含む *n* 因子の交互作用項を作成できます。

交互作用項の個別階層の数は、その項の各変数の階層の数の積になります。尺度変数は、1 階層を含むかのように扱われます。交互作用項の階層の数は、回帰分析モデルで許容される個別階層の最大数に照らしてカウントされます。

変数の選択

変数の選択は、最も有意な変数のみを含むように入力変数の数を減らしていくプロセスです。線形回帰分析モデルとロジスティック回帰分析モデルは、変数の選択を自動的に実行するためのプロパティを備えています。このプロパティを使用すると、SAS Visual Statistics で入力変数に対して変数減少法を実行して、最も有意な変数を判定できます。最も有意な変数を使用したモデリングは、データに過剰適合するモデルの作成を回避するために行います。自動

化された変数選択の実行は実際には、変数の選択を実行しない場合に比べ時間がかかることがあります。

欠損値

デフォルトでは、SAS Visual Statistics は、任意の割り当てられた役割変数で、欠損値を含むすべてのオブザベーションを破棄することにより欠損値を処理します。ただし、線形回帰分析モデル、ロジスティック回帰分析モデル、GLM モデルは、**有用な欠損**プロパティを備えています。場合によっては、オブザベーションに欠損値が含まれているという事実によってモデリングに関連のある情報が提供されることがあります。このプロパティを明示的に選択することで、変数の欠損値を別個の変数としてモデル化できます。尺度変数の場合、欠損値は、平均観測値を使用して推定補完され、欠測を示す指標変数が作成されます。カテゴリ変数では、欠損値は、個別階層とみなされます。

Group BY 変数

Group BY 変数を使用すると、1 つ以上のカテゴリ変数によって定義されているデータセグメントごとにモデルの当てはめを行うことができます。すべての Group BY 変数の階層のそれぞれの固有の組み合わせは、特定のデータセグメントです。たとえば、3 つの階層をもつ Group BY 変数が 1 つある場合は、3 つのデータセグメントがあります。しかし、2 つの Group BY 変数があり、一方の変数に 3 つの階層があり、もう一方の変数に 4 つの階層がある場合は、最大で 12 のデータセグメントがあります。データセグメントは、分類階層の組み合わせにオブザベーションがない場合は作成されません。

SAS Visual Statistics では、**詳細なグループ化**機能を使用する場合を除き、最大数の BY グループを実行します。デフォルトでは、許容される BY グループの最大数は 1024 です。空のデータセグメントは、モデルで許容される BY グループの最大数に照らしてカウントされます。

2 つ以上の Group BY 変数を指定した場合、その結果は変数が **Group By** フィールドに表示されている順番でグループ化されます。

当てはめの要約ウィンドウでは、特定のデータセグメントを選択すると、**Residual Plot** ウィンドウと **Influence Plot** ウィンドウが、指定されたデータセグメントのオブザベーションのみを含むように更新されます。

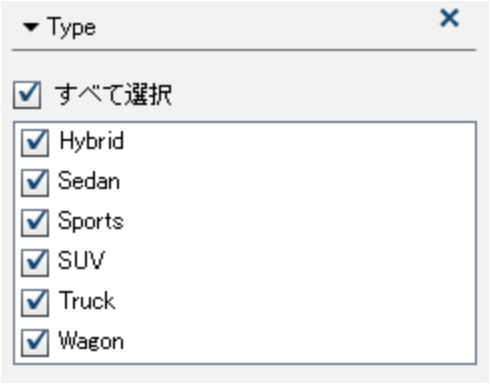
詳細なグループ化ウィンドウは、変数のグループ化に対するさらなる制御機能を備えています。**詳細なグループ化**ウィンドウにアクセスするには、右ペインの **Group By** の横にある**詳細**をクリックします。

Group By フィールドでは、グループ化に使用されている変数を選択できます。指定した尺度変数の集計統計量を表示するには、**Use advanced features** オプションを選択します。**メジャー**フィールドに尺度変数を指定します。**Aggregation** フィールドには、**Average** または **Sum** を計算するかを指定します。**Count** フィールドには、*n* 値の **Top** または **Bottom** が必要であるかどうかを指定します。**Count** の下のフィールドでは、*n* の値を指定できます。

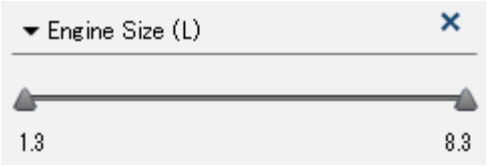
フィルタ変数

モデリングデータをサブセット化するには、フィルタ変数を使用します。モデルで使用されている変数だけでなく、データに含まれている変数もフィルタリングできます。変数のフィルタリングは、現在のモデルのみに適用されます。

カテゴリ変数のフィルタリングを実行する場合は、その変数の階層のリストが表示されます。そのモデルに含める値のみを選択します。次の画像では、すべての階層が利用可能です。



尺度変数のフィルタリングを行う場合、スライダを使用すると、値の範囲を指定できます。三角を使用して、フィルタ変数の下限値と上限値を指定します。



モデルスコアコード

モデルのスコアリングとは、関心の対象となる応答変数を含む可能性のあるデータセットの予測値を生成するプロセスをいいます。スコアコードは、任意の SAS 環境で新しいデータセット

に実行可能な SAS DATA ステップとしてエクスポートされます。いかなる形でもモデルで使用されているすべての変数は、スコアコードに含まれます。これには、交互作用項、Group BY 変数、度数変数、重み変数が含まれます。スコアコードは対話型の決定木には使用できません。

モデルのスコアコードを生成するには、メインメニューで、**ファイル** ▶ **Export** ▶ **モデルのスコアコード**の順に選択します。**モデルスコアコードのエクスポート**ウィンドウで、エクスポートするモデルを選択して、**OK** をクリックします。**名前を付けて保存**ウィンドウで、コードの保存場所に移動し、**保存**をクリックします。

スコアコードは、.sas ファイルとして保存され、任意のワープロプログラムで表示できます。

4

線形回帰分析モデル

線形回帰分析モデルの概要

線形回帰分析は、尺度応答変数の値を 1 つ以上の効果の線形関数として予測を試みます。線形回帰分析モデルでは、最小二乗法を使用してモデルを決定します。最小二乗法では、入力データセットのすべてのオブザベーションの残差平方和を最小化することによって、最良適合線を作成します。残差平方和とは、オブザベーションと最良適合線との間の垂直方向の距離です。最小二乗法には、入力データの分布に関する仮定は必要ありません。

線形回帰分析モデルには、尺度応答変数と少なくとも 1 つの効果変数または交互作用項が必要です。

線形回帰分析モデルのプロパティ

線形回帰分析モデルでは、次のプロパティを使用できます。

名前
このモデルの名前を指定します。

有用な欠損
情報のある欠測アルゴリズムを使用するかどうかを指定します。詳細については、欠損値 (28 ページ)を参照してください。

変数選択の使用
変数の選択を実行するかどうかを指定します。詳細については、変数の選択 (27 ページ)を参照してください。

有意水準
変数をモデルで検討するために必要となる有意水準を指定します。このプロパティは、**変数選択の使用**が選択されている場合のみ使用できます。

評価

- **Use default number of bins** では、デフォルトのビン数を使用するか、独自の値を設定するかを指定します。デフォルトでは、尺度変数は、20 のビンにグループ化されます。
- **Use default number of bins** プロパティが選択されていない場合には、**Number** に、使用するビン数を指定します。5～100 の範囲の整数値を指定する必要があります。
- **Tolerance** には、パーセンタイル値を推定する反復アルゴリズムの収束の決定に使用する許容値を指定します。アルゴリズムの精度を高めるには、より小さな値を指定します。

Show diagnostic plots
Residual Plot ウィンドウ、**評価**ウィンドウおよび **Influence Plot** ウィンドウをモデルペインに表示するかどうかを指定します。

線形回帰分析モデルの結果ウィンドウ

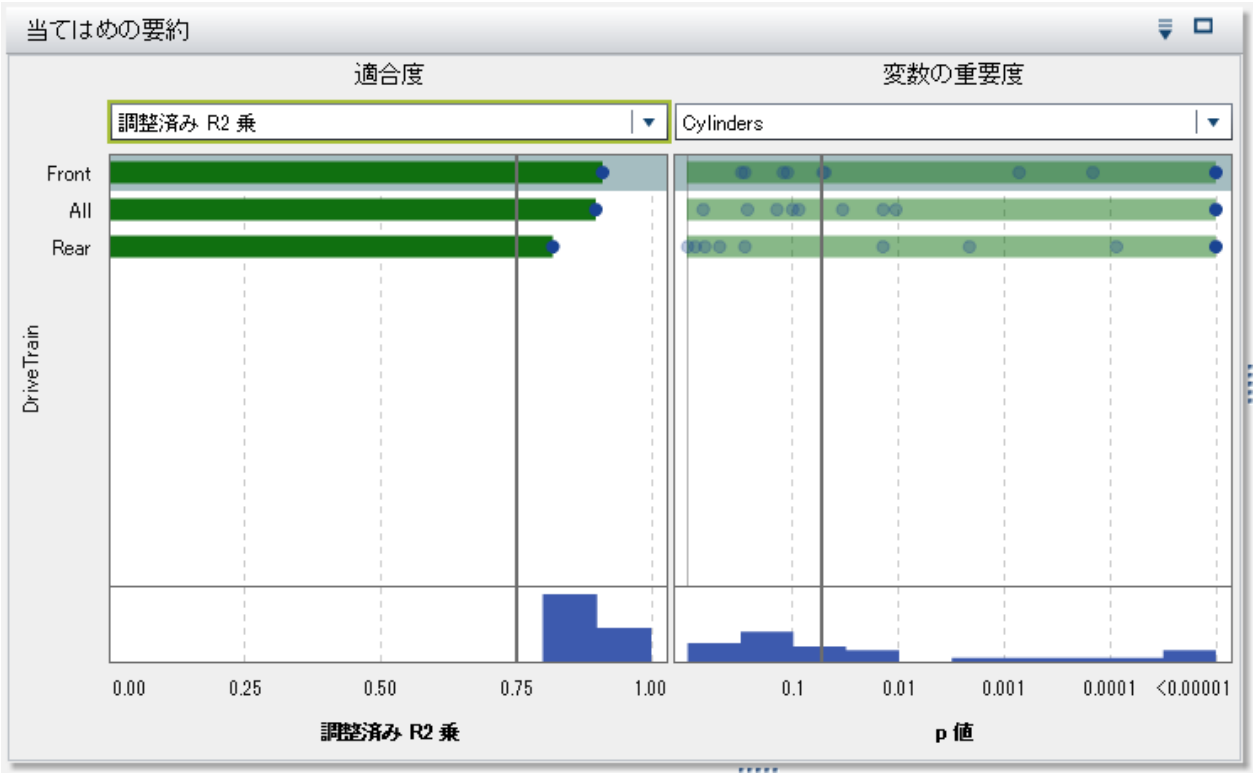
当てはめの要約ウィンドウ

Group By 変数を使用しない場合

当てはめの要約ウィンドウには、*p* 値によって測定された各変数の相対的な重要度がプロットされます。*p* 値は、対数目盛り上にプロットされ、アルファ値(-対数(アルファ)としてプロット)は、垂直線で示されます。アルファ値を調整するには、垂直線をクリックしてドラッグアンドドロップします。*p* 値のヒストグラムは、ウィンドウの下部に表示されます。**当てはめの要約**ウィンドウの例を次に示します。

Group BY 変数を使用する場合

分析に Group BY 変数を含める場合は、**当てはめの要約**ウィンドウには、異なるプロットが表示されます。



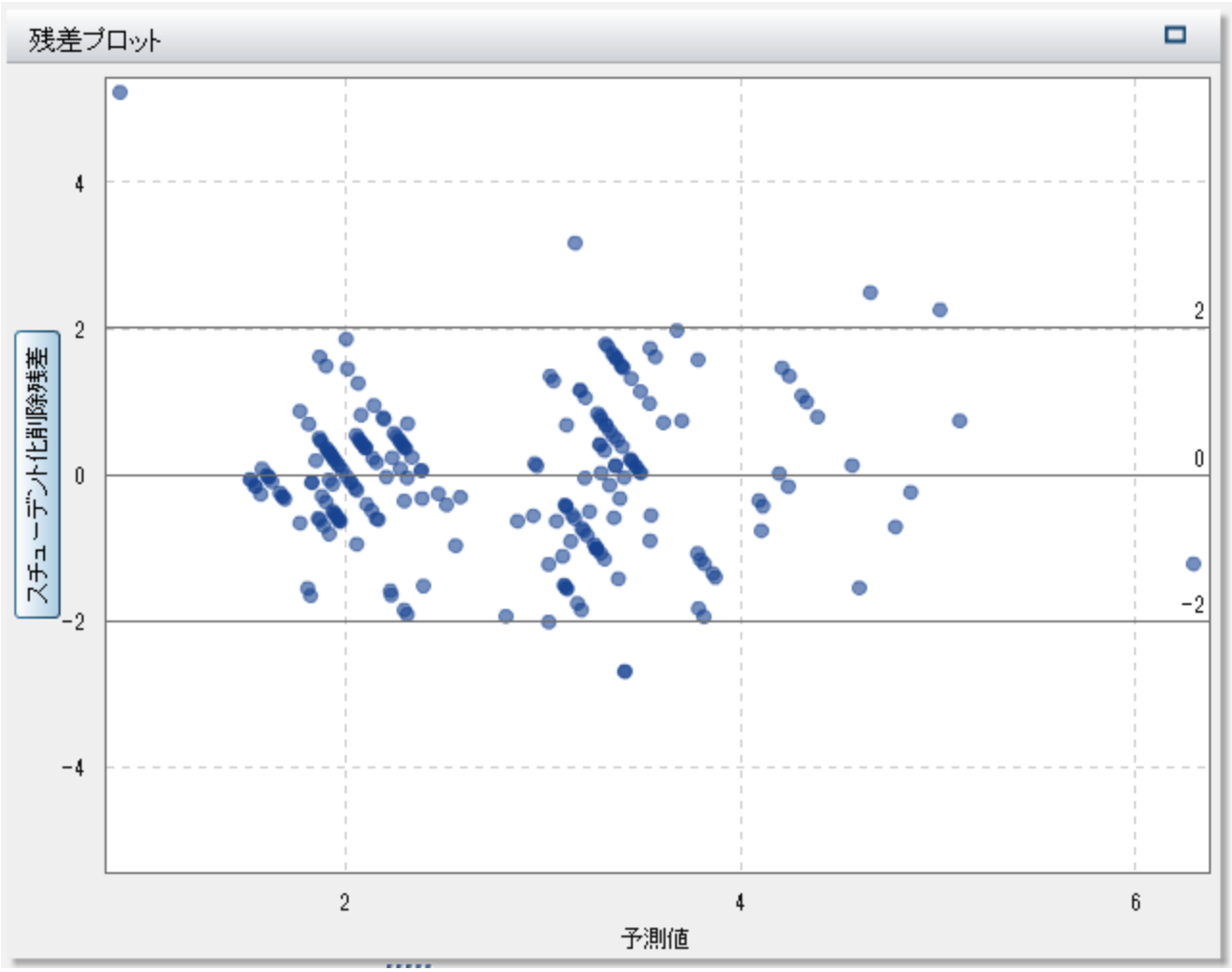
まず、**Variable Importance** プロットに単一の変数のみが表示されていることに注目してください。これは、すべての変数について、変数の重要度が Group BY 変数の各レベルで計算されるからです。異なる効果の変数の重要度を表示するには、ドロップダウンメニューを使用します。次に、Group BY 変数を使用しない場合にはなかった **Goodness of Fit** プロットが表示されていることに注目してください。このプロットは、Grooup BY 変数の各レベルでのモデルによる応答変数の予測の適合度を示すものです。このプロットを使用すると、作成したモデルによる予測の適合度が異なるレベルで大幅に異なるかどうかを判定できます。

プロットの並べ替え方法を指定するには、 アイコンを使用します。

Residual Plot

散布図

オブザベーションの残差とは、応答値の予測値と実測値の差です。次に示すように、デフォルトで **Residual Plot** には、予測値に対する残差の散布図が表示されます。



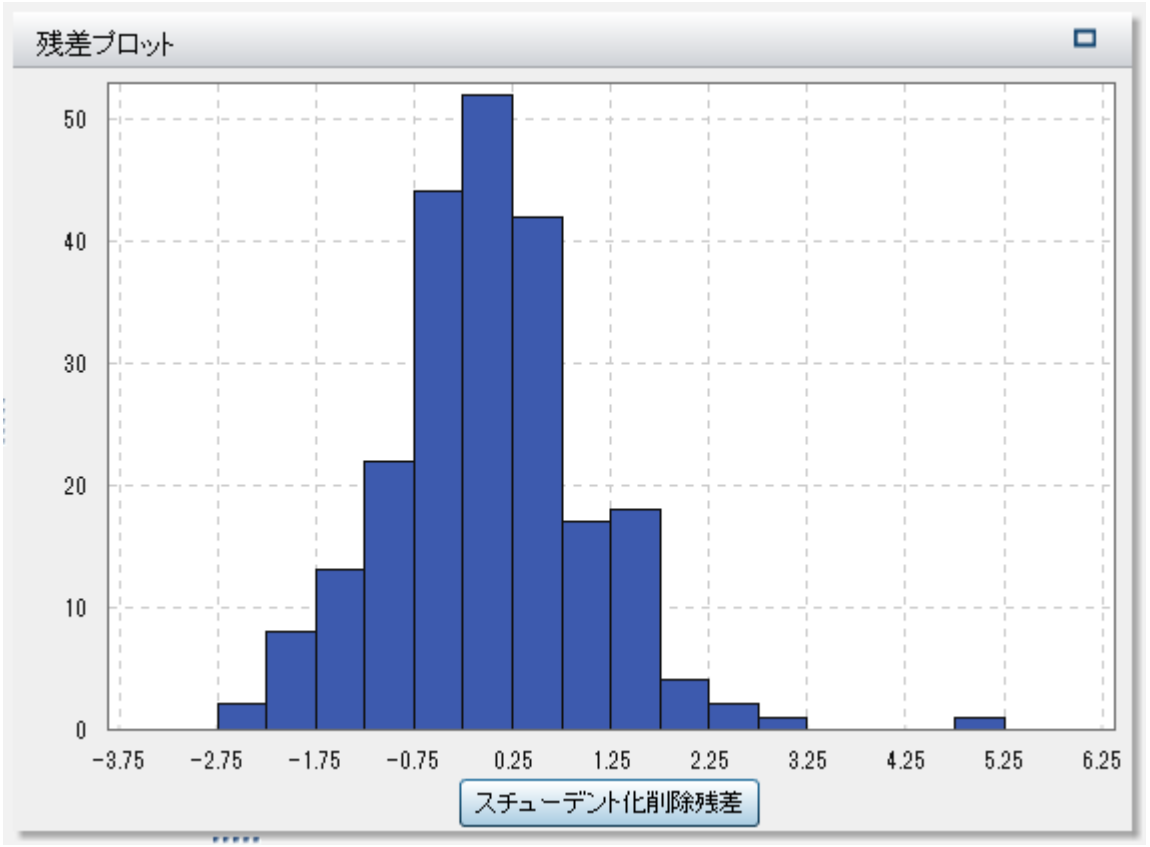
Y 軸上にあるラベルはボタンです。このボタンをクリックすると、Y 軸にプロットされる値を変更できます。Y 軸では、スチューデント削除残差、残差、スチューデント化残差、Press 統計量から選択できます。

残差プロットには、作成したモデルを調べる場合に使えるいくつかの用途があります。第 1 に、残差プロットの明確なパターンは、そのモデルがデータに当てはまっていない可能性があることを示しています。第 2 に、残差プロットでは、予測値に対する残差をプロットする場合に、入力データの非定数分散を検出できます。非定数分散は、予測値が変化するにつれて残差値の相対的な散らばりが変化する場合に明らかです。第 3 に、他の方法と組み合わせると、残差プロットはデータの外れ値を識別する上で役立ちます。

非常に大量のデータセットを使用する場合、残差プロットは、実際のデータのプロットではなく、ヒートマップの形式で表示されます。ヒートマップでは、オブザベーションの実測値がビンに分割され、各ポイントの色は、そのビン内にあるオブザベーションの数を示します。

ヒストグラム

残差プロットのデータをヒストグラムとして表示するには、**Residual Plot** ウィンドウで右クリックして、**Use Histogram** を選択します。残差プロットの Y 軸で利用できた 4 つの値のそれぞれは、ヒストグラムとして利用できます。



プロットされる値を変更するには、X 軸にあるラベルをクリックします。スチューデント削除残差、残差、スチューデント化残差、Press 統計量を選択できます。

残差の分布が正規近似であるか非対称な分布であるかは、ヒストグラムでかなり簡単に確認できます。たとえば、前の画像では、残差は非対称でした。非正規の残差ヒストグラムは、モデルがデータに当てはまっていないことを示している場合があります。

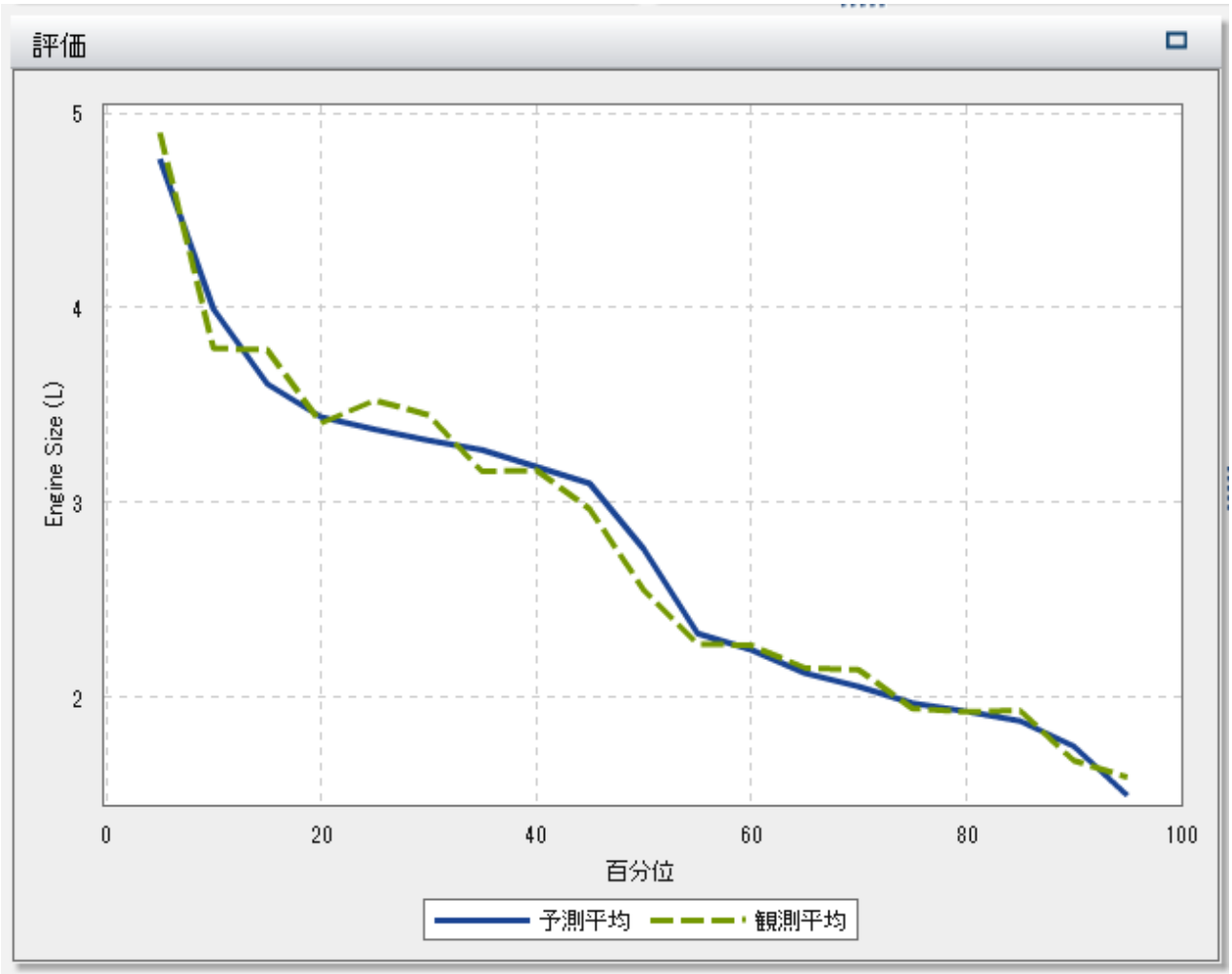
箱ひげ図

現在プロットされている残差の箱ひげ図を表示するには、**Residual Plot** ウィンドウで右クリックして **Plot By** を選択します。次に、箱ひげ図作成時に残差のグループ化に使用するカテゴリ変数を選択します。モデルに含まれているか否かに関係なく、すべてのカテゴリ変数を利用できます。モデルに含まれていない変数の場合は、**Box Plot** ウィンドウを右クリックして、各変数を分類効果または Group BY 変数として割り当てます。**Assign to Classification** メニューのオプションおよび **Assign to Group By** メニューのオプションは、モデルにすでに含まれている変数には利用できません。

外れ値は、デフォルトでは表示されません。外れ値を表示するには、**Box Plot** ウィンドウで右クリックし、**Show Outliers** を選択します。

