



THE  
POWER  
TO KNOW.

# **SAS<sup>®</sup> Visual Statistics 7.1**

## User's Guide

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2014. *SAS® Visual Statistics 7.1: User's Guide*. Cary, NC: SAS Institute Inc.

### **SAS® Visual Statistics 7.1: User's Guide**

Copyright © 2014, SAS Institute Inc., Cary, NC, USA

All rights reserved. Produced in the United States of America.

**For a hard-copy book:** No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

**For a web download or e-book:** Your use of this publication shall be governed by the terms established by the vendor at the time you acquire this publication.

The scanning, uploading, and distribution of this book via the Internet or any other means without the permission of the publisher is illegal and punishable by law. Please purchase only authorized electronic editions and do not participate in or encourage electronic piracy of copyrighted materials. Your support of others' rights is appreciated.

**U.S. Government License Rights; Restricted Rights:** The Software and its documentation is commercial computer software developed at private expense and is provided with RESTRICTED RIGHTS to the United States Government. Use, duplication or disclosure of the Software by the United States Government is subject to the license terms of this Agreement pursuant to, as applicable, FAR 12.212, DFAR 227.7202-1(a), DFAR 227.7202-3(a) and DFAR 227.7202-4 and, to the extent required under U.S. federal law, the minimum restricted rights as set out in FAR 52.227-19 (DEC 2007). If FAR 52.227-19 is applicable, this provision serves as notice under clause (c) thereof and no other notice is required to be affixed to the Software or documentation. The Government's rights in Software and documentation shall be only those set forth in this Agreement.

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

July 2015

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

---

# Contents

<i>Using This Book</i> .....	<i>vii</i>
------------------------------	------------

## PART 1 Introduction to SAS Visual Statistics 1

<b>Chapter 1 / About SAS Visual Statistics</b> .....	<b>3</b>
What Is SAS Visual Statistics? .....	3
Benefits of Using SAS Visual Statistics .....	3
Accessing SAS Visual Statistics .....	4
<b>Chapter 2 / The SAS Visual Statistics User Interface</b> .....	<b>7</b>
Your First Look at the SAS Visual Statistics User Interface .....	8
Managing Projects and Models .....	18
Specifying Your Preferences .....	20
Integration with SAS Visual Analytics Explorer .....	20

## PART 2 Model Building 23

<b>Chapter 3 / Modeling Information</b> .....	<b>25</b>
Available Models .....	25
Variables and Interaction Terms .....	26
Variable Selection .....	28
Missing Values .....	28
Group By Variables .....	28
Filter Variables .....	30
Model Score Code .....	31

<b>Chapter 4 / Linear Regression Model</b>	<b>33</b>
Overview of the Linear Regression Model	33
Linear Regression Model Properties	34
Linear Regression Model Results Windows	35
<b>Chapter 5 / Logistic Regression Model</b>	<b>45</b>
Overview of the Logistic Regression Model	45
Logistic Regression Model Properties	46
Logistic Regression Model Results Windows	47
<b>Chapter 6 / Generalized Linear Model</b>	<b>59</b>
Overview of the Generalized Linear Model	59
Generalized Linear Model Properties	60
Generalized Linear Model Results Windows	63
<b>Chapter 7 / Decision Tree</b>	<b>71</b>
Overview of the Decision Tree	71
Decision Tree Properties	72
Information Gain and Gain Ratio Calculations	74
Decision Tree Results Windows	76
<b>Chapter 8 / Cluster</b>	<b>85</b>
Overview of the Cluster Tool	85
Cluster Properties	86
Cluster Results Windows	87
<b>Chapter 9 / Model Comparison</b>	<b>91</b>
Overview of Model Comparison	91
Model Comparison Usage	92
Model Comparison Properties	93
Model Comparison Results Windows	93
<b>Chapter 10 / SAS Visual Statistics Example</b>	<b>97</b>
Overview	97
Create the Project	98
Create a Decision Tree	98

Create a Linear Regression .....	100
Create a GLM .....	103
Perform a Model Comparison .....	105

## PART 3 Administrative Tasks 109

<b>Chapter 11 / Installation and Configuration</b> .....	<b>111</b>
Installation .....	111
Configuration .....	111
<b>Recommended Reading</b> .....	<b>117</b>



# Using This Book

---

---

## Audience

SAS Visual Statistics is designed for use by data miners, statisticians, data scientists, database marketers, and business analysts who need to analyze large sets of diverse data and interactively build and evaluate predictive models to quickly get precise insights.