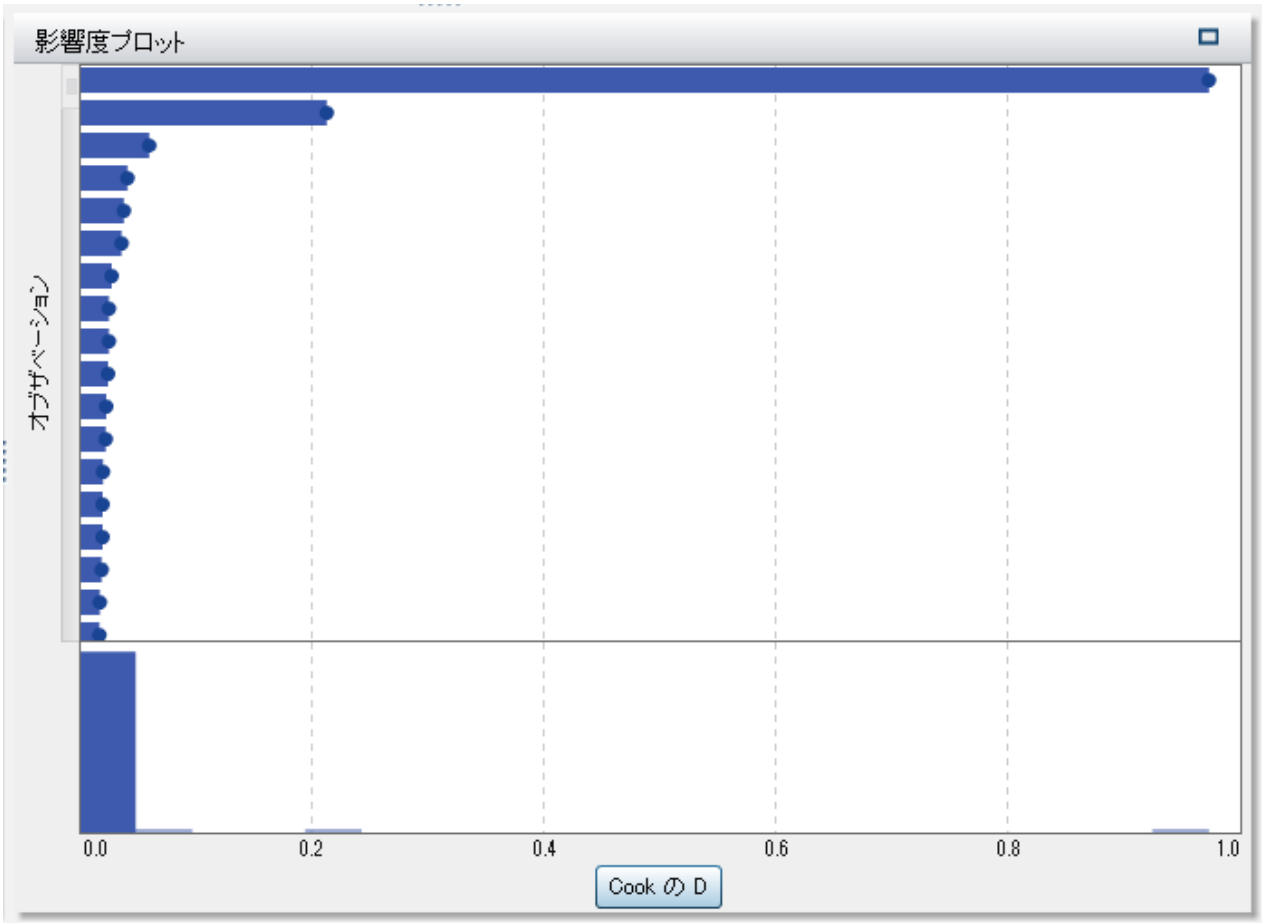
評価

線形回帰分析の場合、**評価**ウィンドウには、ビンに分割されたデータに対する応答の予測値平均と観測値平均がプロットされます。このプロットを使用して、データに対するモデルの適合度を判定します。



Influence Plot

Influence Plot には、オブザベーションごとに計算されるいくつかの測定値が表示されます。次のヒストグラムでは、測定値は表示されているオブザベーションのみに基づいています。入力データに大量のオブザベーションが含まれる場合、オブザベーションはビンに分割されます。デフォルトでは、X 軸には **Cook's D** 値がプロットされます。



その他の利用可能な値は、共分散比、DFFITS、てこ比、尤度変位です。これらの測定値のヒストグラムを表示するには、**Influence Plot** ウィンドウで右クリックして、**Use Histogram** を選択します。このヒストグラムでは、全データセットとそれらのデータに適用されているフィルタが使用されます。

これらの値を使用すると、回帰分析の予測モデルに影響する外れ値や他のデータ点の識別に役立ちます。

当てはめ統計量

線形回帰分析モデルでは、データに対するモデルの適合度の評価に役立ついくつかの評価尺度が計算されます。これらの評価尺度は、モデルペインの上部にあります。利用可能なすべての評価尺度を表示するには、現在表示されている評価尺度をクリックします。

調整済み R2 乗値

調整済み R2 乗値は、効果変数の追加によって生じる結果を説明するために使用します。値は、0~1 の範囲をとります。1 に近いほど望ましい値です。

AIC

赤池情報量規準。小さな値ほど良いモデルであることを示します。AIC 値は負の値になることがあります。AIC は、応答変数の真の分布とモデルで指定された分布との間の差異の Kullback-Leibler 情報量尺度に基づいて導出されます。

AICC

補正赤池情報量規準。AIC のこのバージョンでは、サンプルサイズを説明するために値を調整します。その結果、追加の効果により、AICC には AIC より大きなペナルティ(罰則)が課されます。サンプルサイズが大きくなるに従い、AICC と AIC が収束します。

Average Squared Error

平均平方誤差(ASE)は、平方誤差(SSE)の合計をオブザベーションの数で除算したものです。小さな値ほど、望ましい値です。

F Value for Model

分散を自由度により正規化した後の一元配置分散分析の F 検定の値です。大きな値ほど望ましい値ですが、過剰適合を示すことがあります。

Mean Square Error

平均二乗誤差(MSE)は、SSE を誤差の自由度で除算したものです。誤差の自由度は、ケースの数からモデルの重みの数を減算したものです。このプロセスによって、通常の仮定のもとでノイズの母分散の不偏推定量を算出できます。小さな値ほど、望ましい値です。

オブザベーション

モデルで使用されているオブザベーションの数。

Pr > F

F 統計量に対応する *p* 値です。小さな値ほど、望ましい値です。

R2 乗値

R2 乗値は、モデルがデータにどの程度当てはまっているかを示す指標です。R2 乗値は、0~1 の範囲をとります。1 に近いほど望ましい値です。

Root MSE

MSE の平方根です。

SBC
シュワルツのベイズ規準(SBC)は、ベイズ情報量規準(BIC)とも呼ばれ、モデルの残差平方和と効果の数の増加関数です。応答変数の説明されないばらつきと効果の数によってSBC の値は増加します。このため、SBC が低いほど、説明変数が少ないか、適合度が高い、あるいはその両方を示しています。SBC では、自由度のパラメータに対して AIC より大きなペナルティが課されます。

要約テーブル

モデルペインの上部にある**要約テーブルの表示**をクリックすると、モデルペインの下部に要約パネルが表示されます。要約テーブルには、次の情報が含まれています。

Overall ANOVA
モデル、誤差、修正済みのモデル全体の分散分析結果です。

次元
モデルで使用される効果変数の概要です。このタブでは、そのモデルに選択された尺度および分類効果の数、交差積行列のランク、読み込まれているオブザベーションの数、モデルで使用されているオブザベーションの数などを確認できます。

当てはめ統計量
前のセクションに記載されているすべての当てはめ統計量がリストされています。

Model ANOVA
モデルの分散分析結果です。

Type III 検定
Type III 検定の詳細が表示されています。Type III 検定では、それぞれの部分的な効果の有意性をモデルの他のすべての効果とともに調べます。詳細については、*SAS/STAT User's Guide* の"The Four Types of Estimable Functions"を参照してください。

パラメータ推定値
モデルの各パラメータの推定値を示します。

5

ロジスティック回帰分析モデル

ロジスティック回帰分析モデルの概要

ロジスティック回帰分析モデルは、バイナリ(2 項)分布の応答変数の値を予測するために使用します。ロジスティック回帰分析は、オッズ比の自然対数を説明変数の線形組み合わせとしてモデル化します。この手法により、個々のオブザベーションが目的の階層に属する確率にロジスティック回帰分析モデルを近似させることができます。

ロジスティック回帰分析モデルには、カテゴリ応答変数と少なくとも 1 つの効果変数または交互作用項が必要です。カテゴリ応答変数に、2 つ以上の目的の階層が含まれる場合には、SAS Visual Statistics に、目的の階層の選択を促すプロンプトが表示されます。つまり、SAS Visual Statistics では、その目的の階層にあるすべてのオブザベーションをイベントとして処理し、その他のオブザベーションを非イベントとして処理します。

ロジスティック回帰分析モデルのプロパティ

ロジスティック回帰分析モデルでは、次のプロパティを使用できます。

名前
このモデルの名前を指定します。

有用な欠損
情報のある欠測アルゴリズムを使用するかどうかを指定します。詳細については、欠損値 (28 ページ)を参照してください。

変数選択の使用
変数の選択を実行するかどうかを指定します。詳細については、変数の選択 (27 ページ)を参照してください。

有意水準
変数をモデルで検討するために必要となる有意水準を指定します。このプロパティは、**変数選択の使用**が選択されている場合のみ使用できます。

収束

- **関数収束のオーバーライド**を使用すると、関数の収束値を手動で指定できます。
- **関数収束のオーバーライド**が選択されている場合は、**Value** を使用して、関数の収束値を指定します。より大きな値を指定すると、モデルはより早く収束します。これにより、モデルの学習に費やされる時間を削減できますが、準最適な(最適な水準に達しない)モデルが作成されることがあります。
- **Override gradient convergence** を使用すると、勾配の収束値を手動で指定できます。
- **Override gradient convergence** が選択されている場合は、**Value** を使用して、勾配の収束値を指定します。より大きな値を指定すると、モデルはより早く収束します。これにより、モデルの学習に費やされる時間を削減できますが、準最適な(最適な水準に達しない)モデルが作成されることがあります。

- **最大反復回数**では、モデルの学習中に実行される最大反復回数を指定します。比較的小さな値を指定すると、モデルの学習に費やされる時間を削減できますが、準最適な(最適な水準に達しない)モデルが作成されることがあります。

評価

- **Use default number of bins** では、デフォルトのビン数を使用するか、独自の値を設定するかを指定します。デフォルトでは、尺度変数は、20 のビンにグループ化されます。
- **Use default number of bins** プロパティが選択されていない場合には、**Number** に、使用するビン数を指定します。5～100 の範囲の整数値を指定する必要があります。
- **Prediction cutoff** では、計算される確率がイベントとみなされる値を指定します。
- **Tolerance** には、パーセンタイル値を推定する反復アルゴリズムの収束の決定に使用する許容値を指定します。アルゴリズムの精度を高めるには、より小さな値を指定します。

Show diagnostic plots

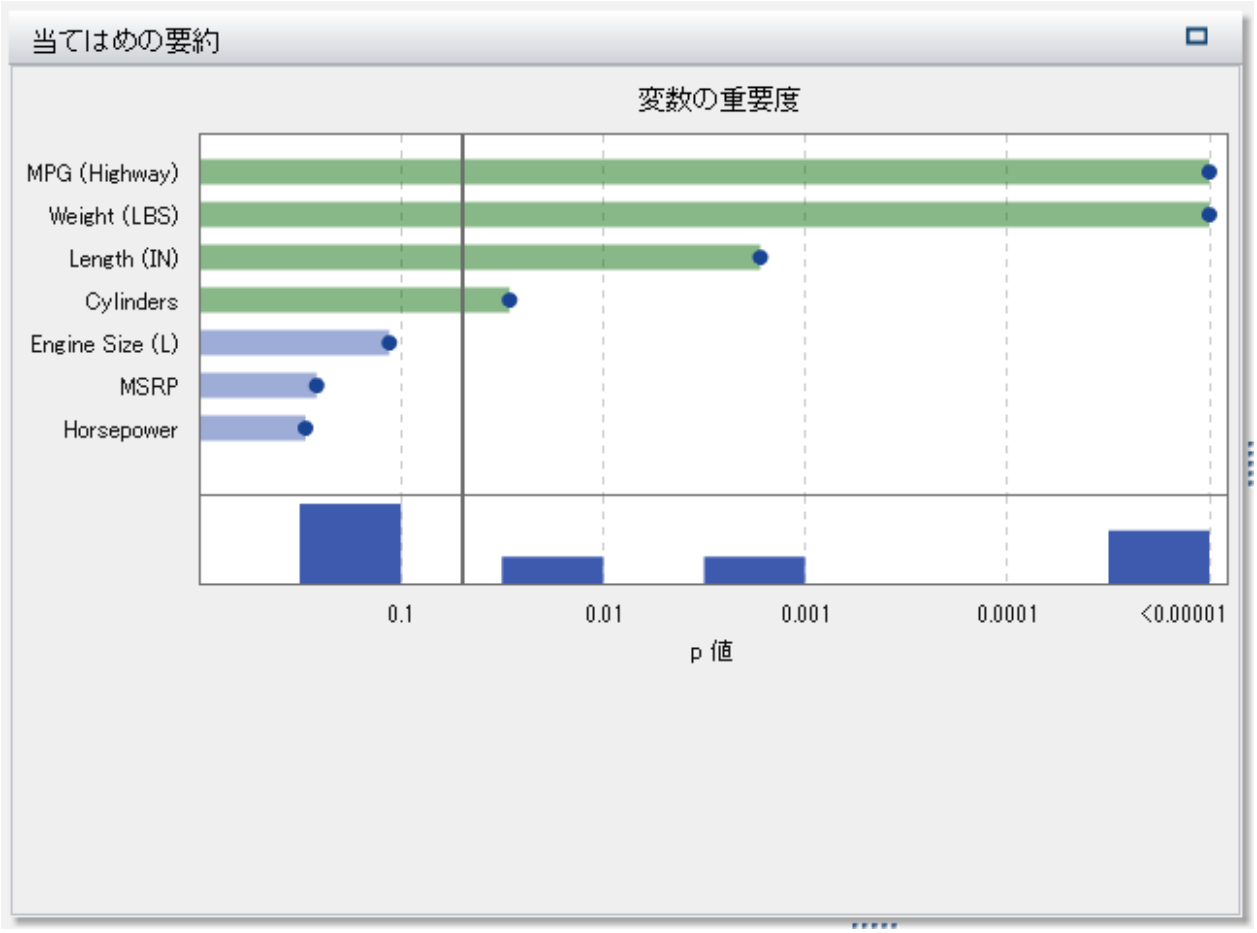
Residual Plot ウィンドウ、**評価**ウィンドウおよび **Influence Plot** ウィンドウをモデルペインに表示するかどうかを指定します。

ロジスティック回帰分析モデルの結果ウィンドウ

当てはめの要約ウィンドウ

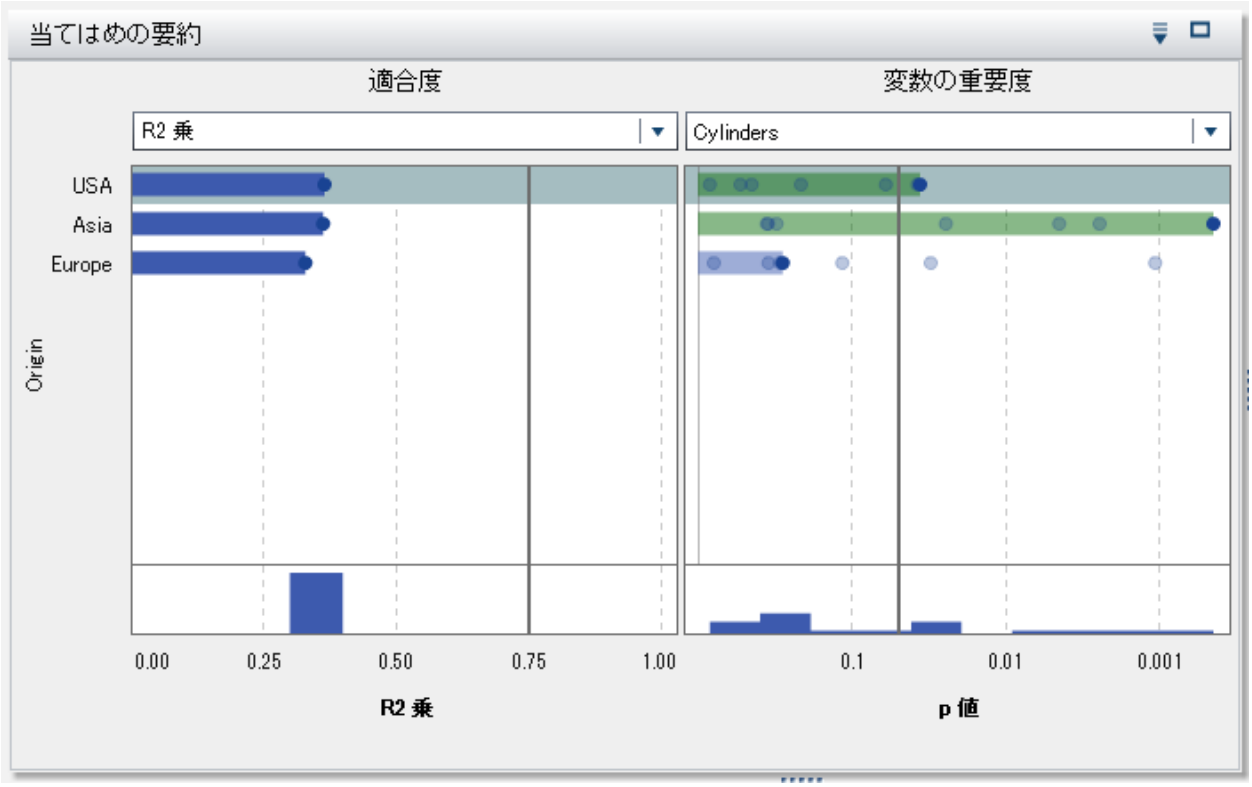
Group By 変数を使用しない場合

当てはめの要約ウィンドウには、*p* 値によって測定された各変数の相対的な重要度がプロットされます。*p* 値は、対数目盛り上にプロットされ、アルファ値(-対数(アルファ)としてプロット)は、垂直線で示されます。アルファ値を調整するには、垂直線をクリックしてドラッグアンドドロップします。*p* 値のヒストグラムは、ウィンドウの下部に表示されます。**当てはめの要約**ウィンドウの例を次に示します。



Group BY 変数を使用する場合

分析に Group BY 変数を含める場合は、**当てはめの要約**ウィンドウには、異なるプロットが表示されます。



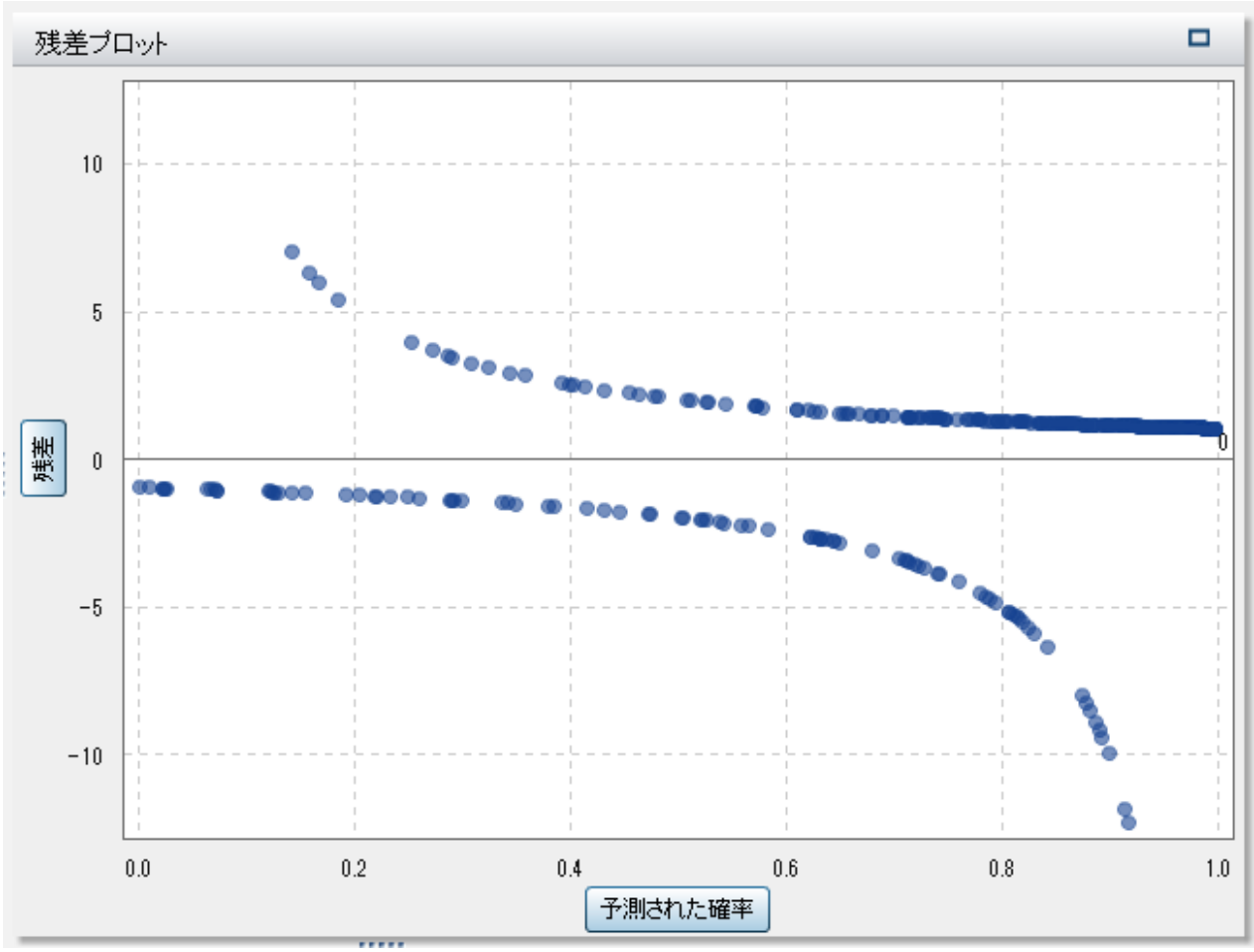
まず、**Variable Importance** プロットに単一の変数のみが表示されていることに注目してください。これは、すべての変数について、変数の重要度が Group BY 変数の各レベルで計算されるからです。異なる効果の変数の重要度を表示するには、ドロップダウンメニューを使用します。次に、Group BY 変数を使用しない場合にはなかった **Goodness of Fit** プロットが表示されていることに注目してください。このプロットは、Grooup BY 変数の各レベルでのモデルによる応答変数の予測の適合度を示すものです。このプロットを使用すると、作成したモデルによる予測の適合度が異なるレベルで大幅に異なるかどうかを判定できます。

プロットの並べ替え方法を指定するには、 アイコンを使用します。

Residual Plot

散布図

オブザベーションの残差とは、応答値の予測値と実測値の差です。次に示すように、デフォルトで **Residual Plot** には、予測確率に対する残差の散布図が表示されます。



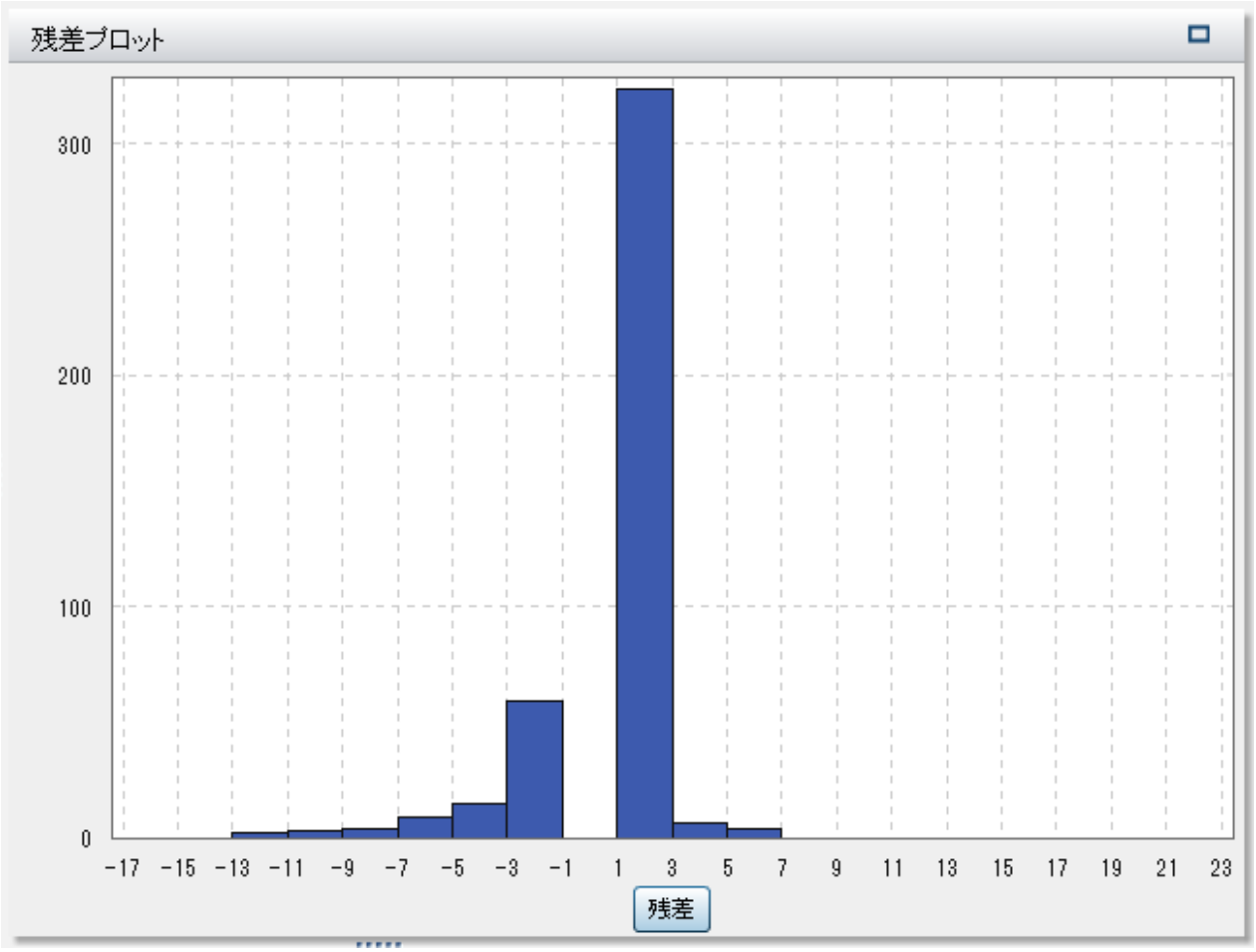
Y 軸と X 軸上にあるラベルはボタンです。これらのボタンのいずれかをクリックすると、その軸にプロットされている値を変更できます。Y 軸の場合、残差または Pearson 残差、逸脱残差および標準化 Pearson 残差をプロットできます。X 軸には、予測確率および線形予測子をプロットできます。標準化 Pearson 残差は、Pearson のカイ 2 乗検定に対する個々の寄与度です。

残差プロットには、作成したモデルを調べる場合に使えるいくつかの用途があります。第 1 に、残差プロットの明確なパターンは、そのモデルがデータに当てはまっていない可能性があることを示しています。第 2 に、残差プロットでは、予測確率に対する残差をプロットする場合に、入力データの非定数分散を検出できます。非定数分散は、予測確率が変化するにつれて残差値の相対的な散らばりが変化する場合に明らかです。第 3 に、他の方法と組み合わせると、残差プロットはデータの外れ値を識別する上で役立ちます。

非常に大量のデータセットを使用する場合、残差プロットは、実際のデータのプロットではなく、ヒートマップの形式で表示されます。ヒートマップでは、オブザベーションの実測値がビンに分割され、各ポイントの色は、そのビン内にあるオブザベーションの数を示します。

ヒストグラム

残差プロットのデータをヒストグラムとして表示するには、**Residual Plot** ウィンドウで右クリックして、**Use Histogram** を選択します。残差プロットの Y 軸で利用できた 4 つの値のそれぞれは、ヒストグラムとして利用できます。



プロットされる値を変更するには、X 軸にあるラベルをクリックします。残差、Pearson 残差、逸脱残差、または標準化 Pearson 残差を選択できます。

残差の分布が正規近似であるか非対称な分布であるかは、ヒストグラムでかなり簡単に確認できます。たとえば、前の画像では、残差は非対称でした。非正規の残差ヒストグラムは、モデルがデータに当てはまっていないことを示している場合があります。

箱ひげ図

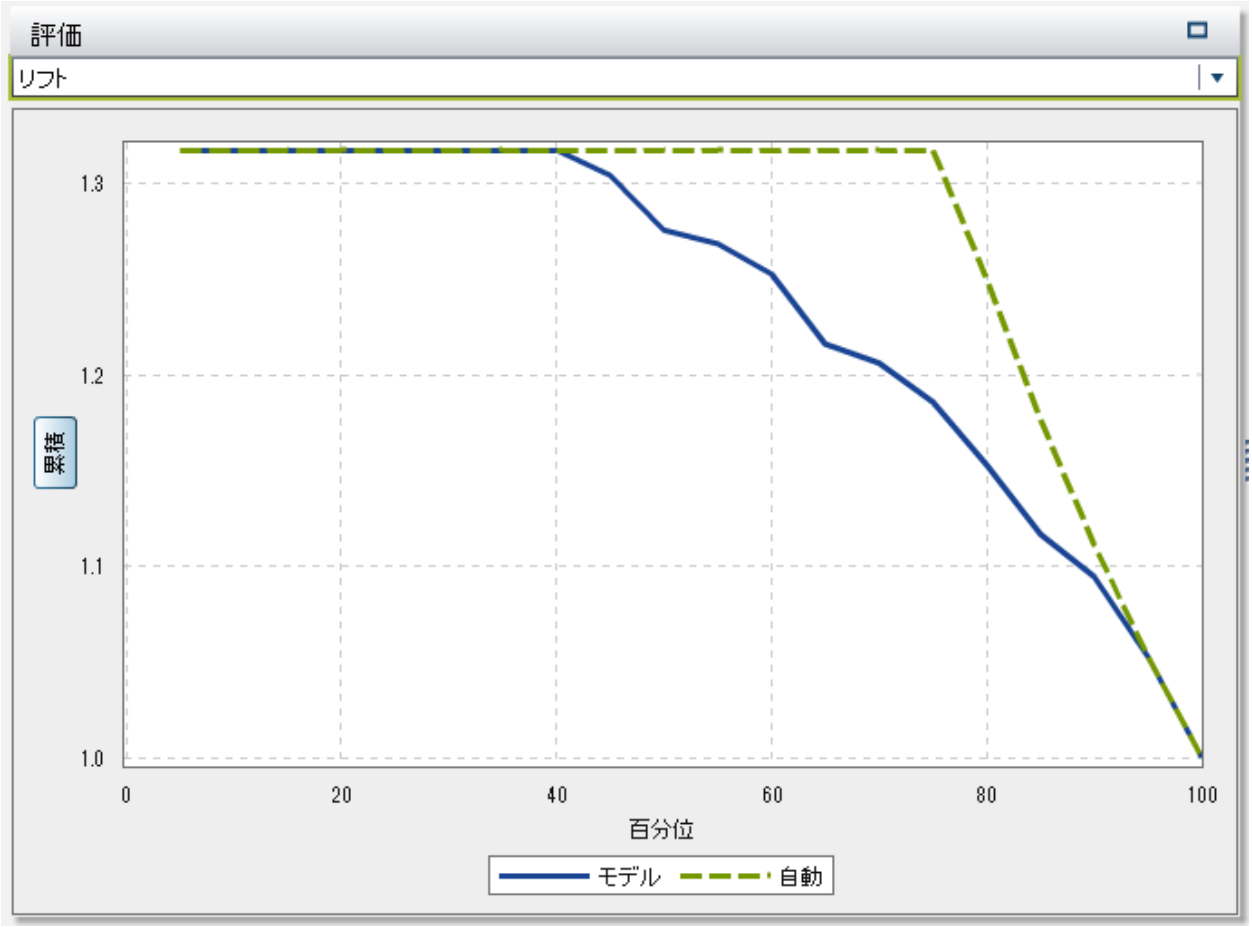
現在プロットされている残差の箱ひげ図を表示するには、**Residual Plot** ウィンドウで右クリックして **Plot By** を選択します。次に、箱ひげ図作成時に残差のグループ化に使用するカテゴリ変数を選択します。モデルに含まれているか否かに関係なく、すべてのカテゴリ変数を利用できます。モデルに含まれていない変数の場合は、**Box Plot** ウィンドウを右クリックして、各変数を分類効果または Group BY 変数として割り当てます。**Assign to Classification** メニューのオプションおよび **Assign to Group By** メニューのオプションは、モデルにすでに含まれている変数には利用できません。

外れ値は、デフォルトでは表示されません。外れ値を表示するには、**Box Plot** ウィンドウで右クリックし、**Show Outliers** を選択します。

評価

リフト

*リフト*とは、モデルの応答の平均割合に対し、パーセンタイルビン内に捕捉された応答のパーセントの比率です。同様に、*累積リフト*とは、現在のパーセンタイルビンのすべてのデータを使用して計算されたものです。デフォルトのリフトチャートには、モデルの累積リフトが表示されます。累積リフトと累積以外のリフトを切り換えるには、リフトチャートで右クリックして、**Switch Lift Type** を選択します。



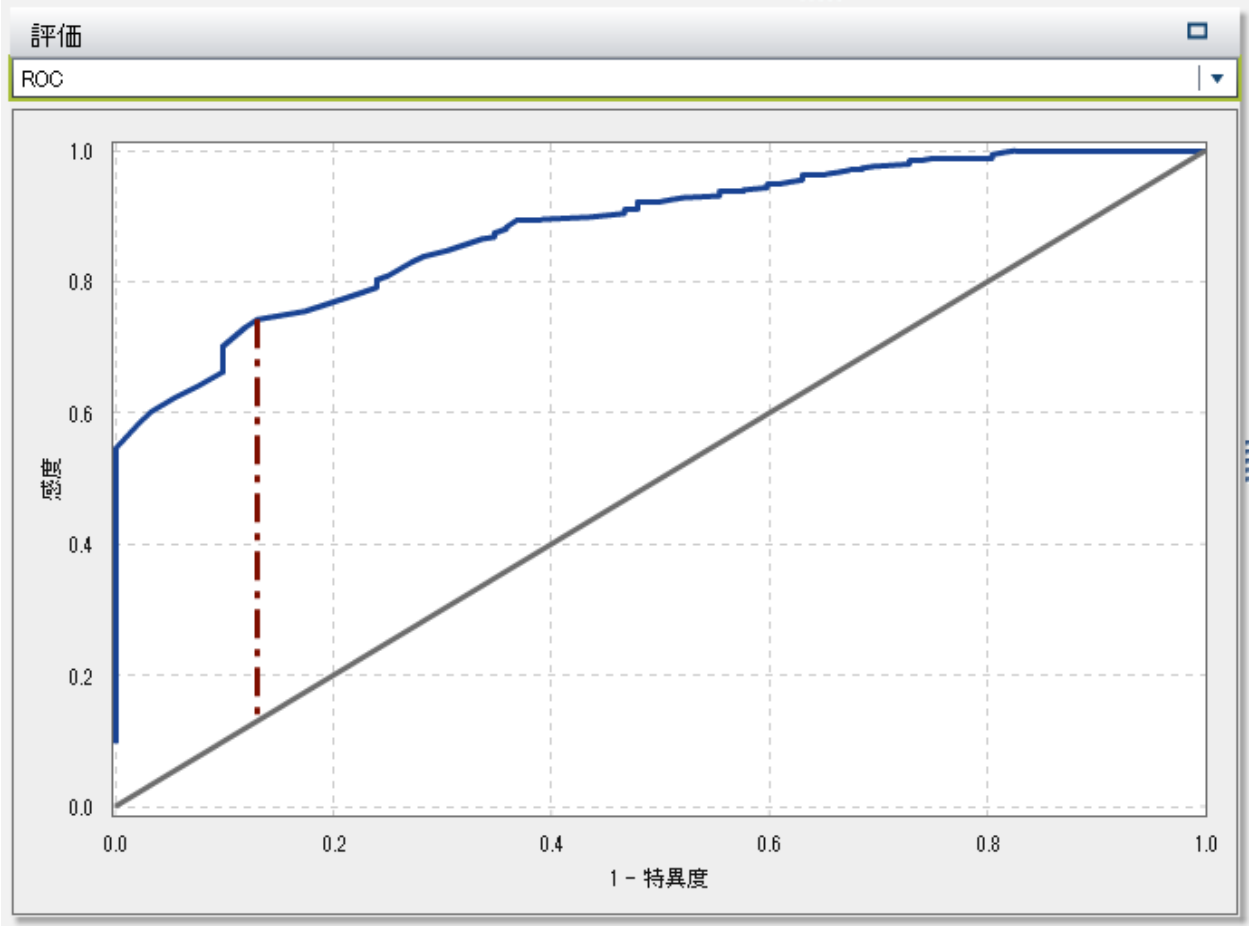
比較においては、リフトチャートは、入力データに関する完全な知識に基づく最良のモデルがプロットされます。

ROC

受信者操作特性(ROC)チャートは、偽陽性および偽陰性の分類を回避するモデルの能力を示します。偽陽性の分類とは、あるオブザベーションで、実際にはイベントがない(疾患がない: 陰性)ときに、イベントがある(疾患がある: 陽性)と識別されることをいいます(第一種(Type 1)過誤とも呼ばれます)。偽陰性の分類とは、あるオブザベーションで、実際にはイベントがある(疾患がある: 陽性)ときに、イベントがない(疾患がない: 陰性)と識別されることをいいます(第二種(Type II)過誤とも呼ばれます)。

このモデルの *特異度*は、真の陰性率です。偽陽性率を導出するには、1 から特異度を減算します。**1 – Specificity** というラベルが付けられた偽陽性率は、ROC チャートの X 軸です。モ

デルの *感度*は、真の陽性率です。これは、ROC チャートの Y 軸です。したがって、ROC チャートでは、偽陽性率の変化に伴う真の陽性率の変化がプロットされます。



良い ROC チャートは、最初に非常に急な勾配があり、すぐに横ばいになります。すなわち、オブザベーションの誤分類より、かなり多い数のオブザベーションが正しく分類されていることがわかります。偽陽性も偽陰性もない完璧なモデルの場合、ROC チャートは(0,0)で開始し、(0,1)に垂直に推移してから、(1,1)で水平になります。この例では、1 つの誤分類が発生するまでは、モデルはすべてのオブザベーションを正しく分類しています。

ROC チャートには、ROC チャートの解釈に役立つ 2 つの線が含まれています。最初の線は、1 の勾配を持つベースラインモデルです。この線は、オブザベーションを誤分類するのと同じ比率で正しく分類するモデルを模倣しています。理想的な ROC チャートは、ベースラインモデルと ROC チャート間の距離を最大化します。オブザベーションを正しく分類するよりも多い比率で誤分類するモデルは、ベースラインモデルの基準に達していません。2 番目の線は、偽

陽性率の垂直線です。この線では、ROC チャートとベースラインモデルの Kolmogorov-Smirnov 値間の差異が最大になります。

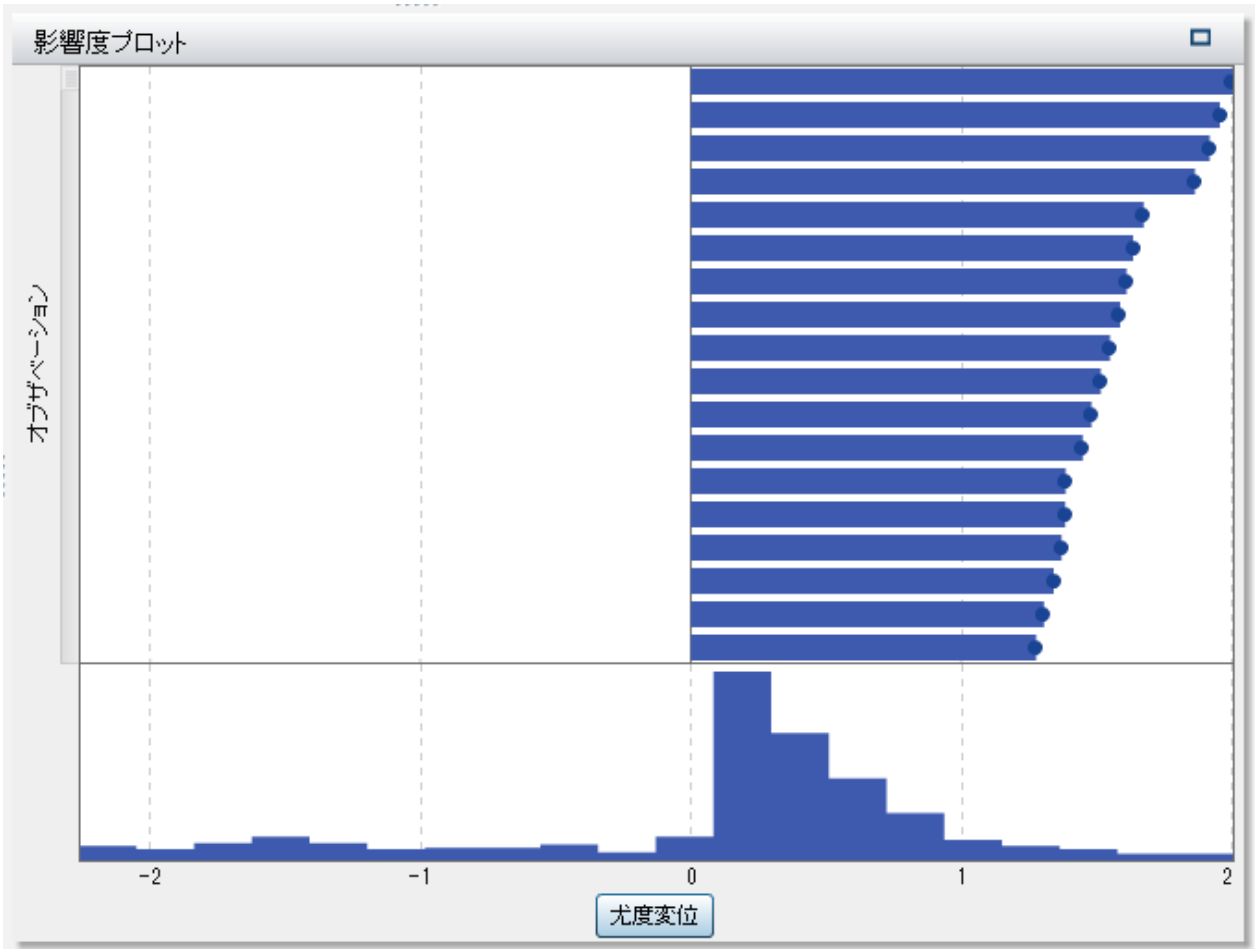
誤分類

誤分類プロットには、応答変数の値ごとに、正しく分類されたオブザベーションと誤分類されたオブザベーションの数が示されています。次に示すように応答変数がバイナリ分布でない場合は、ロジスティック回帰分析モデルでは、イベントでないすべての水準が等しいとみなされます。

誤分類が著しく多い場合は、モデルがデータに当てはまっていないことを示していることがあります。

Influence Plot

Influence Plot には、オブザベーションごとに計算されるいくつかの測定値が表示されます。入力データに大量のオブザベーションが含まれる場合、オブザベーションはビンに分割されます。デフォルトでは、X 軸には**尤度変位**値がプロットされます。



利用可能なその他の値は、CBAR、逸脱度の変化、Pearson 値の変化です。

これらの値を使用すると、回帰分析の予測モデルに影響する外れ値や他のデータ点の識別に役立ちます。

当てはめ統計量

ロジスティック回帰分析モデルでは、データに対するモデルの適合度の評価に役立ついくつかの評価尺度が計算されます。これらの評価尺度は、モデルペインの上部にあります。利用可能なすべての評価尺度を表示するには、現在表示されている評価尺度をクリックします。

-2 Log Likelihood

尤度関数では、想定し得るすべてのパラメータ値を使用したサンプルの観測値の確率が推定されます。対数尤度は、文字通り尤度関数の対数です。尤度関数値は、対数尤度の-2 倍です。小さな値ほど、望ましい値です。

AIC

赤池情報量規準。小さな値ほど良いモデルであることを示します。AIC 値は負の値になることがあります。AIC は、応答変数の真の分布とモデルで指定された分布との間の差異の Kullback-Leibler 情報量尺度に基づいて導出されます。

AICC

補正赤池情報量規準。AIC のこのバージョンでは、サンプルサイズを説明するために値を調整します。その結果、追加の効果により、AICC には AIC より大きなペナルティ(罰則)が課されます。サンプルサイズが大きくなるに従い、AICC と AIC が収束します。

BIC

ベイズ情報量規準(BIC)は、シュワルツのベイズ規準(SBC)とも呼ばれ、モデルの残差平方和と効果の数の増加関数です。応答変数の説明されないばらつきと効果の数によって BIC の値は増加します。このため、BIC が低いほど、説明変数が少ないか、適合度が高い、あるいはその両方を示しています。BIC では、自由度のパラメータに対して AIC より大きなペナルティが課されます。

Max-rescaled R-Square

R2 乗値の観測値を取得し得る最大の R2 乗値で除算した値です。この値は、複数のカテゴリ独立変数がある場合に有効です。値は、0~1 の範囲をとります。1 に近いほど望ましい値です。

オブザベーション

モデルで使用されているオブザベーションの数。

R2 乗値

R2 乗値は、モデルがデータにどの程度当てはまっているかを示す指標です。R2 乗値は、0~1 の範囲をとります。1 に近いほど望ましい値です。

要約テーブル

モデルペインの上部にある**要約テーブルの表示**をクリックすると、モデルペインの下部に要約パネルが表示されます。要約テーブルには、次の情報が含まれています。

次元

モデルで使用される効果変数の概要です。このタブでは、そのモデルに選択された尺度および分類効果の数、交差積行列のランク、読み込まれているオブザベーションの数、モデルで使用されているオブザベーションの数などを確認できます。

反復履歴

関数および勾配の収束結果です。このタブには、反復、関数および勾配が収束する位置が示されます。

収束

収束の理由を示します。

当てはめ統計量

前のセクションに記載されているすべての当てはめ統計量がリストされています。

Type III 検定

Type III 検定の詳細が表示されています。Type III 検定では、それぞれの部分的な効果の有意性をモデルの他のすべての効果とともに調べます。詳細については、*SAS/STAT User's Guide* の"The Four Types of Estimable Functions"を参照してください。

パラメータ推定値

モデルの各パラメータの推定値を示します。

応答プロファイル

モデルのイベント(陽性)とイベントでない(陰性)のカウントを表示します。

6

一般化線形モデル

一般化線形モデルの概要

一般化線形モデル(GLM)は、従来の線形モデルを拡張したものです。GLM を使用すると、非線形リンク関数を介して母平均を線形予測子に依存させることができます。GLM では、分布とリンク関数を指定する必要があります。指定する分布は、応答変数の分布と一致している必要があります。リンク関数は、応答変数を効果変数に相関させるために使用します。

GLM では、尺度応答変数と少なくとも 1 つの効果変数または交互作用項が必要です。分布によって、尺度応答変数に範囲要件が課されます。各分布の範囲要件を次の表に示します。

分布	範囲要件
ベータ分布	値は、0～1 の範囲(0 と 1 を含まない)である必要があります。

分布	範囲要件
バリナリ分布	2 つの個別値
指数分布	負でない実数
ガンマ分布	負でない実数
幾何分布	正の整数
逆 Gauss 分布	正の実数
負の二項分布	負でない整数
正規分布	実数
ポアソン分布	負でない整数

一般化線形モデルのプロパティ

GLM では、次のプロパティを使用できます。

名前
このモデルの名前を指定します。

有用な欠損
情報のある欠測アルゴリズムを使用するかどうかを指定します。詳細については、欠損値 (28 ページ)を参照してください。

分布
応答変数のモデル化に使用する分布を指定します。

リンク関数
線形モデルを応答変数の分布に相関させるために使用するリンク関数を指定します。利用可能なリンク関数は、分布ごとに異なります。各分布に利用できるリンク関数を次の表に示します。

分布	利用可能なリンク関数
ベータ分布	ロジット、プロビット、両対数(Log-log)、C 両対数(C-log-log)
バリナリ分布	ロジット、プロビット、両対数(Log-log)、C 両対数(C-log-log)
指数分布	対数、恒等
ガンマ分布	対数、恒等、逆数
幾何分布	対数、恒等
逆 Gauss 分布	べき乗(-2)、対数、恒等
負の二項分布	対数、恒等
正規分布	対数、恒等
ポアソン分布	対数、恒等

収束

- **関数収束のオーバーライド**を使用すると、関数の収束値を手動で指定できます。
- **関数収束のオーバーライド**が選択されている場合は、**Value** を使用して、関数の収束値を指定します。より大きな値を指定すると、モデルはより早く収束します。これにより、モデルの学習に費やされる時間を削減できますが、準最適な(最適な水準に達しない)モデルが作成されることがあります。
- **Override gradient convergence** を使用すると、勾配の収束値を手動で指定できます。
- **Override gradient convergence** が選択されている場合は、**Value** を使用して、勾配の収束値を指定します。より大きな値を指定すると、モデルはより早く収束します。これにより、モデルの学習に費やされる時間を削減できますが、準最適な(最適な水準に達しない)モデルが作成されることがあります。

- **最大反復回数**では、モデルの学習中に実行される最大反復回数を指定します。比較的小さな値を指定すると、モデルの学習に費やされる時間を削減できますが、準最適な(最適な水準に達しない)モデルが作成されることがあります。

注: 勾配の収束または関数の収束条件を指定した場合、モデルが、指定した条件に到達する前に内部的な収束条件に基づいて収束する可能性があります。収束の理由については、要約テーブルの**収束**タブで参照できます。

評価

- **Use default number of bins** では、デフォルトのビン数を使用するか、独自の値を設定するかを指定します。デフォルトでは、尺度変数は、20 のビンにグループ化されます。
- **Use default number of bins** プロパティが選択されていない場合には、**Number** に、使用するビン数を指定します。5～100 の範囲の整数値を指定する必要があります。
- **Tolerance** には、パーセンタイル値を推定する反復アルゴリズムの収束の決定に使用する許容値を指定します。アルゴリズムの精度を高めるには、より小さな値を指定します。