# Part 1

## Introduction to SAS Visual Statistics

### *Chapter 1*

***About SAS Visual Statistics* ..... 3**

### *Chapter 2*

***The SAS Visual Statistics User Interface* ..... 7**



## 1

## About SAS Visual Statistics

<i>What Is SAS Visual Statistics?</i> .....	3
<i>Benefits of Using SAS Visual Statistics</i> .....	3
<i>Accessing SAS Visual Statistics</i> .....	4

---

## What Is SAS Visual Statistics?

SAS Visual Statistics is an add-on to SAS Visual Analytics that enables you to develop and test models using the in-memory capabilities of SAS LASR Analytic Server. SAS Visual Analytics Explorer (the explorer) enables you to explore, investigate, and visualize data sources to uncover relevant patterns. SAS Visual Statistics extends these capabilities by creating, testing, and comparing models based on the patterns discovered in the explorer. SAS Visual Statistics can export the score code, before or after performing model comparison, for use with other SAS products and to put the model into production.

---

## Benefits of Using SAS Visual Statistics

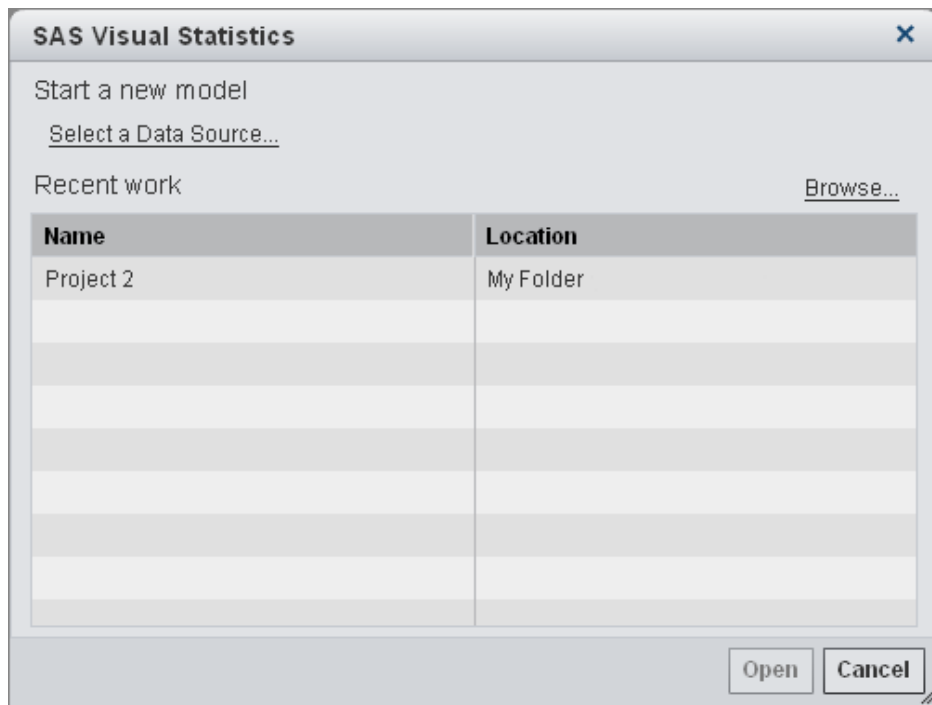
SAS Visual Statistics enables you to rapidly create powerful statistical models in an easy-to-use, web-based interface. After you have created two or more competing models for your data, SAS Visual Statistics provides a model comparison tool. The

model comparison tool enables you to evaluate the relative performance of two or more models against each other and to choose a champion model. A wide variety of model selection criteria is available. Regardless of whether you choose to perform a model comparison, you are able to export model score code for any model that you create. With exported model score code, you can easily apply your model to new data.

## Accessing SAS Visual Statistics

SAS Visual Statistics uses the standard sign-in window for SAS applications. To display the sign-in window, use the URL that is supplied by your system administrator. For example, you might enter `http://host/SASVisualStatistics`.

After you have entered the appropriate URL to access SAS Visual Statistics, you must sign in using the user ID and password provided by your system administrator. When you sign in to SAS Visual Statistics, the Welcome window appears. In the Welcome window, you can choose to create a new project or open a recent project.



The Welcome window enables you to perform the following tasks:

- Create a new model by clicking **Select a Data Source**. The Data sources window appears.
- Open an existing model. Select from your recent work or click **Browse** to select any work.

To exit SAS Visual Statistics, click the **Sign Out** link in the upper right corner of the SAS Visual Statistics user interface.

By default, if there is no activity for a specified period of time or the connection to the server is lost, you are signed out automatically. When this happens, your progress is not saved. You must start again from your most recent saved state. Your system administrator specifies the inactivity period and whether to return to the application or to display the sign-in window after your session times out.