Show diagnostic plots

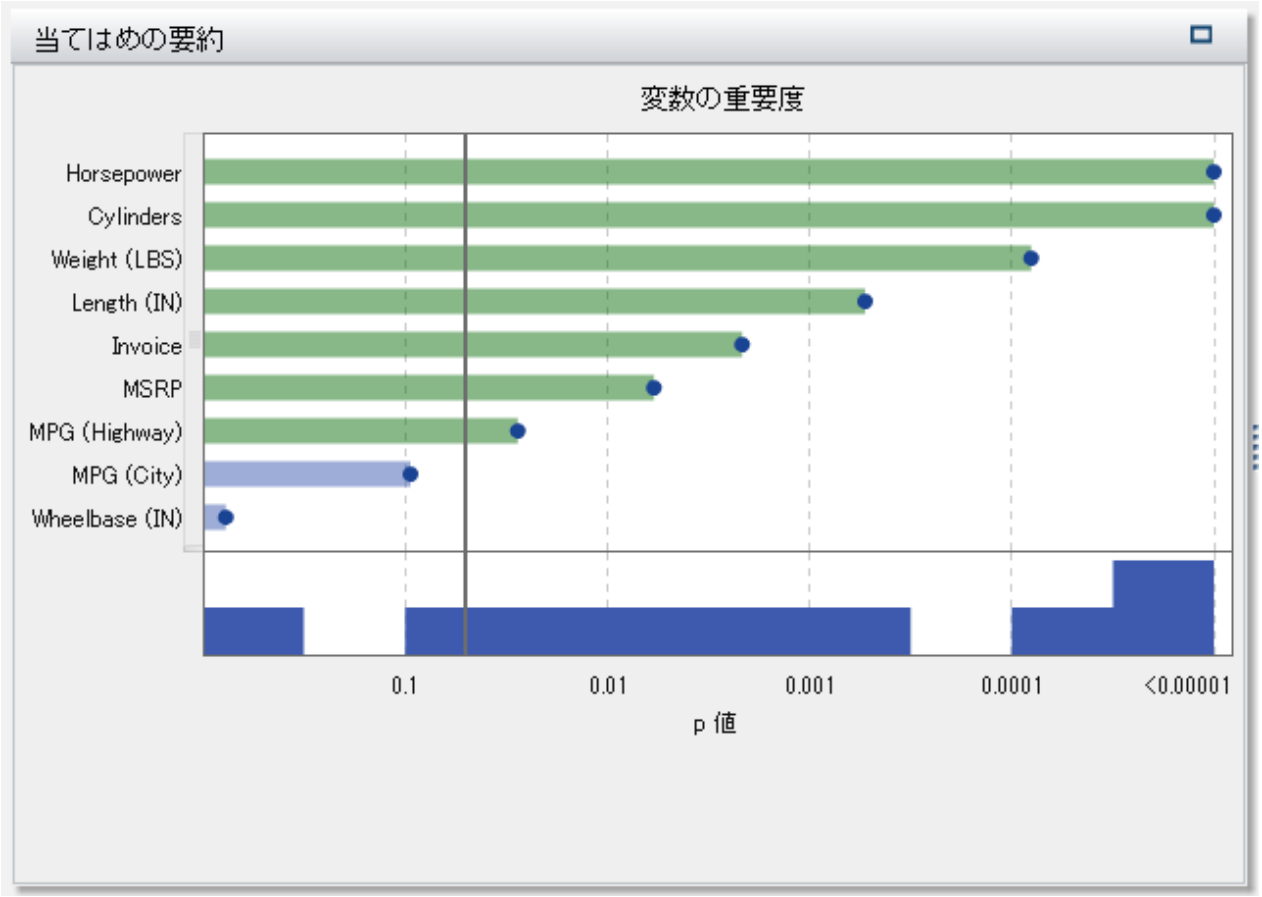
Residual Plot ウィンドウおよび**評価**ウィンドウをモデルペインに表示するかどうかを指定します。

一般化線形モデルの結果ウィンドウ

当てはめの要約ウィンドウ

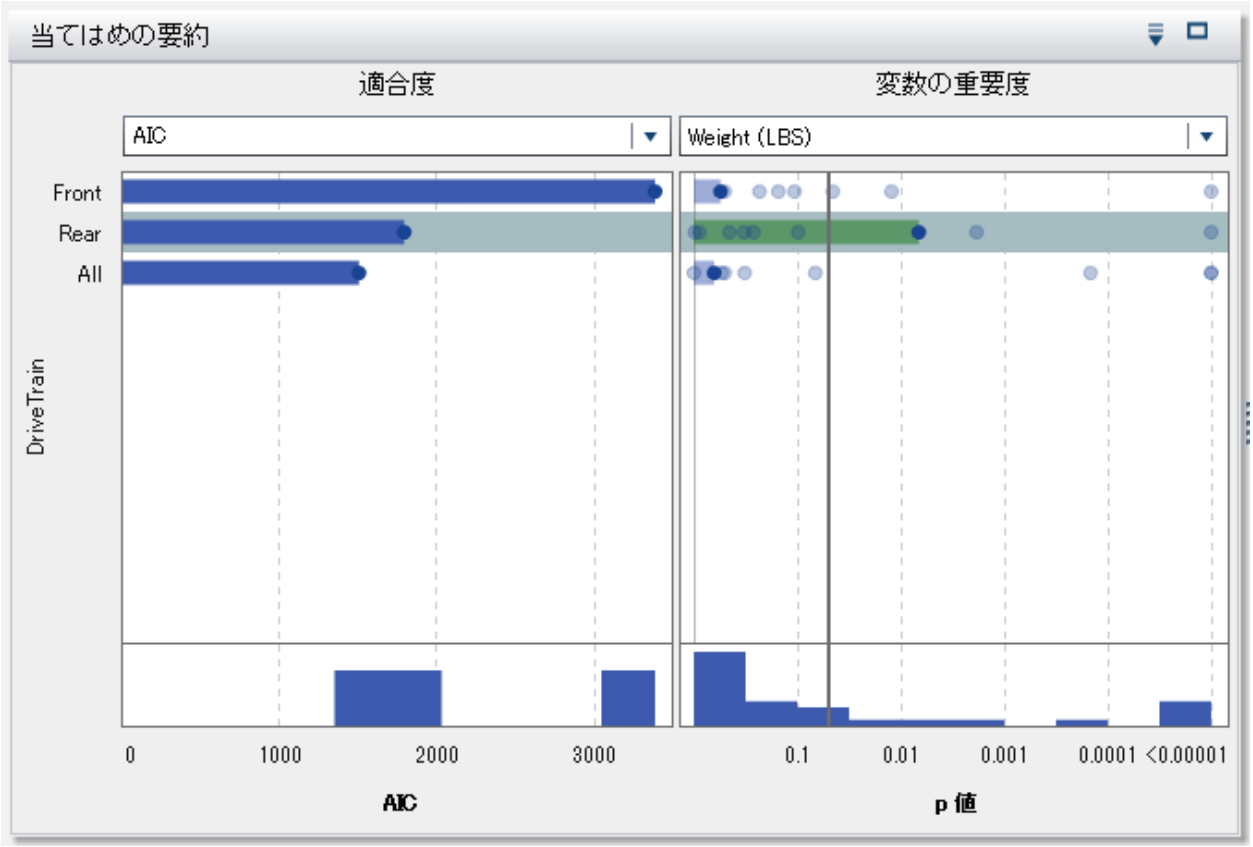
Group By 変数を使用しない場合

当てはめの要約ウィンドウには、*p* 値によって測定された各変数の相対的な重要度がプロットされます。*p* 値は、対数目盛り上にプロットされ、アルファ値(-対数(アルファ)としてプロット)は、垂直線で示されます。アルファ値を調整するには、垂直線をクリックしてドラッグします。*p* 値のヒストグラムは、ウィンドウの下部に表示されます。**当てはめの要約**ウィンドウの例を次に示します。



Group BY 変数を使用する場合

分析に Group BY 変数を含める場合は、**当てはめの要約**ウィンドウには、異なるプロットが表示されます。



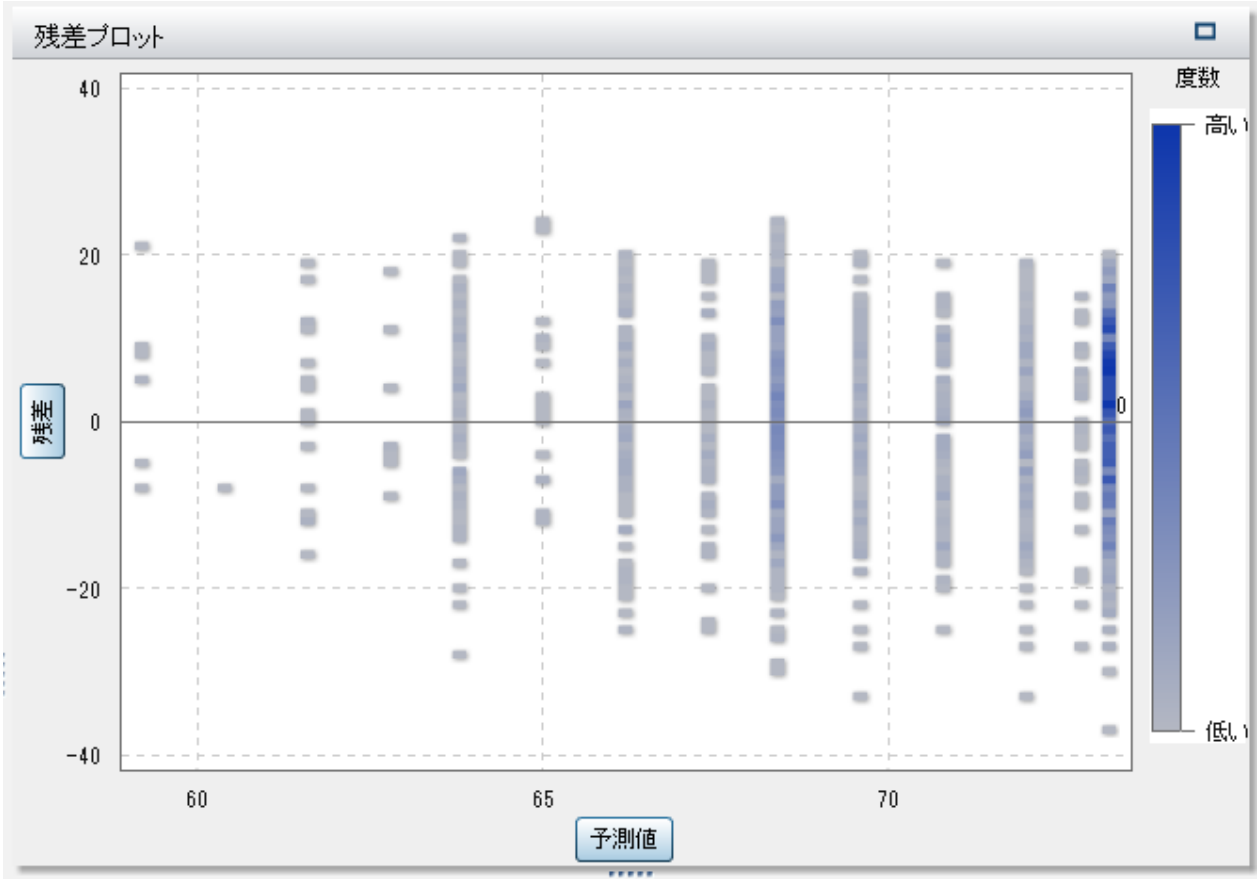
まず、**Variable Importance** プロットに単一の変数のみが表示されていることに注目してください。これは、すべての変数について、変数の重要度が Group BY 変数の各レベルで計算されるからです。異なる効果の変数の重要度を表示するには、ドロップダウンメニューを使用します。次に、Group BY 変数を使用しない場合にはなかった **Goodness of Fit** プロットが表示されていることに注目してください。このプロットは、Grooup BY 変数の各レベルでのモデルによる応答変数の予測の適合度を示すものです。このプロットを使用すると、作成したモデルによる予測の適合度が異なるレベルで大幅に異なるかどうかを判定できます。

プロットの並べ替え方法を指定するには、 アイコンを使用します。

Residual Plot

散布図

オブザベーションの残差とは、応答値の予測値と実測値の差です。次に示すように、デフォルトで **Residual Plot** には、予測値に対する残差の散布図が表示されます。



Y 軸と X 軸上にあるラベルはボタンです。これらのボタンのいずれかをクリックすると、その軸にプロットされている値を変更できます。Y 軸の場合、残差または標準化 Pearson 残差をプロットできます。標準化 Pearson 残差は、Pearson のカイ 2 乗検定に対する個々の寄与度です。X 軸の場合は、予測値または線形予測子をプロットできます。

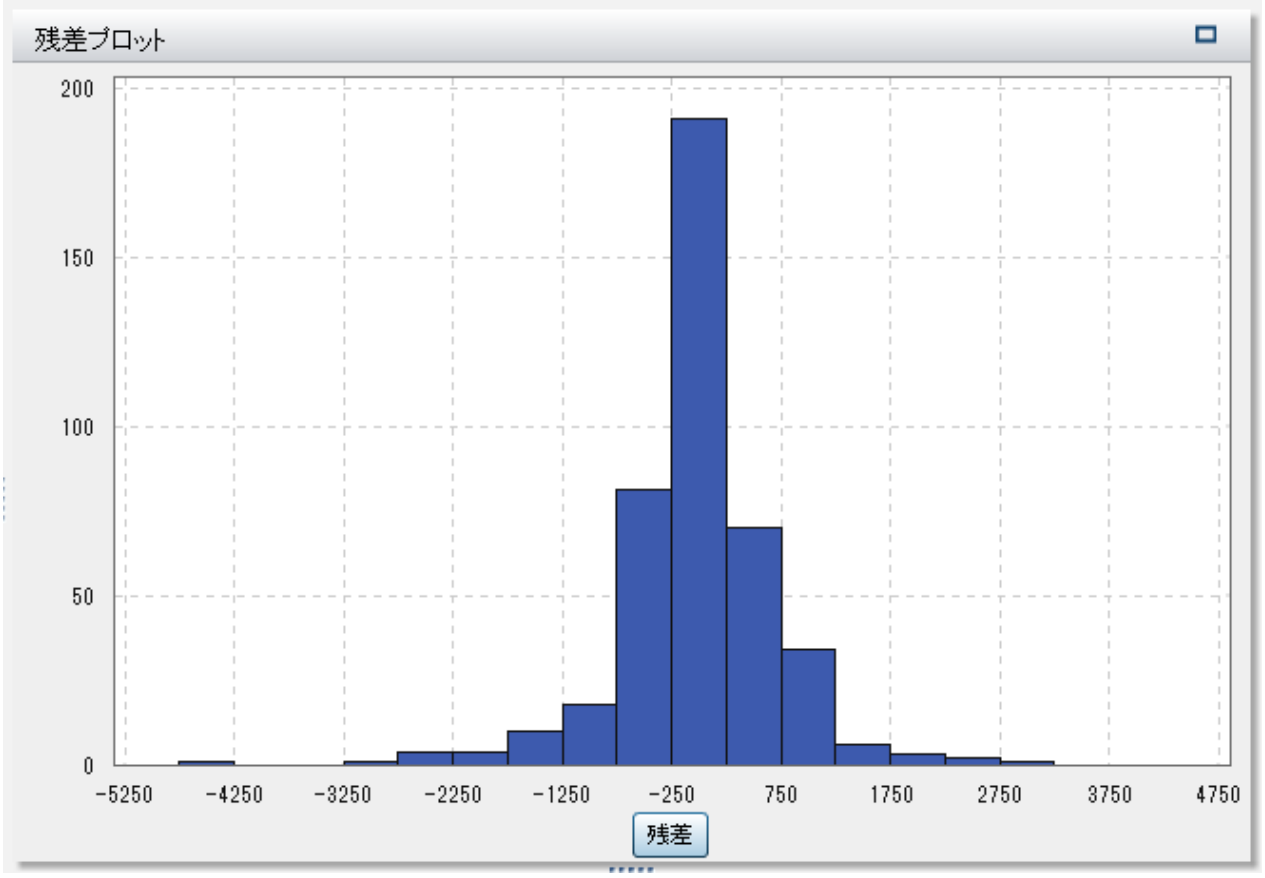
残差プロットには、作成したモデルを調べる場合に使えるいくつかの用途があります。第 1 に、残差プロットの明確なパターンは、そのモデルがデータに当てはまっていない可能性があることを示しています。第 2 に、残差プロットでは、予測値に対する残差をプロットする場合に、入力データの非定常分散を検出できます。非定数分散は、予測値が変化するにつれて残差値の相対的な散らばりが変化する場合に明らかです。第 3 に、他の方法と組み合わせると、残差プロットはデータの外れ値を識別する上で役立ちます。

非常に大量のデータセットを使用する場合、残差プロットは、実際のデータのプロットではなく、ヒートマップの形式で表示されます。ヒートマップでは、オブザベーションの実測値がビンに分

割され、各ポイントの色は、そのビン内にあるオブザベーションの数を示します。前述のヒートマップを参照してください。

ヒストグラム

残差プロットのデータをヒストグラムとして表示するには、**Residual Plot** ウィンドウで右クリックして、**Use Histogram** を選択します。残差プロットの Y 軸で利用できた各値は、ヒストグラムとして利用できます。



プロットされる値を変更するには、X 軸にあるラベルをクリックします。残差または標準化 Pearson 残差を選択できます。

残差の分布が正規近似であるか非対称な分布であるかは、ヒストグラムでかなり簡単に確認できます。たとえば、前の画像では、標準化 Pearson 残差は対称でした。非正規の残差ヒストグラムは、モデルがデータに当てはまっていないことを示している場合があります。

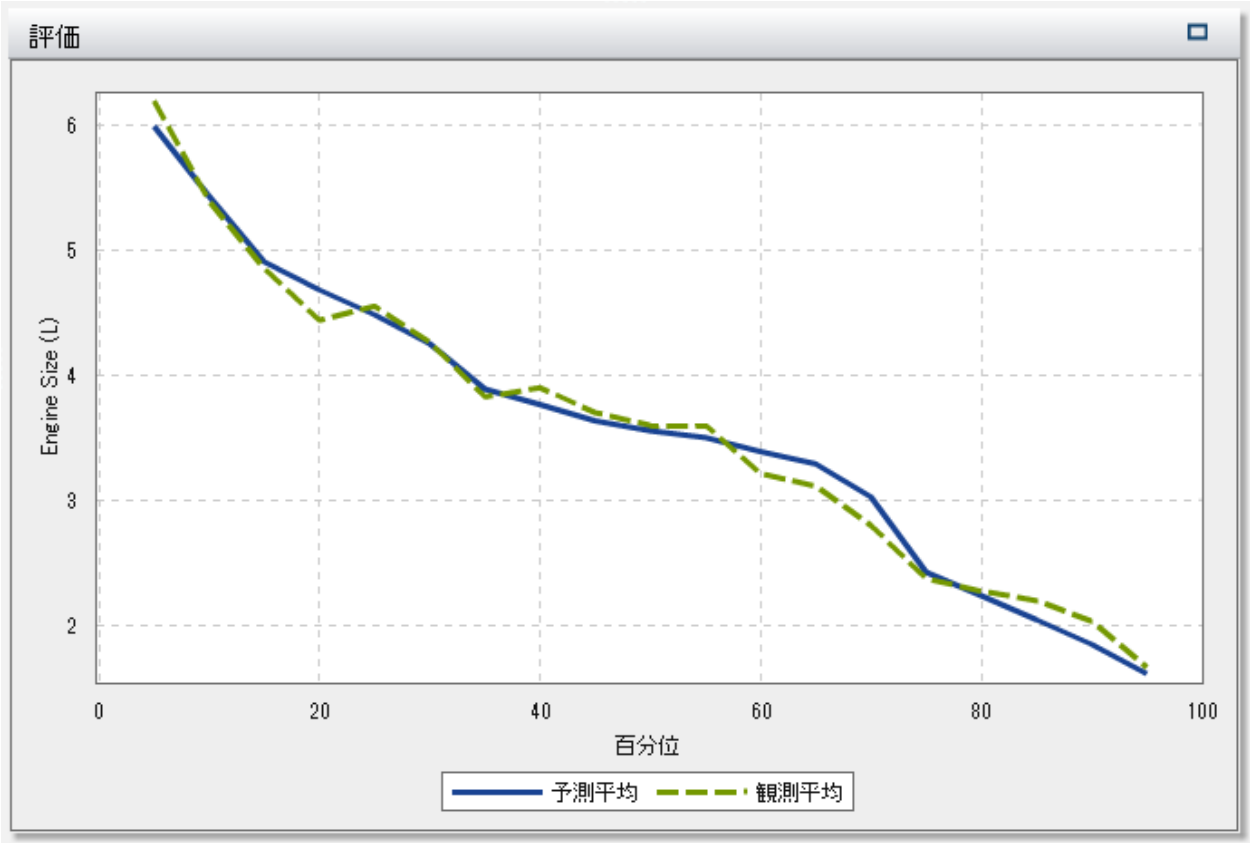
箱ひげ図

現在プロットされている残差の箱ひげ図を表示するには、**Residual Plot** ウィンドウで右クリックして **Plot By** を選択します。次に、箱ひげ図作成時に残差のグループ化に使用するカテゴリ変数を選択します。モデルに含まれているか否かに関係なく、すべてのカテゴリ変数を利用できます。モデルに含まれていない変数の場合は、**Box Plot** ウィンドウを右クリックして、各変数を分類効果または Group BY 変数として割り当てます。**Assign to Classification** メニュー項目および **Assign to Group By** メニュー項目は、モデルにすでに含まれている変数には利用できません。

外れ値は、デフォルトでは表示されません。外れ値を表示するには、**Box Plot** ウィンドウで右クリックし、**Show Outliers** を選択します。

評価

GLM の場合、**評価**ウィンドウには、ビンに分割されたデータに対する応答の予測値平均と観測値平均がプロットされます。モデルに強いバイアスがあるかを検出するには、このプロットを使用します。予測値平均と観測値平均の大きな違いは、バイアスを示すことができることです。



当てはめ統計量

GLM では、データに対するモデルの適合度の評価に役立ついくつかの評価尺度が計算されます。これらの評価尺度は、モデルペインの上部にあります。利用可能なすべての評価尺度を表示するには、現在表示されている評価尺度をクリックします。利用可能な評価尺度は、次のとおりです。

-2 Log Likelihood

尤度関数では、想定し得るすべてのパラメータ値を使用したサンプルの観測値の確率が推定されます。対数尤度は、文字通り尤度関数の対数です。この値は対数尤度の-2 倍です。小さな値ほど、望ましい値です。

AIC

赤池情報量規準。小さな値であるほど、良いモデルであることを示しています。AIC 値は、2 つのモデルにほとんど同じ数のオブザベーションがある場合にのみ比較する必要があり

ます。AIC 値は、負の値になることがあります。AIC は、応答変数の真の分布とモデルで指定された分布との間の差異の Kullback-Leibler 情報量尺度に基づいて導出されます。

AICC

補正赤池情報量規準。AIC のこのバージョンでは、比較的小さなサンプルサイズを説明するために値を調整します。その結果、追加の効果により、AICC には AIC より大きなペナルティ(罰則)が課されます。サンプルサイズが大きくなるに従い、AICC と AIC が収束します。

BIC

ベイズ情報量規準(BIC)は、シュワルツのベイズ規準(SBC)とも呼ばれ、モデルの残差平方和と効果の数の増加関数です。応答変数の説明されないばらつきと効果の数によって BIC の値は増加します。このため、BIC が低いほど、説明変数が少ないか、適合度が高い、あるいはその両方を示しています。BIC では、自由度のパラメータに対して AIC より大きなペナルティが課されます。

オブザベーション

モデルで使用されているオブザベーションの数。

要約テーブル

モデルペインの上部にある**要約テーブルの表示**をクリックすると、モデルペインの下部に要約パネルが表示されます。要約テーブルには、次の情報が含まれています。

次元

モデルで使用される効果変数の概要です。このタブでは、そのモデルに選択された尺度および分類効果の数、交差積行列のランク、読み込まれているオブザベーションの数、モデルで使用されているオブザベーションの数などを確認できます。

反復履歴

関数および勾配の反復結果です。このタブには、目的(尤度)関数の値、その値の変化および最大勾配が示されています。

収束

収束の理由を示します。

当てはめ統計量

前のセクションに記載されているすべての当てはめ統計量がリストされています。

Type III 検定

Type III 検定の詳細が表示されています。Type III 検定では、それぞれの部分的な効果の有意性をモデルの他のすべての効果とともに調べます。詳細については、*SAS/STAT User's Guide* の“The Four Types of Estimable Functions”を参照してください。

パラメータ推定値

モデルの各パラメータの推定値を示します。

7

決定木

決定木の概要

決定木では、各オブザベーションに適用された一連のルールに基づいて入力データの階層状のセグメントが作成されます。各ルールにより、1 つの効果の値に基づいて、セグメントにオブザベーションが割り当てられます。ルールは順次適用され、各セグメント内にセグメントの階層が作成されます。この階層はツリーと呼ばれ、各セグメントは*ノード*と呼ばれます。最初のセグメントには、データセット全体が含まれています。このセグメントは、*ルート(根)ノード*と呼ばれます。ノードとそのすべての後続ノードは、*ブランチ(枝)*を形成しています。最後のノードは、*リーフ(葉)*と呼ばれます。リーフごとに、応答変数に関する決定が行われ、そのリーフ内のすべてのオブザベーションに適用されます。厳密な決定内容は、応答変数によって変わります。

決定木には、尺度応答変数またはカテゴリ応答変数、および少なくとも 1 つの予測子が必要です。予測子には、カテゴリ変数または尺度変数を使用できますが、交互作用項は使用できません。

決定木では、対話モードに入ることにより、決定木の学習または刈り込みを手動で行うことができます。対話モードでは、使用中の応答変数または予測子は変更できません。また、モデルのスコアコードはエクスポートできません。対話モードに入るには、**Tree** ウィンドウで、決定木への変更を開始するか、右ペインの**役割**タブで**対話型モードの使用**をクリックします。対話モードを終了するには、**役割**タブで**非対話型モードの使用**をクリックします。

注: 対話モードを終了すると、行った変更はすべて失われます。

決定木のプロパティ

決定木では、次のプロパティを使用できます。

名前
このモデルの名前を指定します。

最大枝数
ノードを分岐する場合に許容されるブランチ(枝)の最大数を指定します。

最大レベル
決定木の最大深度を指定します。

Leaf size
1 つのリーフノードに割り当てることができるオブザベーションの最大数を指定します。

Response bins
尺度応答変数のカテゴリ分けに使用するビン数を指定します。

Predictor bins
尺度変数である予測子のカテゴリ分けに使用するビン数を指定します。

Pruning
ツリーの刈り込みアルゴリズムの刈り込みの強度を指定します。より刈り込みが強いアルゴリズムは、小さな決定木を作成します。大きな値ほど、より刈り込みが強いということになります。

Rapid growth
情報利得比と k 平均法による高速検索方式を使用して決定木を拡大できます。無効になっている場合は、情報利得と貪欲法による検索方式が使用されます。この場合、一般的に大きなツリーが生成されるため、より多くの時間を要します。

Include missing
欠損値のあるオブザベーションを含めることができます。カテゴリ変数の場合、欠損値は変数自身の階層に割り当てられます。尺度変数の場合、欠損値は、最小の利用可能なマシン値(負の無限大)に割り当てられます。

予測因子の再利用
予測子に基づいて、同じブランチ(枝)で 2 つ以上の分岐が可能となります。

Frequency
ノードに含まれるオブザベーションの数または割合をノードで通知するかを指定します。

評価

- **Use default number of bins** では、デフォルトのビン数を使用するか、独自の値を設定するかを指定します。デフォルトでは、尺度変数は、20 のビンにグループ化されます。
- **Use default number of bins** プロパティが選択されていない場合には、**Number** に、使用するビン数を指定します。5～100 の範囲の整数値を指定する必要があります。
- **Prediction cutoff** では、計算される確率がイベントとみなされる値を指定します。
- **Tolerance** には、パーセンタイル値を推定する反復アルゴリズムの収束の決定に使用する許容値を指定します。アルゴリズムの精度を高めるには、より小さな値を指定します。

Show diagnostic plots
リーフの統計量ウィンドウおよび**評価**ウィンドウをモデルペインに表示するかどうかを指定します。

ツリーの概要を表示
ツリーの概要を表示します。ツリーの概要を使用すると、大きな決定木でもすばやく移動できます。決定木の特定の領域を拡大表示すると、ツリーの概要には、決定木全体が表示され、拡大表示している領域が強調表示されます。決定木の表示を変更するには、強調表示されている領域をクリックしてドラッグします。決定木全体を表示するには、ツリーの概要

の左上隅にあるアイコンをクリックします。ツリーの概要を最小化するには、ツリーの概要の左上隅にあるアイコンをクリックします。

情報利得と利得比の計算

Rapid growth プロパティが有効になっている場合、ノードの分岐の一部は、情報利得ではなく、情報利得比によって決定されます。このセクションでは、情報利得と情報利得比の算出方法のほか、その利点と欠点について説明します。これらの説明では、属性は、分類変数または尺度変数のビンの特定の測定レベルとみなしています。

情報利得法では、どの属性が情報利得を最大にするかに基づいて分岐を選択します。利得は、ビット単位で測定されます。この方法を使用すると、良い結果が得られますが、属性数が多い変数での分岐が有利になります。情報利得比法では、分岐の値を組み込んで、その分岐に実際に価値のある情報利得の比率を決定します。最大の情報利得比をもつ分岐が選択されます。

情報利得の計算は、学習データの情報の決定から始まります。応答値の情報 *r* は、次の式で計算されます。

$$-\log_2\left(\frac{freq(r,\ T)}{|\ T\ |}\right)$$

T は、学習データを表し、|*T*|は、オブザベーションの数を表しています。学習データの推定情報を決定するには、想定し得るすべての応答値について、この式を合計します。

$$I\left(T\right) = -\sum_{j=1}^{n} \frac{freq(r_j,\ T)}{|\ T\ |} \times \log_2\left(\frac{freq(r_j,\ T)}{|\ T\ |}\right)$$

ここで、*n* は、応答値の総数です。この数は、学習データの*エントロピー*としても参照されます。

次に、*m* 個の想定し得る属性を持つ変数 X での分岐 *S* について検討します。この分岐で提供される推定情報は、次の方程式により計算されます。

$$I_S\left(T\right) = \sum_{j=1}^{m} \frac{|\ T_j\ |}{|\ T\ |} \times I\left(T_j\right)$$

この方程式で、T_j は、$j^{番目}$の属性を含むオブザベーションを表します。

分岐 S の情報利得は、次の方程式により計算されます。

$$G(S) = I(T) - I_S(T)$$

情報利得比では、分岐情報値の導入による情報利得計算の修正が試行されます。分岐情報は、次の方程式により計算されます。

$$SI\left(S\right) = -\sum_{j=1}^{m} \frac{|T_j|}{|T|} \times \log_2\left(\frac{|T_j|}{|T|}\right)$$

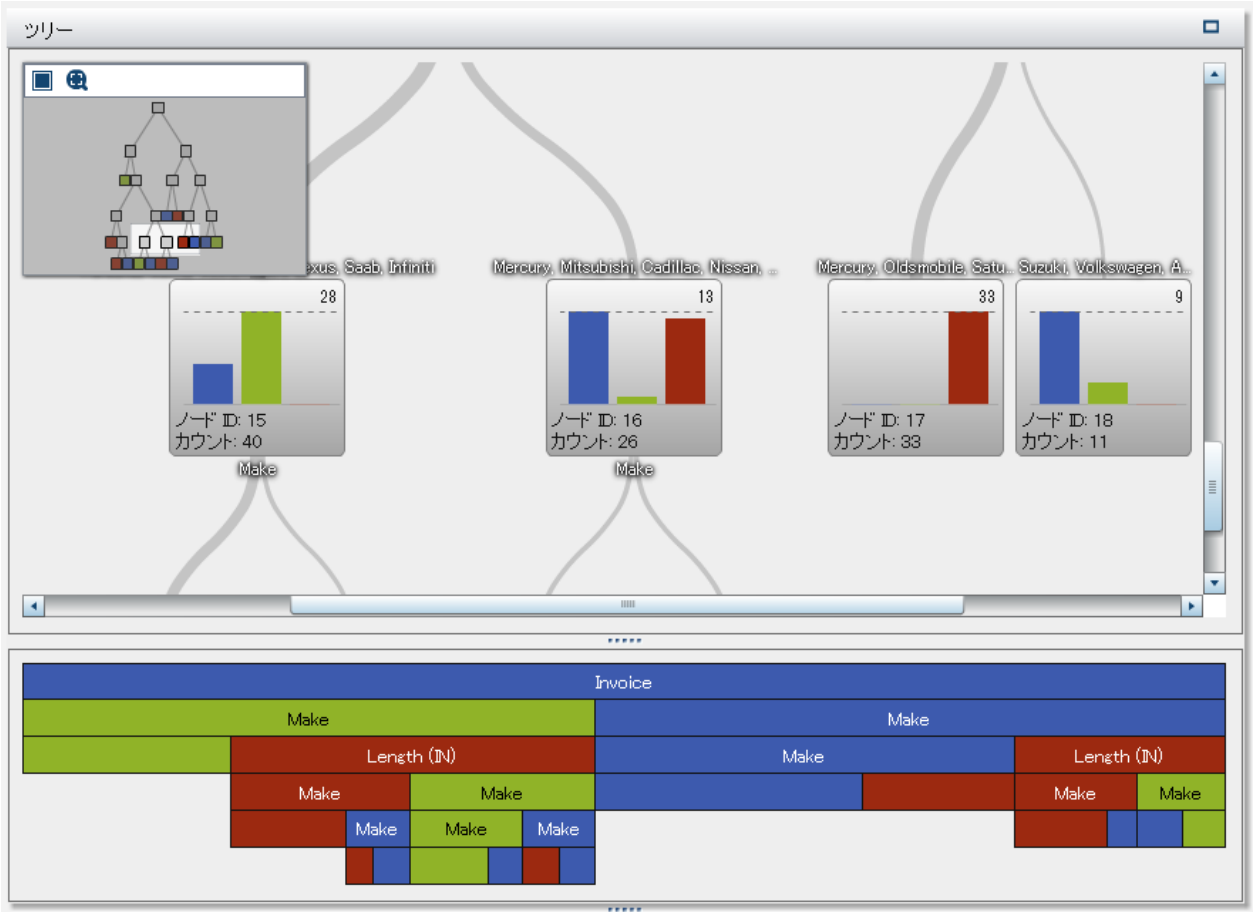
情報利得比とは、名前が示すとおり、分岐情報に対する情報利得の比です。

$$GR\left(S\right) = \frac{G(S)}{SI(S)}$$

決定木の結果ウィンドウ

Tree

Tree ウィンドウには、決定木、ツリーの概要、ツリーマップ(つららプロットともいう)が含まれています。



この決定木では、応答変数は 3 つの階層を持つカテゴリ変数です。ツリーの概要には、現在ユーザーが拡大表示している決定木の右側中央の領域が示されています。

ヒント 決定木を移動するには、マウスとキーボードを使用します。**Shift** キーを押しながら、**Tree** ウィンドウの任意の場所をクリックして、決定木をウィンドウ内で移動します。お使いのマウスのスクロールホイールを使用して決定木を拡大(ズームイン)および縮小(ズームアウト)します。拡大するには、スクロールアップし、縮小するには、スクロールダウンします。ズーム操作の中心は、カーソルの位置になります。

ツリーマップのノードの色は、そのノードの予測水準を示しています。ノードを決定木またはツリーマップで選択すると、対応するノードが他の場所で選択されます。リーフノードを選択すると、そのノードは、**リーフの統計量**ウィンドウで選択されます。凡例は、モデルペインの下部にあります。

応答変数が尺度変数である場合、予測したビンを示すにはグラデーションを使用します。濃い色ほど、大きな値であることを示しています。

ポップアップメニューを開くには、**Tree** ウィンドウで、ノードの外部を右クリックします。このメニューの最初の項目は、**リーフ ID 変数を派生**です。この項目をクリックすると、SAS Visual Statistics でオブザベーションごとにそのリーフ ID を含むカテゴリ変数が作成されます。この変数は、他のモデルで効果として使用できます。

別のポップアップメニューを開くには、ノードの内部を右クリックします。利用可能なメニューオプションは、リーフノードをクリックしたかどうかにより異なります。

リーフノードの場合、次のメニューオプションのいずれかを選択できます

Split

ディシジョンツリーの分割ウィンドウを開きます。ノードの分岐に使用する変数を選択するには、このウィンドウを使用します。選択した変数に基づいてノードを分岐するには、**OK** をクリックします。ノードを分岐しない場合は、**キャンセル**をクリックします。変数は、対数価値を基準として降順に並べ替えられます。

Split Best

Rapid growth が有効になっている場合は、最大の情報利得比を持つ変数に基づいてノードを分岐します。また、**Rapid growth** が無効になっている場合は、最大の情報利得を持つ変数に基づいてノードを分岐します。

学習

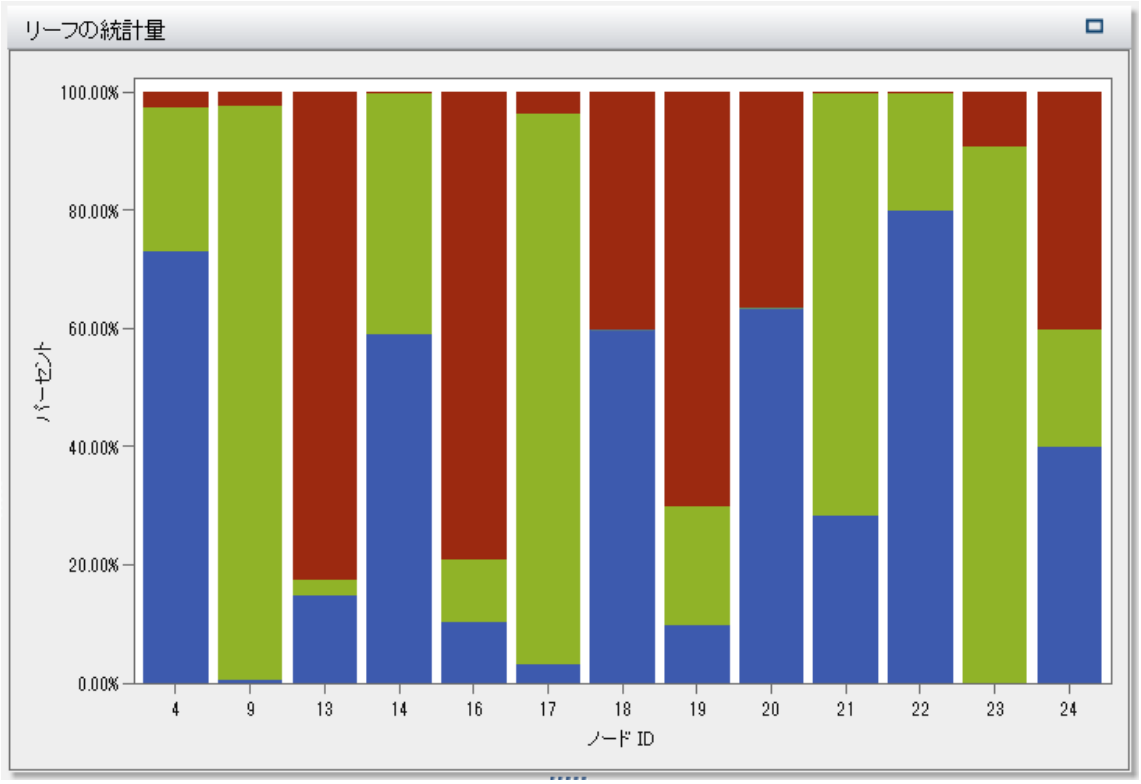
ディシジョンツリーの学習ウィンドウを開きます。リーフノードを 2 階層以上超えて学習するには、このウィンドウを使用します。最初に、学習に使用するすべての変数を選択します。**ディシジョンツリーの学習**ウィンドウで選択した変数のみ、学習で利用できます。**サブツリーの最大深さ**プロパティで学習の最大深度を指定します。決定木の学習を実行するには、**OK** をクリックします。

他のノードの場合、選択したノードに続くすべてのノードを削除するには、**剪定**を選択します。これにより、選択したノードはリーフノードに変わります。ノードを刈り込んだ後に元に戻すには、**復元**を選択します。

リーフの統計量

リーフの統計量ウィンドウでは、各リーフノードの各オブザベーションの割合をプロットします。ノードで最も一般的な水準は、そのノードに割り当てられている予測値です。2 つ以上のほぼ

同じ水準に相当する統計量を含むリーフノードは、追加の学習を実行することにより効果が得られます。

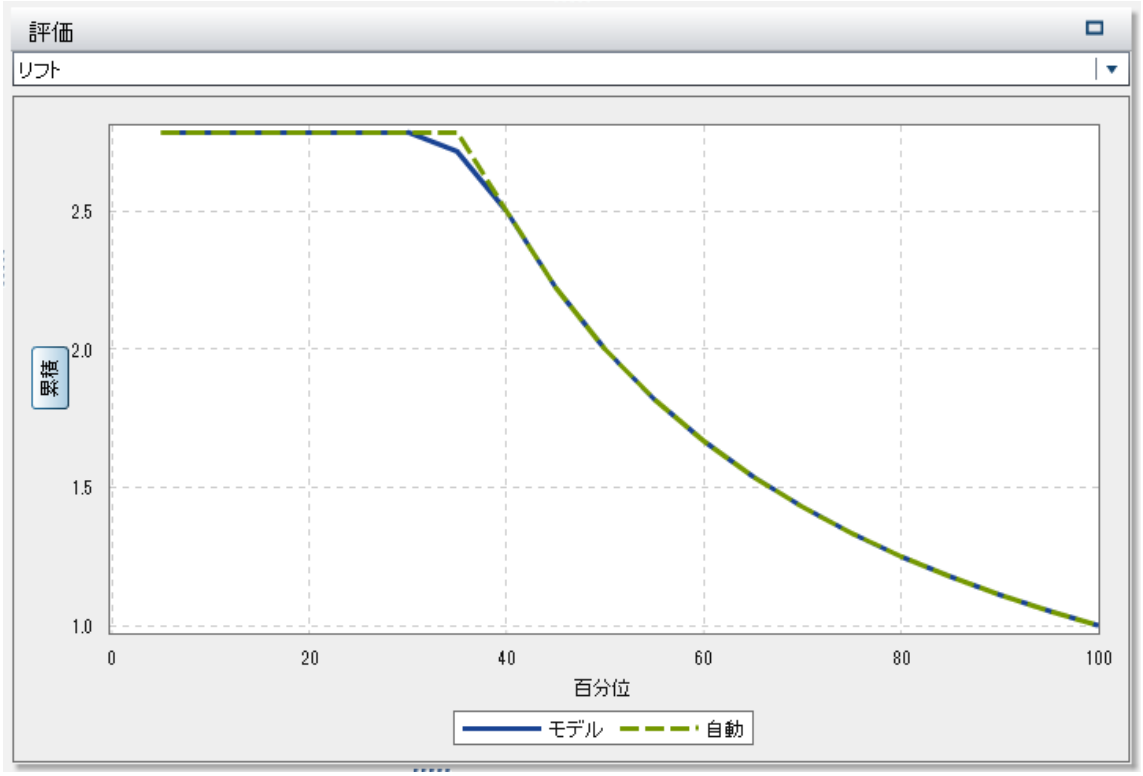


リーフの統計量ウィンドウで任意の列を選択すると、**Tree** ウィンドウで、対応するリーフが選択されます。

評価

リフト

*リフト*とは、モデルの応答の平均割合に対し、パーセンタイルビン内に捕捉された応答のパーセントの比率です。同様に、*累積リフト*とは、現在のパーセンタイルビンのすべてのデータを使用して計算されたものです。



リフトチャートには、**Model** 線が 1 より大きい場合に適合度が高いことが示されています。最初の 10～20 パーセンタイルでは、モデルは応答平均値と比較しても、高いリフト値を示しています。

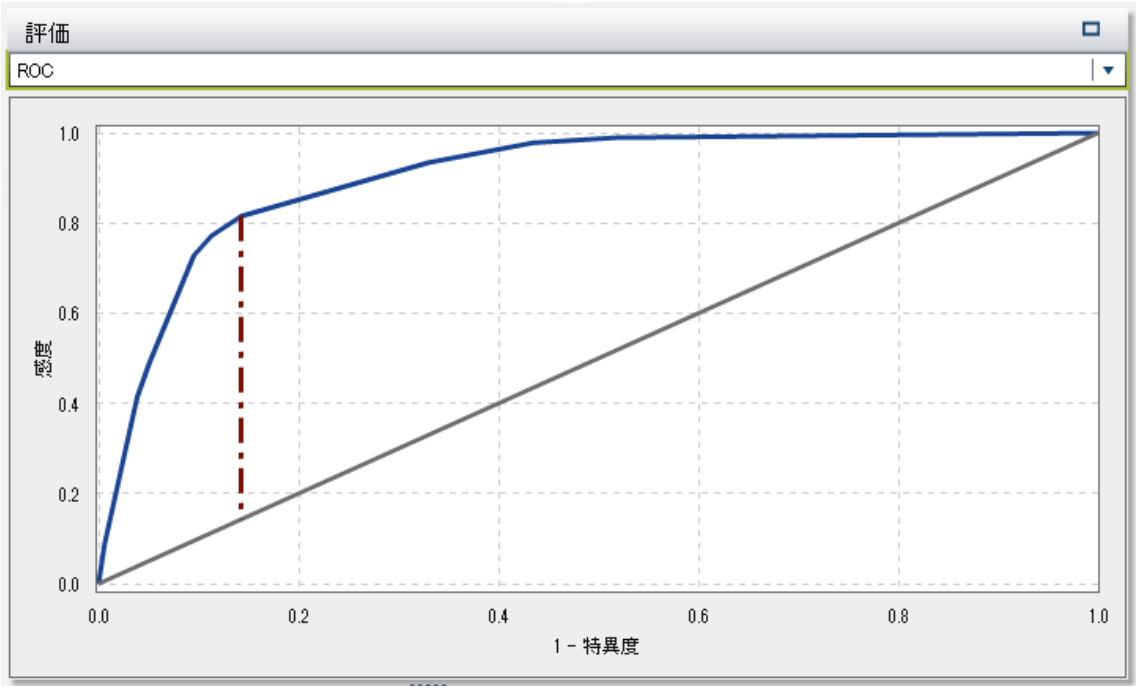
リフトチャートが下がり始める位置は、モデルの予測値が減少する位置を示しています。決定木では、この位置で、次のリーフまたはリーフグループが著しく弱くなっています。

比較においては、リフトチャートは、入力データに関する完全な知識に基づく最良のモデルがプロットされます。

ROC

受信者操作特性(ROC)チャートは、偽陽性および偽陰性の分類を回避するモデルの能力を示します。偽陽性の分類とは、あるオブザベーションで、実際にはイベントがない(疾患がない: 陰性)ときに、イベントがある(疾患がある: 陽性)と識別されることをいいます(第一種(Type 1)過誤とも呼ばれます)。偽陰性の分類とは、あるオブザベーションで、実際にはイベントがある(疾患がある: 陽性)ときに、イベントがない(疾患がない: 陰性)と識別されることをいいます(第二種(Type II)過誤とも呼ばれます)。

このモデルの *特異度*は、真の陰性率です。偽陽性率を導出するには、1 から特異度を減算します。**1 – Specificity** というラベルが付けられた偽陽性率は、ROC チャートの X 軸です。モデルの *感度*は、真の陽性率です。これは、ROC チャートの Y 軸です。したがって、ROC チャートでは、偽陽性率の変化に伴う真の陽性率の変化がプロットされます。

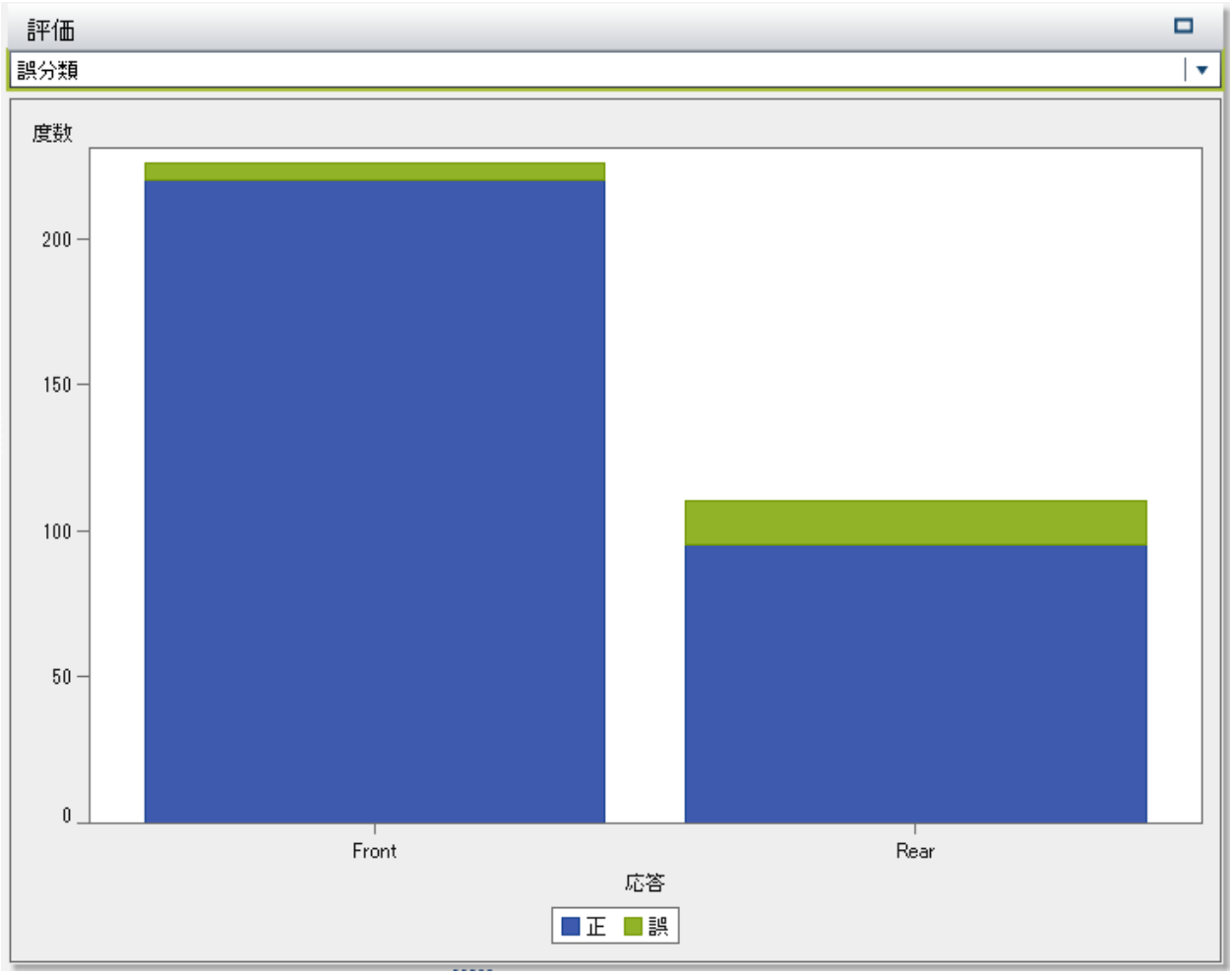


良い ROC チャートは、最初に非常に急な勾配があり、すぐに横ばいになります。すなわち、オブザベーションの誤分類より、かなり多い数のオブザベーションが正しく分類されていることがわかります。偽陽性も偽陰性もない完璧なモデルの場合、ROC チャートは(0,0)で開始し、(0,1)に垂直に推移してから、(1,1)で水平になります。この例では、1 つの誤分類が発生するまでは、モデルはすべてのオブザベーションを正しく分類しています。

ROC チャートには、ROC チャートの解釈に役立つ 2 つの線が含まれています。最初の線は、1 の勾配を持つベースラインモデルです。この線は、オブザベーションを誤分類するのと同じ比率で正しく分類するモデルを模倣しています。理想的な ROC チャートは、ベースラインモデルと ROC チャート間の距離を最大化します。オブザベーションを正しく分類するよりも多い比率で誤分類するモデルは、ベースラインモデルの基準に達していません。2 番目の線は、偽陽性率の垂直線です。この線では、ROC チャートとベースラインモデルの Kolmogorov-Smirnov 値間の差異が最大になります。

誤分類

誤分類プロットには、正しく分類されたオブザベーションと誤分類されたオブザベーションの数が示されています。

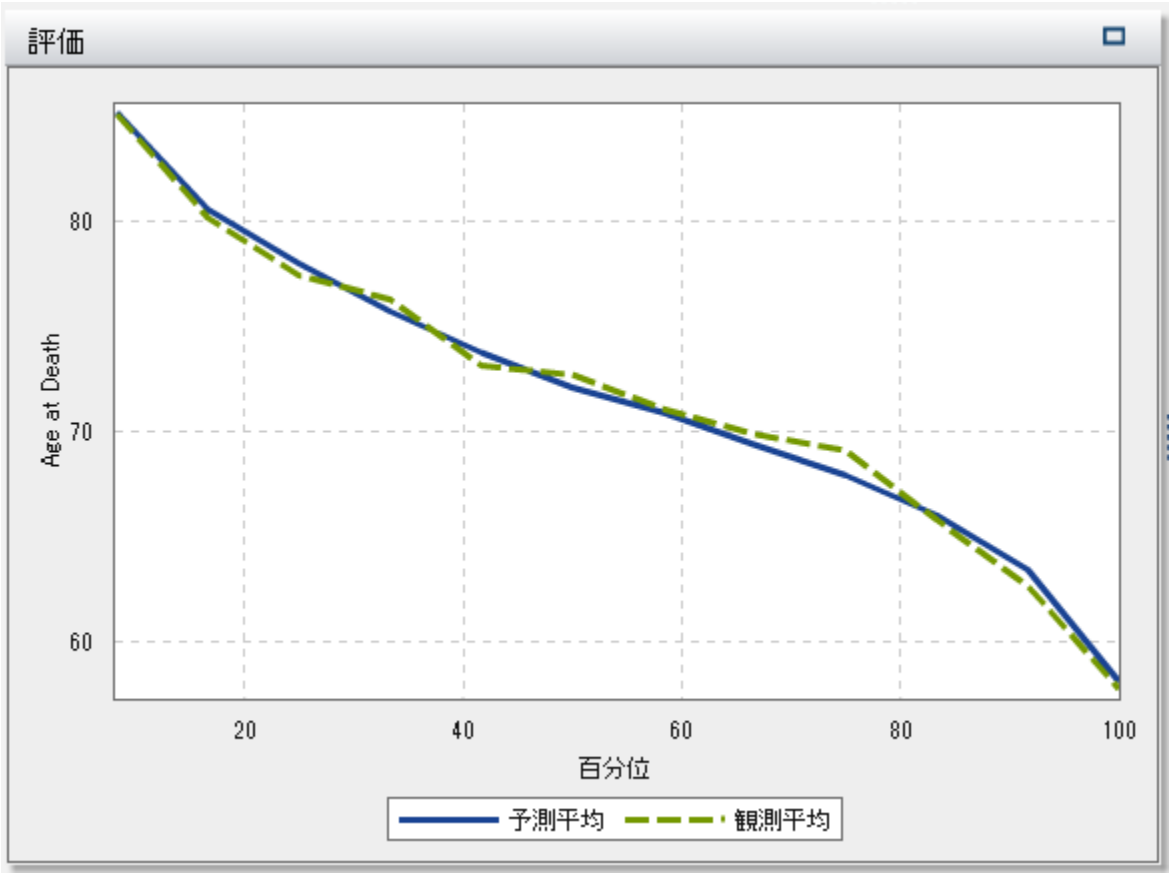


誤分類が著しく多い場合は、モデルがデータに当てはまっていないことを示していることがあります。

データ内でイベントがあるに対するイベントがないの比率が比較的大きい場合、誤分類プロットに多数の真の陽性率と偽の陽性率が示されていることがあります。この場合は、モデルでほとんどのオブザベーションがイベントがあると予測され、間違って分類されている場合よりも正しく分類されている場合の方が多くなります。

予測平均値対観測平均値

Response bins の数が 10 より大きい数に設定されている場合は、**評価**ウィンドウに予測平均値と観測平均値がプロットされます。



要約テーブル

モデルペインの上部にある**要約テーブルの表示**をクリックすると、モデルペインの下部に要約パネルが表示されます。要約テーブルには、次の情報が含まれています。

ノード統計

決定木の各ノードの要約統計量を提供します。利用可能な統計量には、**奥行**、**親 ID**、**子の数**、**種類**、**オブザベーション**、**% オブザベーション**、**欠損数**、**ゲイン**、**予測値**、**分割**、各ビン内のオブザベーションの数および割合があります。

ノードルール

決定木のノードごとに使用する並べ替えルールを提供します。利用可能な変数はすべてテーブルに列としてリストされています。任意のルールがノードの変数またはその親ノードのいずれかに適用された場合は、このテーブルにリストされます。リストされない場合、そのエントリはブランクになります。

8

クラスタ

クラスタツールの概要

クラスタリングとは、データをセグメント化する手法で、オブザベーションをデータによって示唆されるグループに分類します。各クラスタのオブザベーションは、測定可能な形で類似している傾向にあり、異なるクラスタのオブザベーションは、似ていない傾向があります。各オブザベーションは、最大で 1 つのクラスタに割り当てられます。クラスタリング分析では、他のツールで使用するクラスタ ID 変数を生成できます。

クラスタツールには、少なくとも 2 つの尺度変数が入力として必要です。交互作用項またはカテゴリ変数は指定できません。

クラスタのプロパティ

クラスタツールでは、次のプロパティを使用できます。

名前
このモデルの名前を指定します。

クラスターマトリックス

- **Number of clusters** では、生成するクラスタの数を指定します。
- **Seed** では、最初のクラスタ割り当て時に使用される乱数ジェネレータのシード値を指定します。
- **Initial assignment** では、クラスタの初期割り当ての作成に使用する方法を指定します。次の方法を利用できます。
  - **Forgy** では、k データ点をランダムに選択し、k クラスタの重心として使用するように指定します。
  - **Random** では、クラスタにオブザベーションをランダムに割り当てます。
- **Visible roles** では、**クラスターマトリックス**に表示する効果の数を指定します。有効な値は、2～6 の範囲の整数(2 と 6 を含む)です。

  値 *n* を指定する場合、**役割**タブの **Variables** テーブルにリストされている最初の *n* 個の効果が表示されます。**クラスターマトリックス**にプロットされる効果のペアを変更するには、分析から効果を削除してから、すぐに変更を組み入れます。同じ入力データを使用しているため、クラスタリングの結果には変更はありません。ただし、新しい効果は、**Variables** テーブルのリストの下部に追加されます。
- **Variable standardization** では、平均値 0 と標準偏差 1 をもつように効果変数を変換します。このプロパティは、デフォルトで有効になっており、要約テーブルに表示される結果に影響を及ぼします。**クラスターマトリックス**ウィンドウおよび **Parallel Coordinates** ウィンドウには、最初の変数が表示されます。

Parallel Coordinates

- **ビン数**では、並列座標ポリラインプロットを生成する際に使用するビン数を指定します。
- **Maximum polylines** では、並列座標アルゴリズムによって生成されるポリラインの最大数を指定します。

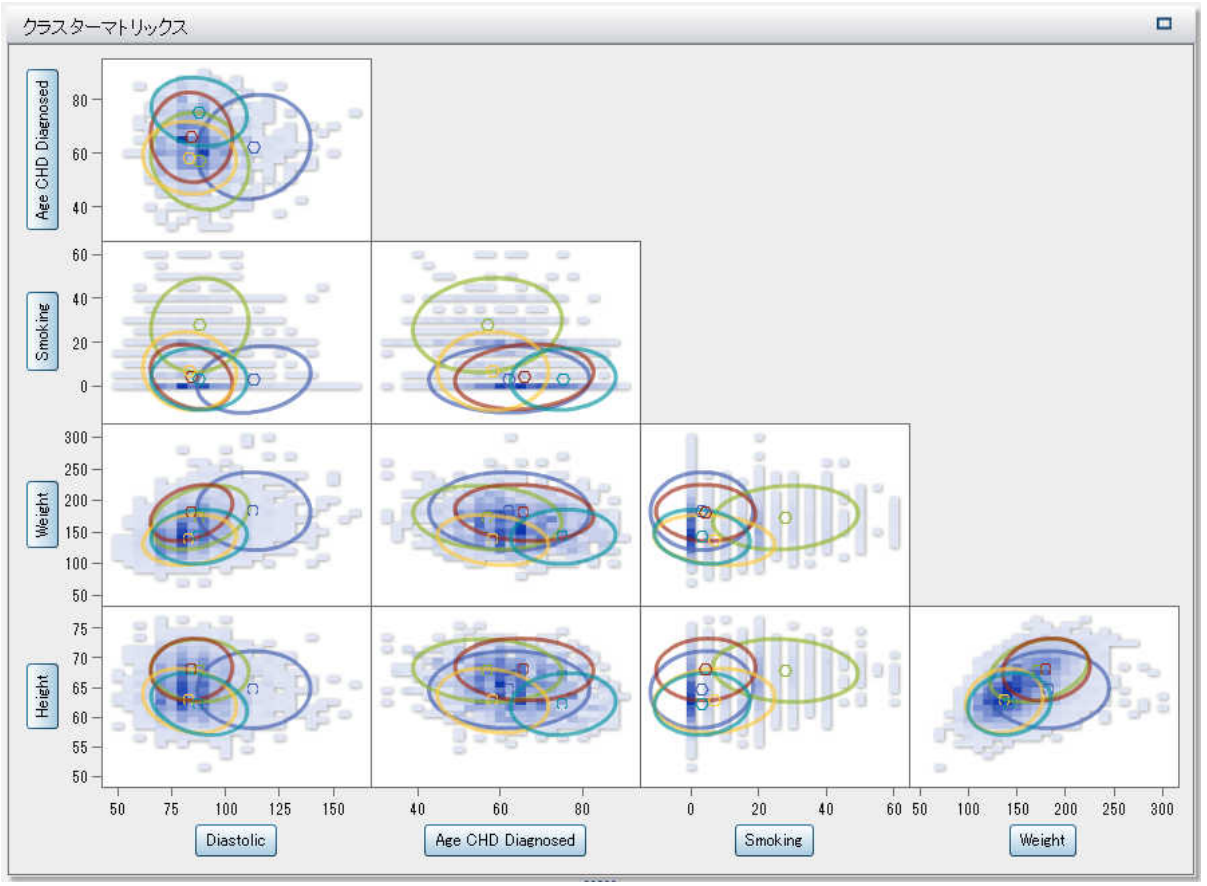
Show ellipses
クラスタの投影楕円を**クラスターマトリックス**に表示します。

Show centroids
クラスターマトリックスに重心を表示します。

クラスタ結果ウィンドウ

クラスターマトリックス

クラスターマトリックスには、指定された数の効果のペア上に各クラスタの 2 次元の投影が表示されます。これらの投影は、プロットされた効果ペア内でクラスタの類似性や相違点を特定する上で有効です。



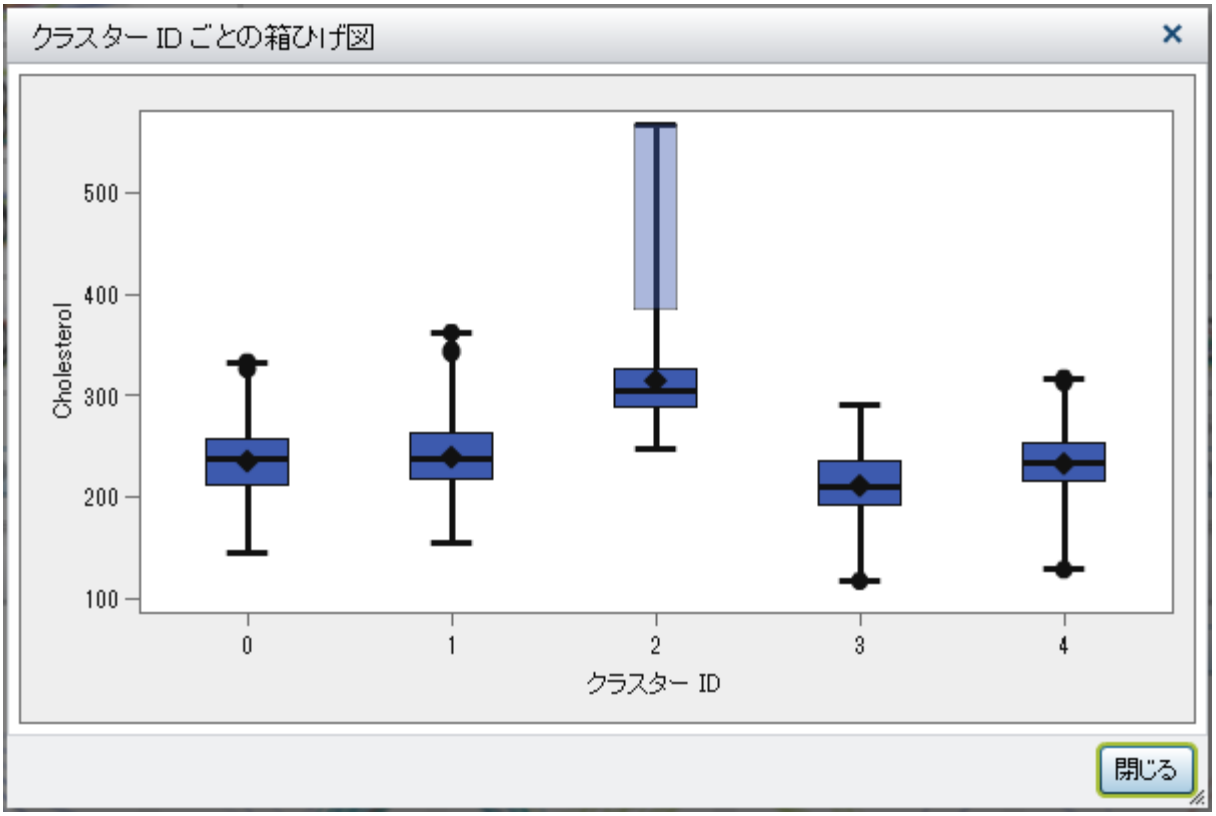
各クラスタには、一意の色が割り当てられています。各クラスタは、*n* 空間では一意であり、2 次元の投影は重なります。ヒートマップが使用されない場合は、個々のオブザベーションは所属クラスタを示すために色分けされます。

任意の効果ペアのプロットをより大きく表示するには、そのプロット内を右クリックし、**Explore** をクリックします。**Explore** ウィンドウでは、オブザベーションの表示や選択を簡単に行えます。オブザベーションを選択すると、選択したオブザベーションに重なっているクラスタも選択されます。

注意すべきことは、各オブザベーションは 1 つのクラスタだけに属することができるということです。ただし、**クラスターマトリックス**では、投影を 2 次元でしか表示できないため、複数のクラスタが 1 つのオブザベーションに重なることがあります。

クラスターマトリックスプロットを右クリックして、ポップアップメニューを開きます。このメニューの最後の項目は、**Derive a Cluster ID Variable** です。この項目を選択すると、SAS Visual Statistics によって各オブザベーションのクラスタ ID を含むカテゴリ変数が作成されます。この変数は、他のモデルで効果として使用できます。

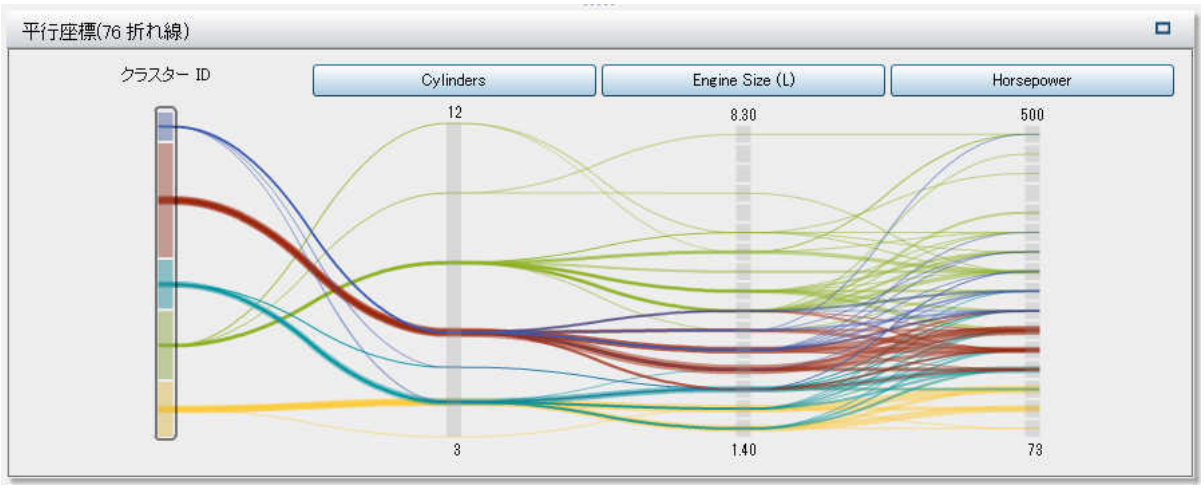
オブザベーションをクラスタによって区分化する変数の箱ひげ図を表示できます。関心の対象である変数を含むプロット内を右クリックし、**Plot *variable_name* by Cluster ID** を選択します。選択したプロットの各変数には、メニュー項目があります。



箱ひげ図は、任意の変数に対してクラスタの類似性がどの程度あるかを判断するために使用します。

Parallel Coordinates

Parallel Coordinates プロットは、データとクラスタのパターンを示します。このプロットでは、クラスタ ID が 1 番左側に記載され、各変数は、ビンに分割された値の範囲を垂直に表示した 1 つの列です。色分けされたポリラインが各クラスタから描かれ、変数ごとにクラスタに含まれる値の範囲を示します。



最初はわかりにくいですが、**Parallel Coordinates** プロットは、データに関するさまざまな推論に使用できます。このプロットを調整して、所属クラスタまたは 1 つ以上の変数の特定の範囲あるいはその両方に基づいて、データを検証できます。

複数のクラスタがある場合は、各クラスタのデータ分類がどのようになっているか判断しにくいことがあります。単一のクラスタのみのポリラインを表示するには、一番左側でクラスタ ID を選択します。他のすべてのクラスタのポリラインがグレイ表示されます。これで、1 つのクラスタに焦点を合わせて検証できます。**Ctrl** キーを押しながら、複数のクラスタをクリックすれば、これらのクラスタのみを表示できます。

変数を選択するには、その変数名の上または近くをクリックします。この操作を行うと、そのポリラインの色のグラデーションが変化します。大きな値は、小さな値よりも色が濃くなります。表示されている値の範囲を調整するには、変数の上の範囲または下の範囲をクリックしてドラッグします。この手順を複数の変数について繰り返します。

これらの 2 つの機能を組み合わせることで、関心のある特定のクラスタおよび変数の範囲の表示を制限できます。

要約テーブル

モデルペインの上部にある**要約テーブルの表示**をクリックすると、モデルペインの下部に要約パネルが表示されます。要約テーブルには、次の情報が含まれています。

- **Cluster Summary** には、クラスタごとの要約統計量が示されています。利用可能な統計量には、**オブザベーション**、**RMS of STD**、**Within-Cluster SS**、**Min centroid-to-observation**、**Max centroid-to-observation**、**Nearest Cluster**、**Centroid Distance** があります。

9

モデルの比較

モデルの比較の概要

モデル比較ツールを使用すると、さまざまなベンチマーク基準を使って、競合モデルのパフォーマンスを比較できます。利用可能な比較基準は、分析で使用するモデルと応答変数によって異なります。モデルの比較を行うには、比較する前に少なくとも 1 つの他のモデルの学習をしておく必要があります。

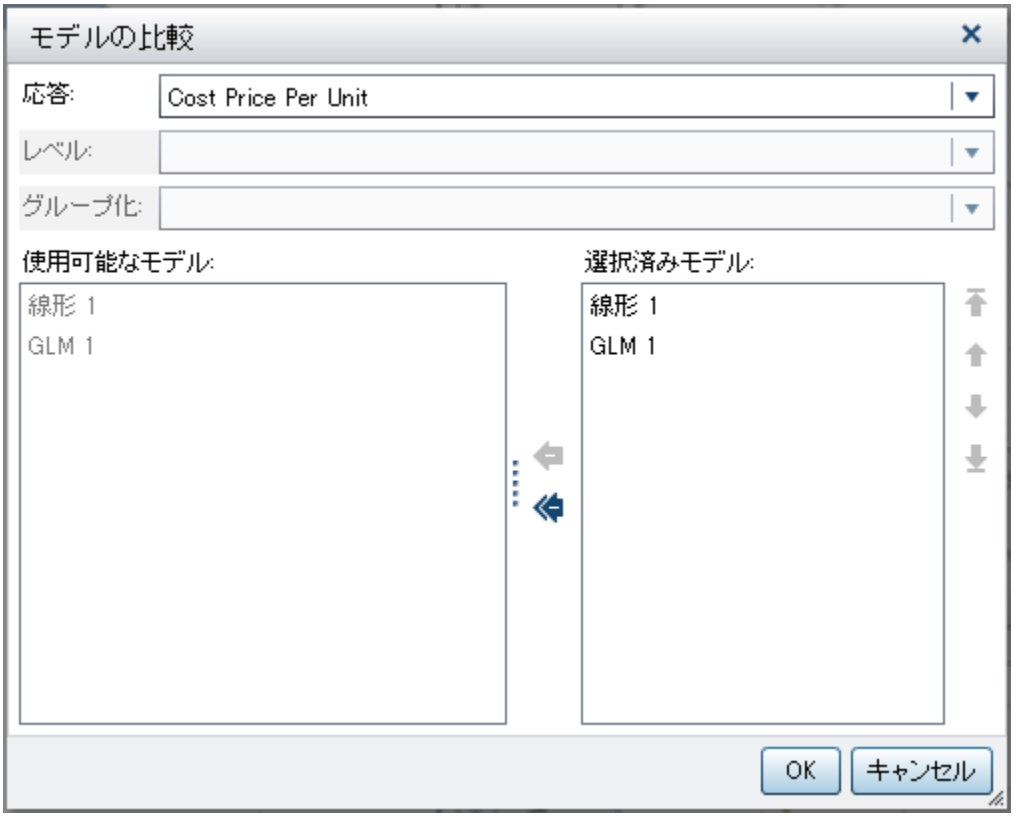
注: モデルの比較は、SAS Visual Statistics で保存されません。プロジェクトを閉じてからまた開いて任意のモデルの比較を参照する場合は、その比較を再作成する必要があります。

モデルの比較を実行する前に、すべてのモデルを初期化および更新してください。任意のモデルで **Auto-update model** プロパティが無効になっている場合は、別のモデルと比較する前にそのモデルを手動で更新する必要があります。モデルは、学習が完了するまで初期化されたとはみなされません。

比較の作成後にモデルを変更した場合、そのモデルの比較に変更は反映されません。

モデルの比較の使用方法

ツールバーでアイコンをクリックすると、**モデルの比較**ウィンドウが表示されます。



モデルの比較ウィンドウでは、関心の対象となる応答変数、目的の階層、Group BY 変数、比較対象となるモデルを指定できます。少なくとも 1 つの応答変数と少なくとも 2 つのモデルを指定できます。

注: 応答変数、目的の階層、Group BY 変数が同一である場合のみ、2 つ以上のモデルを比較できます。

モデルの比較のプロパティ

モデルの比較では、次のプロパティを使用できます。

名前
この比較の名前を指定します。

当てはめ統計量
当てはめ統計量ウィンドウにプロットし、チャンピオンモデルの決定に使用する比較基準を指定します。利用可能な当てはめ統計量は、比較するモデルによって異なります。

誤差平方和(SSE)当てはめ統計量の場合、線形回帰分析モデルとロジスティック回帰分析モデルでは、重み付き SSE を使用します。一般化線形モデルでは、重み付けなしの SSE を使用します。

Prediction Cutoff
任意のオブザベーションがモデル化されているイベントであるかどうかを判定するために使用するカットオフ確率を指定します。

Percentile
可能な場合に、指定されている当てはめ統計量をプロットする位置のパーセンタイル値を指定します。

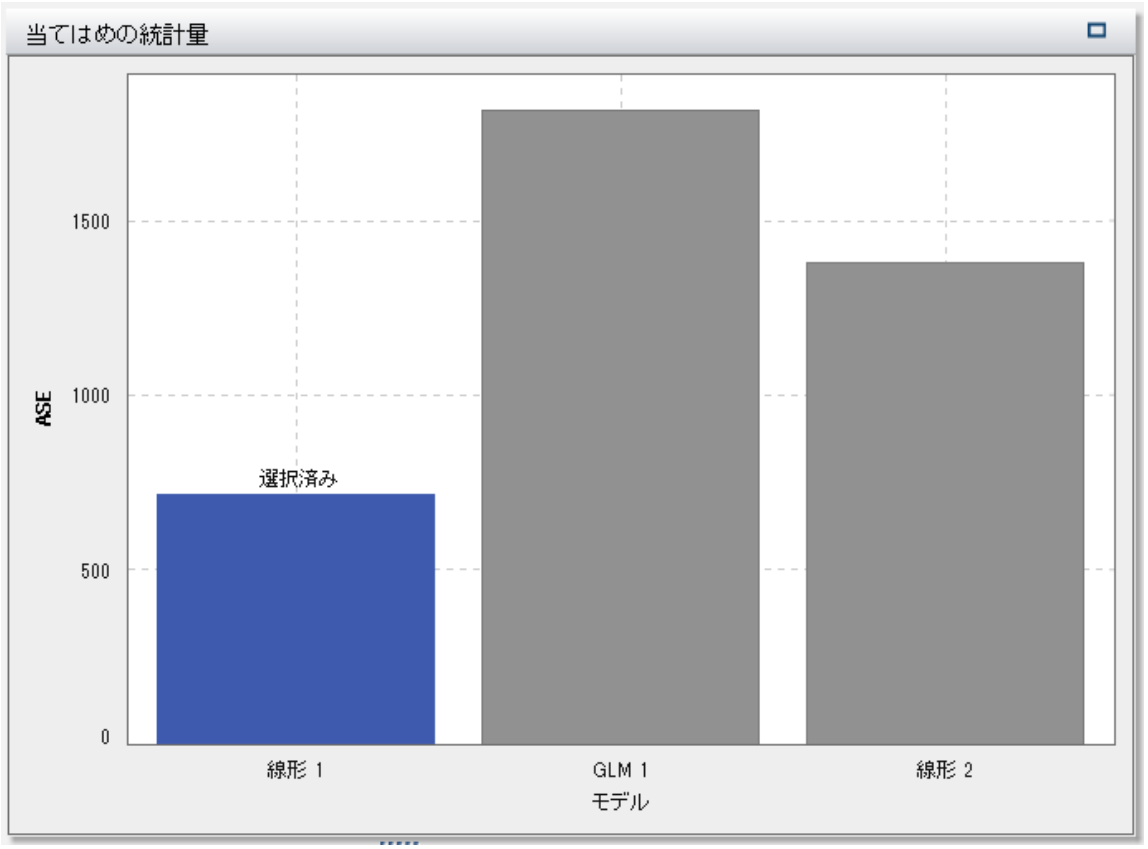
モデル比較の結果ウィンドウ

評価

利用可能な評価プロットは、比較するモデルによって異なります。分類モデルの場合、表示されるプロットは、リフト、ROC、誤分類です。数値モデルの場合、表示されるプロットは、応答観測値および応答予測値です。

当てはめ統計量

当てはめ統計量プロットには、**当てはめ統計量**プロパティで指定した基準が表示されます。次の画像では、線形回帰分析モデルと GLM モデルの平均観測値がプロットされています。プロットには、チャンピオンモデルが示されています。チャンピオンモデルは、他のモデルとは異なる表示で示されます。



要約テーブル

モデルペインの上部にある**要約テーブルの表示**をクリックすると、モデルペインの下部に要約パネルが表示されます。要約テーブルには、次の情報が含まれています。

統計量

比較する各モデルの要約統計量を示します。**選択済み**列の値、**はい**または**いいえ**は、どのモデルが**当てはめ統計量**プロパティで指定されている基準に基づいてモデル比較ツールによって選定された望ましいモデルに該当するかを示します。ただし、要約テーブルにリストされている統計量は、**Fit statistic** プロパティにリストされているものとは異なることがあります。

Variable Importance

どの変数が比較されている各モデルに対して最大の影響を及ぼすかを示します。

10

SAS Visual Statistics の使用例

概要

この例では、SASHELP.HEART にあるフレーミンハム心疾患研究データセットを使用して、線形回帰分析モデルと一般化線形モデル(GLM)を比較します。その目的は、健康因子の集合に基づいて、人の死亡年齢を予測することにあります。これらの因子には、性別、体重、身長、喫煙の有無、血圧などがあります。この例は、最適なモデルの構築方法ではなく、SAS Visual Statistics の使い方の説明に重点を置いています。

また、この例では、ユーザーが SASHELP.HEART データセットにアクセスできることを前提として説明します。各所在場所からの個々のデータのアクセス方法についての説明は、この例では省略しています。このデータセットへのアクセスについては、システム管理者にお問い合わせください。

プロジェクトの作成

この例では、すでに SAS Visual Analytics にサインオンし、ホームページが表示されていることを前提として説明します。ホームページで、**コンテンツの作成**グループの**分析モデルの作成**アイコンをクリックします。これにより、SAS Visual Statistics が表示されます。ここでは、最近使用したプロジェクトを開くか、新規プロジェクトを作成するかを選択できます。新規プロジェクトを作成するには、**新規モデルの開始**の下にある**データソースの選択**をクリックします。

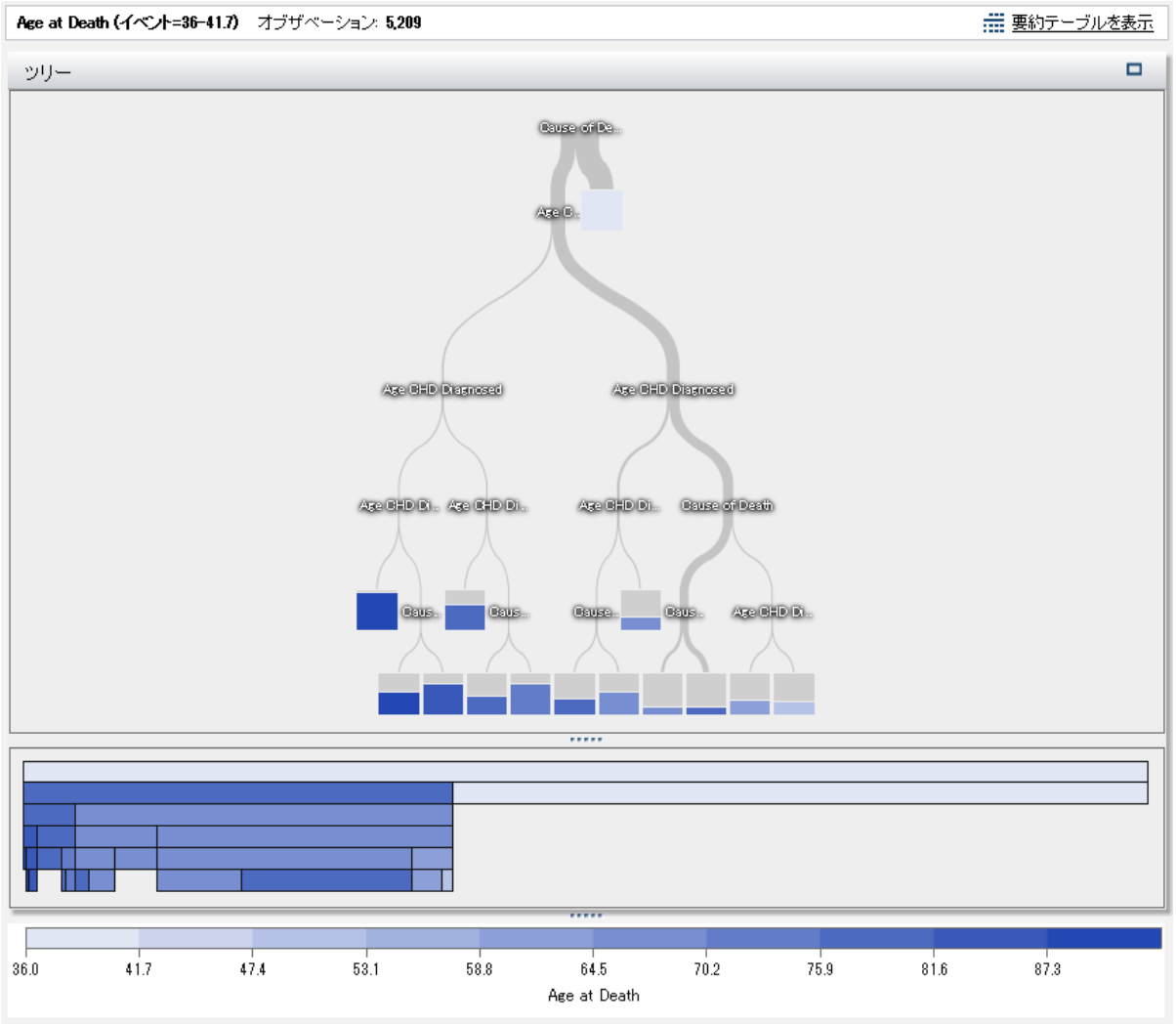
このプロジェクトのデータソースを選択できるウィンドウが表示されます。SASHELP.HEART に対応するデータソースを選択します。**開く**をクリックします。

デフォルトで、このプロジェクトは **Project 1** という名前が付けられ、SAS Visual Statistics の左上の隅に表示されます。この例の作成を続行する前に、このプロジェクトの名前を変更して保存します。メインメニューから、**ファイル** ▶ **保存**の順に選択します。**名前を付けて保存**ウィンドウが開きます。**SAS Folders** ペインで、書き込み権限のある場所に移動します。通常は、**マイフォルダ**に作業を保存できます。**名前**フィールドで、`Heart Study` と入力し、**保存**をクリックします。

デフォルトでは、線形回帰分析モデルは即時使用可能となっています。このデフォルトのモデルの種類は、**プリファレンス**ウィンドウで変更できます。ただし、この例では、リーフ ID 変数を導出するために決定木を作成します。このリーフ ID 変数は、後に線形回帰分析や GLM で使用します。

決定木の作成

決定木を作成するには、ツールバーで、アイコンをクリックします。**データ**ペインで、**Age at Death** 変数を右ペインの**応答**フィールドにドラッグアンドドロップします。**データ**ペインで、**Diastolic**、**Weight**、**Height**、**Cholesterol**、**Age CHD Diagnosed**、**Sex**、**Cause of Death** を選択します。これらの項目をモデルペインにドラッグアンドドロップします。決定木が自動的に更新されます。



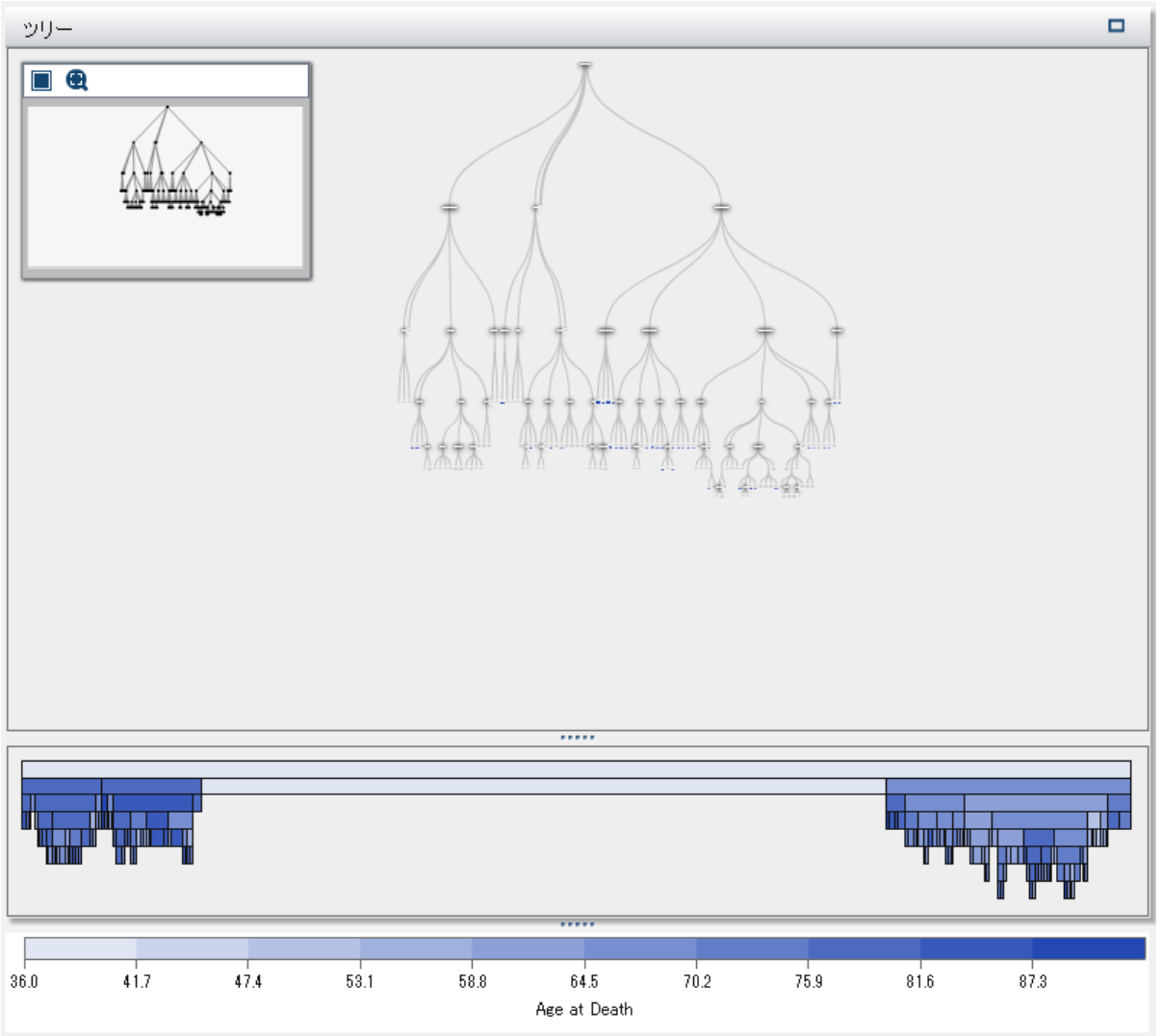
要約バーで、**要約テーブルの表示**をクリックします。要約テーブルで、**ノードルール**タブを選択します。使用されている予測子は、**Age CHD Diagnosed** および **Cause of Death** のみであることがわかります。決定木のプロパティを調整して、予測子を追加できます。

右ペインで、**プロパティ**タブをクリックします。最も顕著なプロパティの変更は、**予測因子の再利用**です。このプロパティの選択を解除すると、各予測子変数は、1 つの分岐でのみ使用されます。この例では、予測子を再利用することにより各ノードで最良の分岐が作成されることを前提としています。これは、どのデータについても必ず言えるということではありません。

かわりに、**最大レベル**の値を 10 に設定します。これで決定木は、デフォルトの 6 階層ではなく、最大の深さである 10 階層になります。要約テーブルの**ノードルール**タブでは、すべての予測子が少なくとも一度使用されます。

最大枝数の値を 4 に設定します。これにより、リーフのないノードは、最大で 4 つの新しいノードに分岐できます。

ツリーの概要を表示プロパティを選択します。**ツリーの概要**ウィンドウで、アイコンをクリックし、決定木全体を**ツリーの概要**ウィンドウに収めます。各ノードは見えにくくなりますが、決定木を次のように表示できます。



Tree ウィンドウで、右クリックし、**リーフ ID 変数を派生**を選択します。**Leaf ID 1** という名前の新しいカテゴリ変数が**データ**ペインに表示されます。

プロジェクトを保存します。

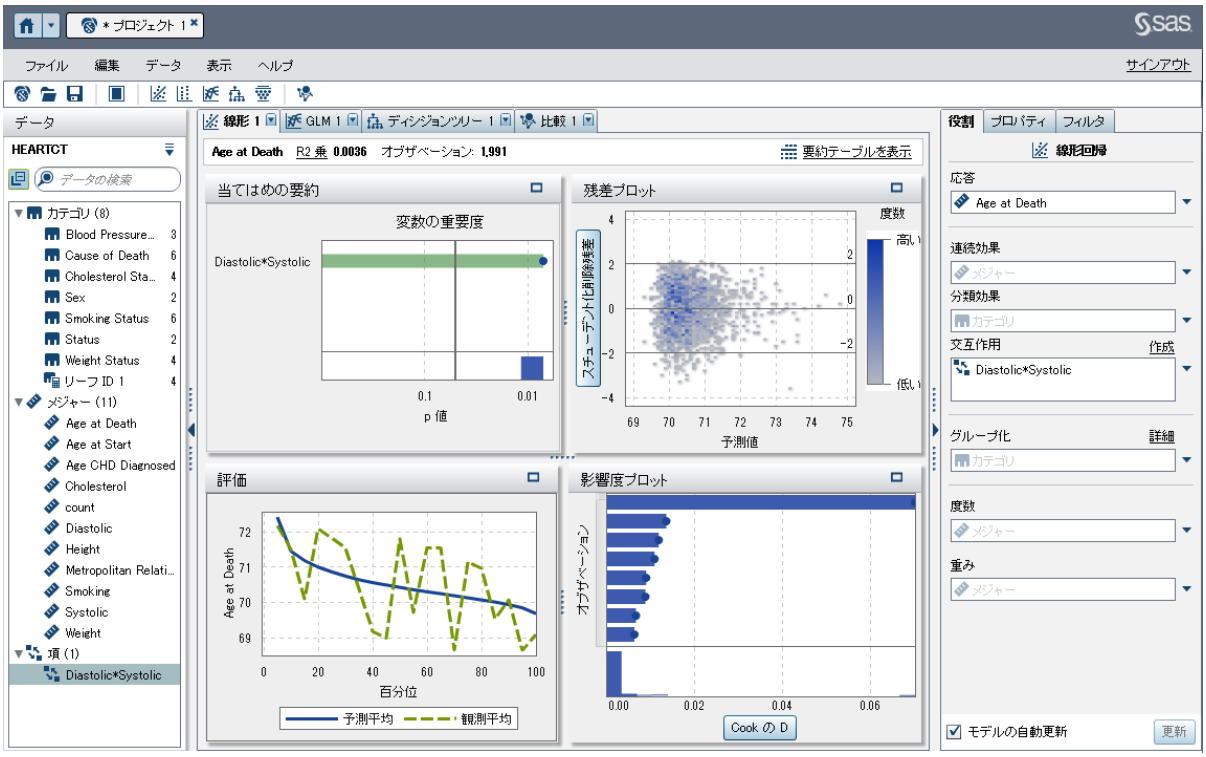
線形回帰分析の作成

モデルペインで、プロジェクトの作成時に作成した線形回帰分析モデルを選択します。

この例では、関心の対象となる変数は、**Age at Death** であり、**データ**ペインの**メジャー**セクションにリストされる最初の変数となります。この変数を応答変数にする必要があるため、**Age at Death** をクリックし、**データ**ペインからモデルペインにドラッグアンドドロップします。これで、**Age at Death** が、**役割**タブの**応答**フィールドに表示されます。

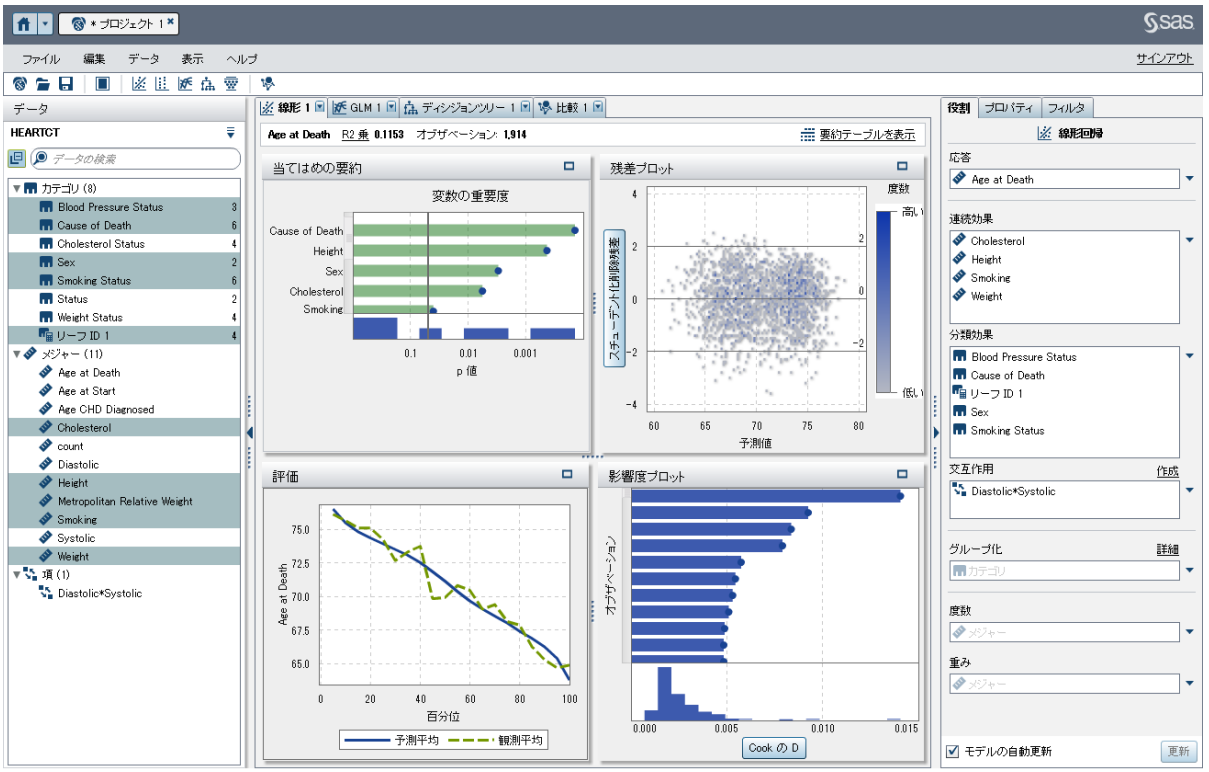
次のステップでは、分析に含める効果変数または交互作用項を選択します。利用可能なすべての変数を効果変数にするための 1 つの選択肢は、SAS Visual Statistics で自動的に変数を選択することです。しかし、この方法は、計算資源の観点からいつも実行可能であるとは限りません。この例では、交互作用項を作成して効果変数として使用し、2、3 の他の変数を効果変数として追加します。

最高血圧と最低血圧は、相互に作用する変数であると考えられるため、これらの変数に対して交互作用項を作成します。**データ**ペインで、**Diastolic** を選択します。Ctrl キーを押しながら、**Systolic** をクリックします。両方の変数が選択されます。**Systolic** を右クリックして、**1 つの交互作用の作成**を選択します。交互作用項 **Diastolic*Systolic** が、**データ**ペインの**項**グループに表示されます。



Diastolic*Systolic をクリックして、モデルペインにドラッグアンドドロップします。その単一の効果に基づいて、モデルが作成されます。これは、右ペインの **Auto-update model** オプションが選択されているためです。モデルに対して変更が行われるたび、この線形回帰分析は自動的に更新されます。多数の変更を行うことが予想される場合や、サーバーのパフォーマンス上の問題などが発生している場合は、**Auto-update model** オプションの選択を解除してください。自動更新を無効にした場合、モデルを更新するには、右ペインの**更新**をクリックする必要があります。

次に、いくつかの効果をモデルに追加します。Ctrl キーを押しながら、**Blood Pressure Status**、**Cause of Death**、**Leaf ID 1**、**Sex**、**Smoking Status**、**Cholesterol**、**Height**、**Smoking**、**Weight** を選択します。これらの変数をモデルペインにドラッグアンドドロップします。線形回帰分析によってこれらの効果が追加されて更新されます。



右ペインで、**プロパティ**タブを選択します。このモデルでは、**有用な欠損**および**変数選択の使用**は、選択しません。**有用な欠損**を無効にするということは、欠損値を含むオブザベーションは、分析に含めないことを意味します。**変数選択の使用**を無効にするということは、モデルに対する有意性の如何にかかわらず、すべての変数をモデルで使用することを意味します。このモデルでは、デフォルトのプロパティ設定を維持します。

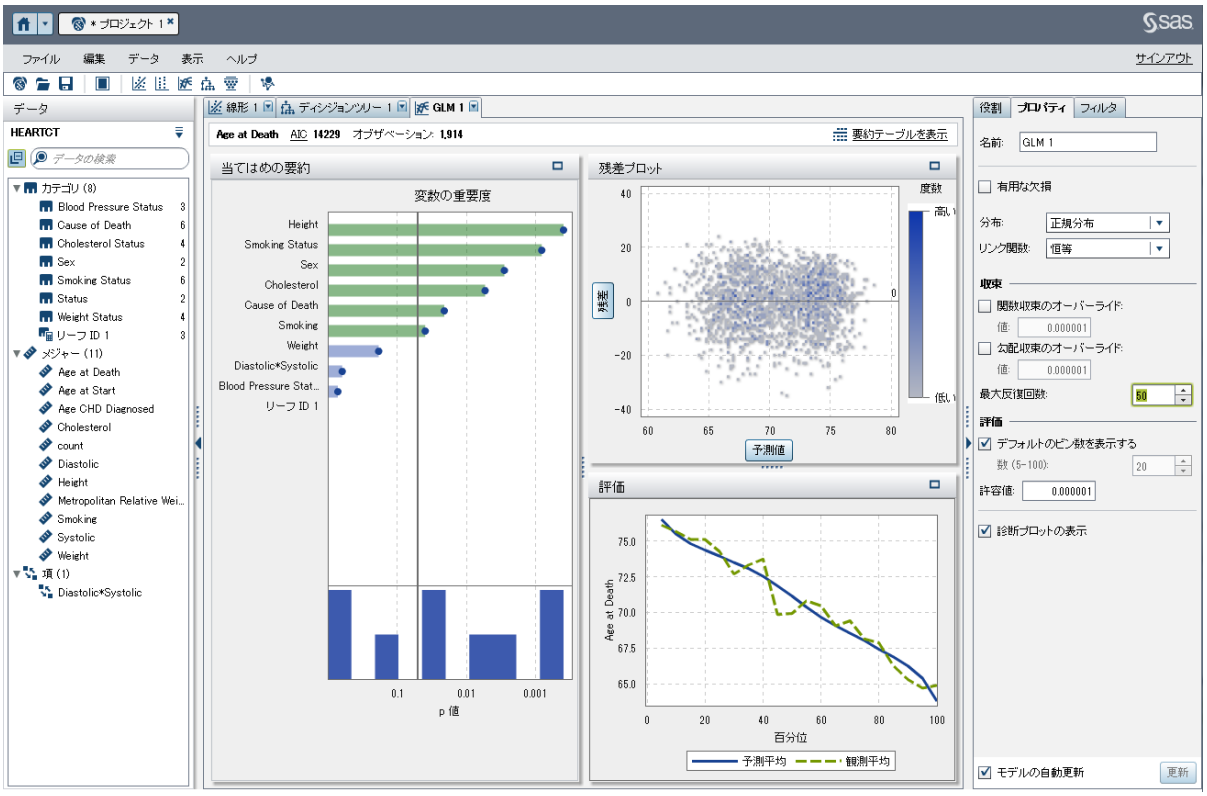
当てはめの要約ウィンドウには、**Cause of Death**、**Leaf ID 1** および **Height** の 3 つの効果がこのモデルで最も重要な効果であることが示されています。

評価ウィンドウには、平均値の観測値と予測値は、ほとんどのビンでほぼ同じであることが示されています。

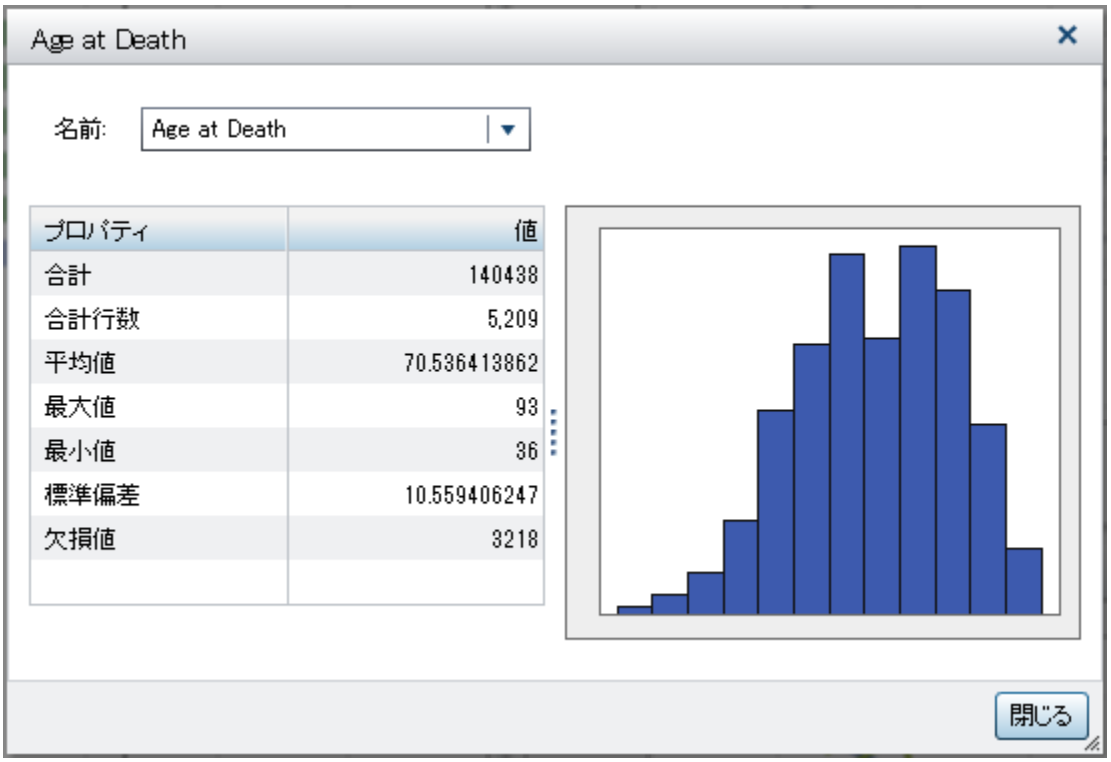
プロジェクトを保存します。

GLM の作成

GLM を新規作成するには、ツールバーで、アイコンをクリックします。**データ**ペインで、**Age at Death** 変数を右ペインの**応答**フィールドにドラッグアンドドロップします。**データ**ペインで、Ctrl キーを押しながら、**Blood Pressure Status**、**Cause of Death**、**Leaf ID 1**、**Sex**、**Smoking Status**、**Cholesterol**、**Height**、**Smoking**、**Weight**、**Diastolic*Systolic** を選択します。これらの変数をモデルペインにドラッグアンドドロップします。



右ペインで、**プロパティ**タブをクリックします。**Distribution** プロパティを使用すると、応答変数の分布を指定してその分布に基づいてモデルを構築できます。デフォルトの分布は、**正規分布**です。応答変数に正規分布を適用するかどうかを決定するには、**データ**ペインで、**Age at Death** を右クリックして、**プロパティ**を選択します。

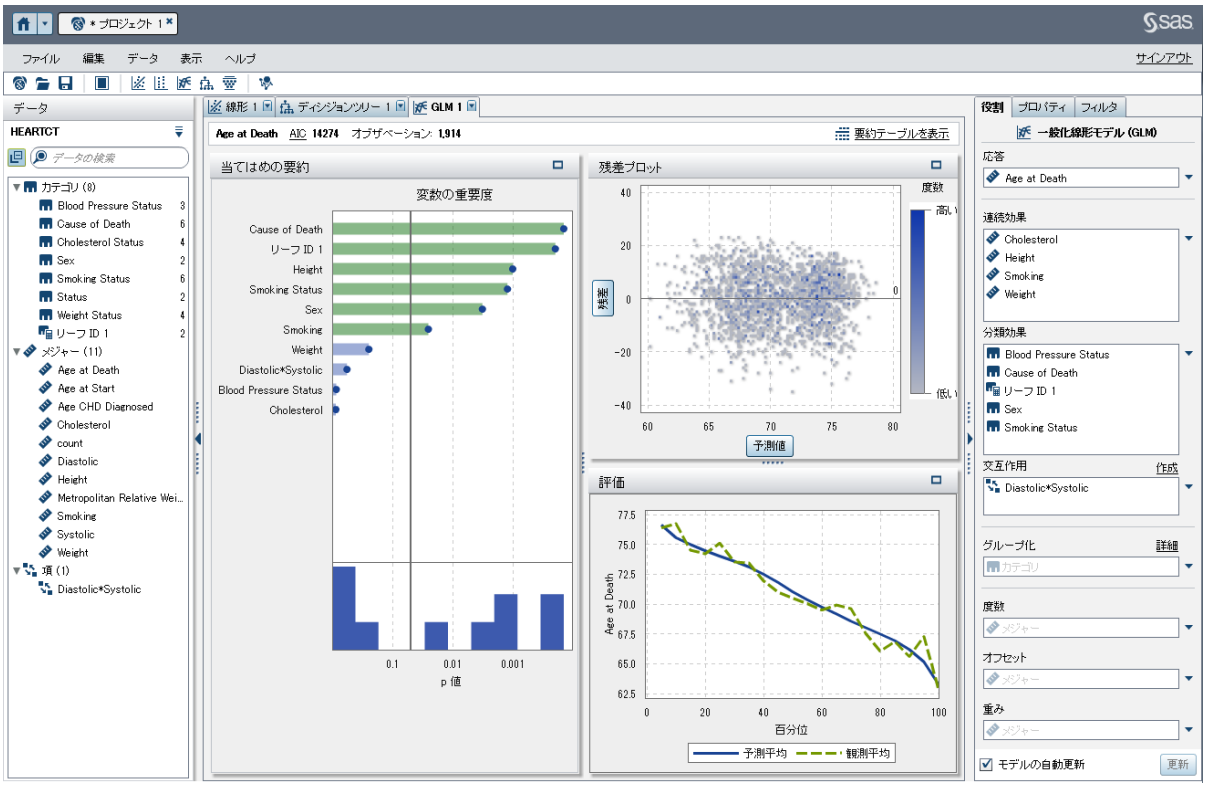


Age at Death は、正規分布に従っていません。その分布は、負の二項分布ではありませんが、この例では負の二項分布を使用します。**分布**で、**負の二項分布**を選択します。次に、**リンク関数**で**恒等**を選択します。

注: さまざまな分布やリンク関数を使用してこの例を繰り返し実行してそのパフォーマンスを比較し、SAS Virtual Statistics の操作に慣れることをお勧めします。

収束状態ウィンドウが表示されます。最大反復回数に到達したにもかかわらず、モデルが収束しない場合には、このウィンドウにより通知されます。この問題を解決するには、**最大反復回数**プロパティの値を増加する必要があります。**収束状態**ウィンドウで、**閉じる**をクリックします。

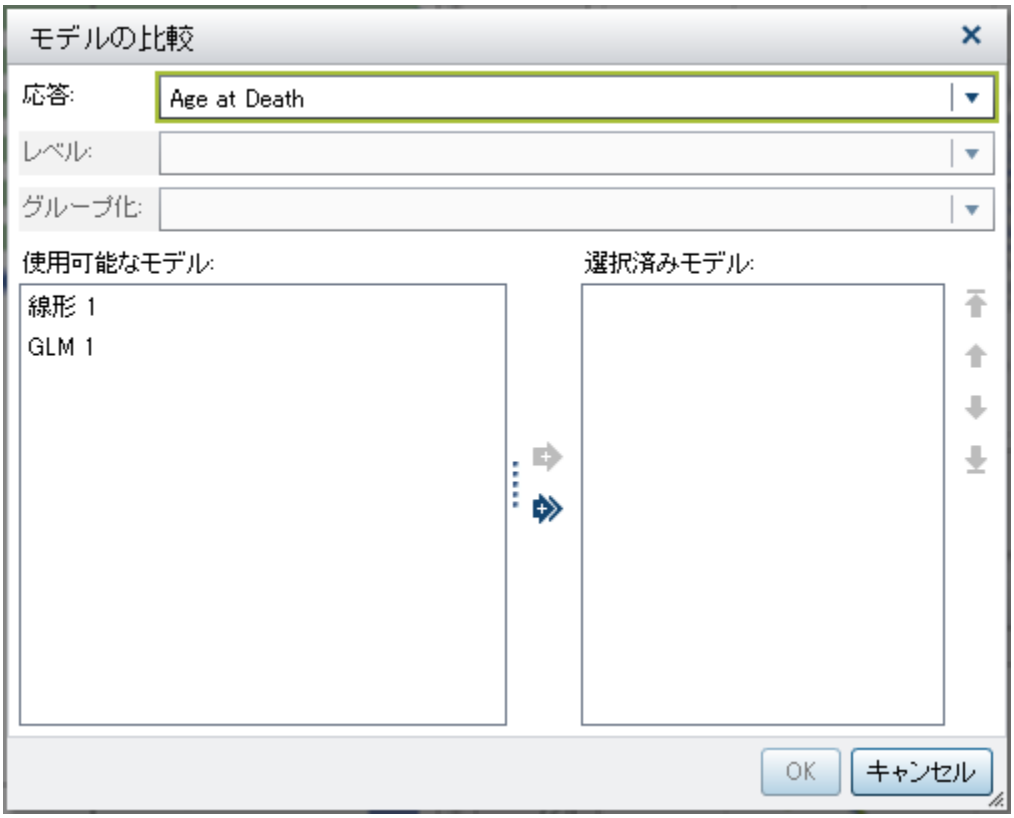
最大反復回数の値を最大値である `100` に設定します。モデルは依然として収束しません。このような場合には、関数の収束条件を確実に収束するように調整します。**関数収束のオーバーライド**を選択します。**Value** を `0.00001` に設定します。これで、モデルは収束します。



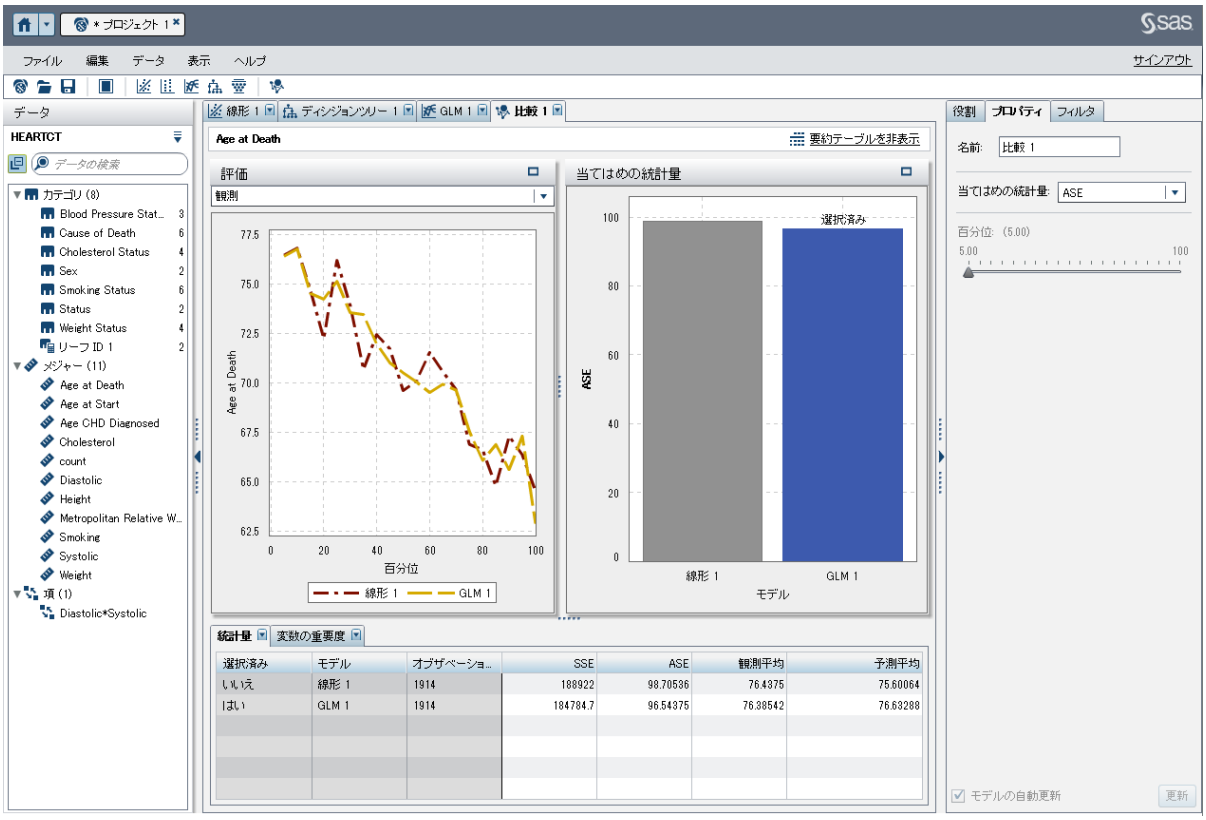
プロジェクトを保存します。

モデルの比較の実行

モデルの比較を新規作成するには、ツールバーで、アイコンをクリックします。**モデルの比較**ウィンドウが表示されます。



応答変数は、すでに **Age at Death** に設定され、**Level** は、**(none)**に設定されています。これらの設定を使用して、利用できるモデルは、**Linear 1** および **GLM 1** です。両方のモデルを選択して比較するには、をクリックします。**OK** をクリックします。



モデルの比較には、デフォルトで、当てはめ統計量の平均平方誤差 **ASE** が使用されます。その他の利用可能な当てはめ統計量は、**SSE** および**観測平均**です。**ASE** または **SSE** が条件である場合には、小さな値であるほど望ましいため、チャンピオンモデルとして **Linear 1** が選択されます。

当てはめ統計量が**観測平均**である場合には、**Percentile** スライダを利用できます。このスライダを使用して、観測平均値と予測平均値を比較する地点のパーセンタイル値を指定します。

評価プロットを表示すると、**Observed** プロットと **Predicted** プロットの両方で、モデルが比較的類似していることが示されます。

チャンピオンモデルが得られたため、そのモデルのスコアコードをエクスポートして、新しいデータをスコアリングします。メインメニューで、**ファイル** ▶ **Export** ▶ **モデルのスコアコード**の順にクリックします。**モデルスコアコードのエクスポート**ウィンドウで、**Linear 1** を選択します。これは、このモデルの ASE と SSE が望ましいからです。**OK** をクリックします。**名前を付けて保存**ウィンドウで、書き込み権限を持つファイルシステムの場所に移動します。モデルスコアコードを保存します。

この例を保存します。

3 部

管理タスク

11

インストールと設定

インストール

SAS Visual Statistics は、SAS Deployment Wizard を介してインストールされます。インストール時には、プロンプトや SAS Visual Statistics に固有のページは表示されません。SAS Visual Statistics のインストール時には、*SAS Visual Analytics: インストールと設定ガイド*およびこのガイドに記載されている関連資料を参照してください。

設定

SAS Visual Statistics の機能

特に SAS Visual Statistics でモデルを構築する必要があるユーザーについては、SAS Visual Statistics へのアクセスに使用する役割を SAS Management Console で別に作成することを強くお勧めします。その主な理由は、SAS Visual Statistics でのモデルの構築は、お使いのコンピューティング環境のパフォーマンスに対する影響が大きいためです。SAS Visual Statistics のモデル構築機能へのアクセスを制限することにより、このリスクを低減できます。

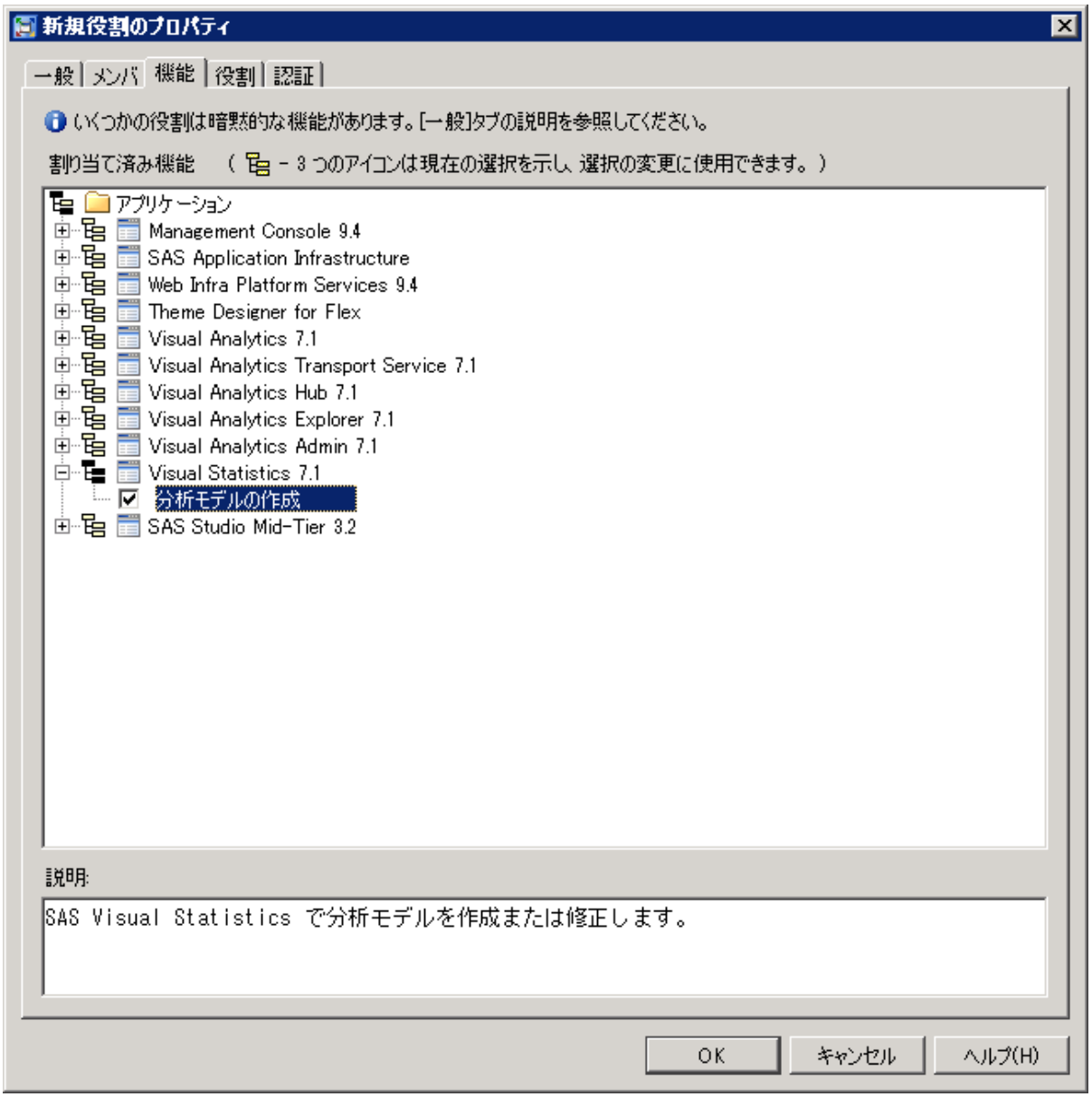
SAS Visual Statistics のモデル構築機能にアクセスを提供する新しい役割を作成するには、次の手順を実行します。

1 SAS Management Console で、**User Manager** プラグインにアクセスします。**User Manager** プラグインを使用する権限と、新しい役割を作成するためのアクセス権限を持っている必要があります。

2 **User Manager** プラグイン内で右クリックして、**新規** ▶ **Role** の順に選択します。**New Role Properties** ウィンドウが表示されます。

3 **General** タブで、**名前**フィールドに、`Visual Statistics` と入力します。**表示名**、**Description** あるいはその両方を追加できます。

4 **機能**タブで、**分析モデルの作成**機能を選択します。



5 **OK** をクリックします。

これで、SAS Visual Statistics のモデル構築機能にアクセスが必要なユーザー用に特定の役割を作成できました。この役割に、**Visual Analytics:分析**機能を追加するには、**Visual Statistics Properties** ウィンドウの **Contributing Roles** タブを使用します。

高カーディナリティデータのしきい値

高カーディナリティデータには、大量の一意の値を含む複数の列があります。たとえば、ユーザー名、電子メールアドレス、銀行口座番号などが、高カーディナリティデータの項目である可能性があります。SAS Visual Statistics は、適切なパフォーマンスを確保し、有意な結果を得るために、高カーディナリティデータに制約を加えるメタデータプロパティ機能を備えています。各制約は、他の制約から独立しており、データやシステムのパフォーマンスのニーズに基づいて調整する必要があります。行った変更を適用するには、SAS Web アプリケーションサーバーを再起動します。

classCardinalityLimit および **groupbyCardinalityLimit** プロパティは、モデル作成時の作業負荷を制御することを目的としたものです。これらのプロパティには 2 つの別々の因果関係がありますが、それらは関連性があります。第 1 に、各プロパティに指定される値よりもモデルに含まれている個別階層が多い場合は、効果変数または Group BY 変数は指定できません。第 2 に、モデルの個別階層の総数が各プロパティに指定される値を超える場合は、そのモデルに効果変数または Group BY 変数は追加できません。

これに対して、**filterCardinalityLimit** および **responseCardinalityLimit** プロパティは、表示される個別階層の数のみに影響を及ぼします。フィルタの条件の選択時またはロジスティック回帰分析モデルの目的の階層を選択時に、この制限によって、有意な数の個別階層のみが表示されるようになります。この値より多くの個別階層を含むフィルタ変数やロジスティック回帰モデル応答変数を指定できますが、上位 *n* 階層のみが表示されます。

これらのメタデータプロパティにアクセスするには、SAS Management Console を開きます。**プラグイン**タブで、**アプリケーション管理** ▶ **構成マネージャ** ▶ **SAS Application Infrastructure** ▶ **Visual Analytics** ▶ **Visual Statistics** の順に選択します。**Visual Statistics** を右クリックし、**プロパティ**を選択します。**詳細**タブを選択します。次の 4 つのメタデータプロパティが表示されます。

vstat.classCardinalityLimit	2048
vstat.filterCardinalityLimit	1024
vstat.groupbyCardinalityLimit	1024
vstat.predictorBinsCardinalityLimit	1024
vstat.predictorCardinalityLimit	1024
vstat.responseCardinalityLimit	1024

classCardinalityLimit
モデルのすべての分類効果および交互作用項に許容される最大個別階層数を決定します。この制約は、モデルに含まれる分類効果および交互作用項の累積総数に課されます。この制約は、モデルの構築または更新時に必ず計算されます。

filterCardinalityLimit
フィルタ変数の個別階層の最大数を決定します。この値は、右ペインの**フィルタ**タブに表示される個別階層の最大数です。フィルタ変数には、この値よりも多くの個別階層を含めることができますが、最初の *n* 階層のみが表示されます。

groupbyCardinalityLimit
Group BY 変数に許容される個別階層の最大数を決定します。この制約は、Group BY 変数の累積総数に課されます。この制約は、モデルの構築または更新時に必ず計算されます。

responseCardinalityLimit
ロジスティック回帰分析モデルの応答変数に表示される個別階層の最大数を決定します。この値は、目的の階層を指定した場合に表示される個別階層の最大数です。この値よりも多い個別階層数を持つ応答変数を指定する場合、最初の *n* 階層のみが表示されます。

カーディナリティ制限は、入力データ、コンピュータシステム、統計モデルへの精通度などに基づいて決定してください。