## 2

# The SAS Visual Statistics User Interface

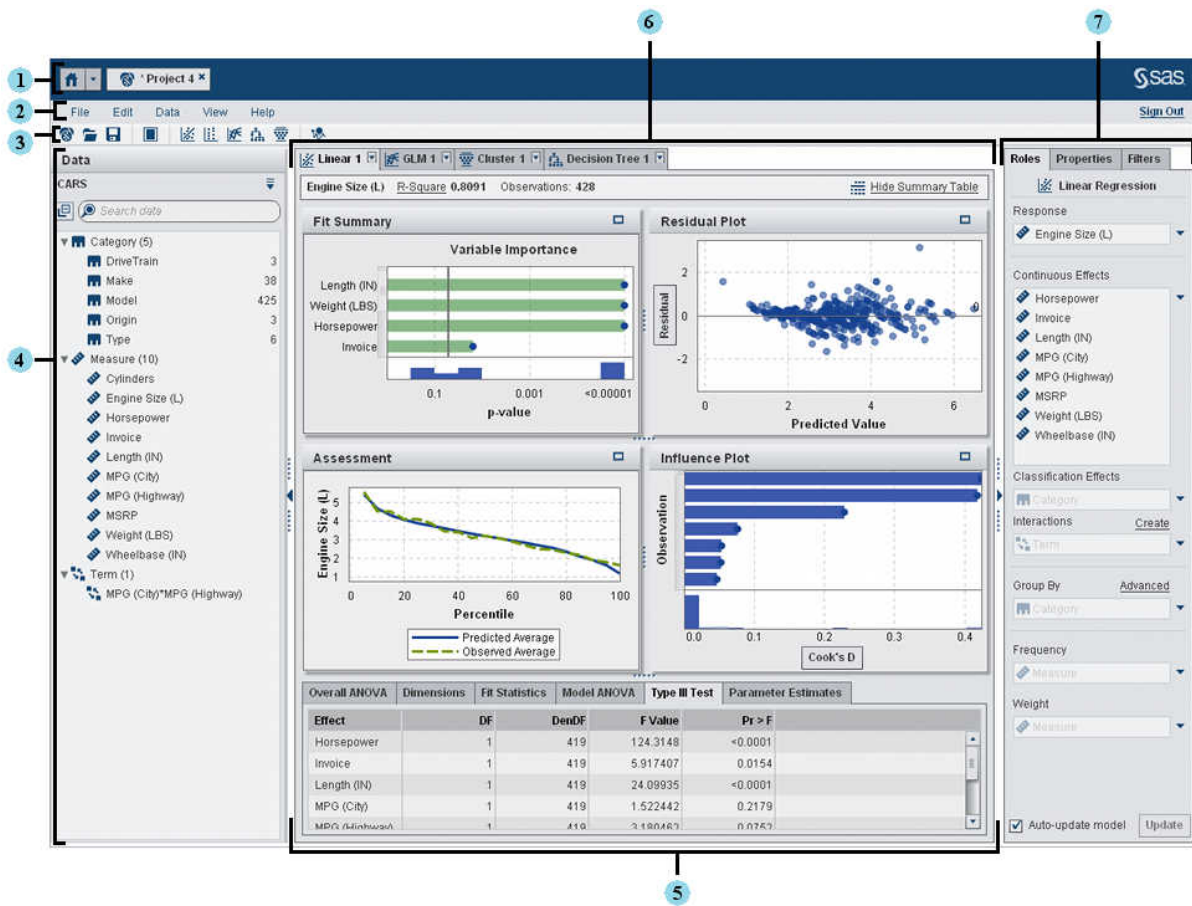
<b><i>Your First Look at the SAS Visual Statistics User Interface</i></b> .....	<b>8</b>
Overview .....	8
Menus and Toolbars .....	9
The Data Pane .....	10
The Right Pane .....	13
The Model Pane .....	14
<b><i>Managing Projects and Models</i></b> .....	<b>18</b>
Projects .....	18
Models .....	19
<b><i>Specifying Your Preferences</i></b> .....	<b>20</b>
<b><i>Integration with SAS Visual Analytics Explorer</i></b> .....	<b>20</b>

# Your First Look at the SAS Visual Statistics User Interface

## Overview

This section covers the user interface components and common navigation tasks for SAS Visual Statistics. Here are the main components of the SAS Visual Statistics user interface:

Figure 2.1 The SAS Visual Statistics User Interface











- 1 The application bar enables you to access the SAS Visual Analytics home page and recent projects.
- 2 The menu bar enables you to access all of the features of SAS Visual Statistics.
- 3 The toolbar enables you to quickly access the most commonly used features of SAS Visual Statistics.
- 4 The **Data** pane displays the variables that are available for analysis.
- 5 The summary table displays detailed statistics for the current model.
- 6 The model pane enables you to access created models and displays results plots for the current model.
- 7 The right pane enables you to access the **Roles**, **Properties**, and **Filters** tabs.






## Menus and Toolbars

From the SAS Visual Statistics main menu, you are able to access all of the features of the application.

The SAS Visual Statistics toolbar enables you to quickly access frequently used tasks.

The following icons are available on the SAS Visual Statistics toolbar:



Icon	Description
	Creates a new project.
	Opens a saved project.
	Saves the current project.
	Maximizes the modeling workspace.
	Returns the modeling workspace to the default view.
	Creates a linear regression model.


Icon	Description
	Creates a logistic regression model.
	Creates a generalized linear model (GLM).
	Creates a decision tree model.
	Creates a cluster model.
	Compares two or more models.

## The Data Pane

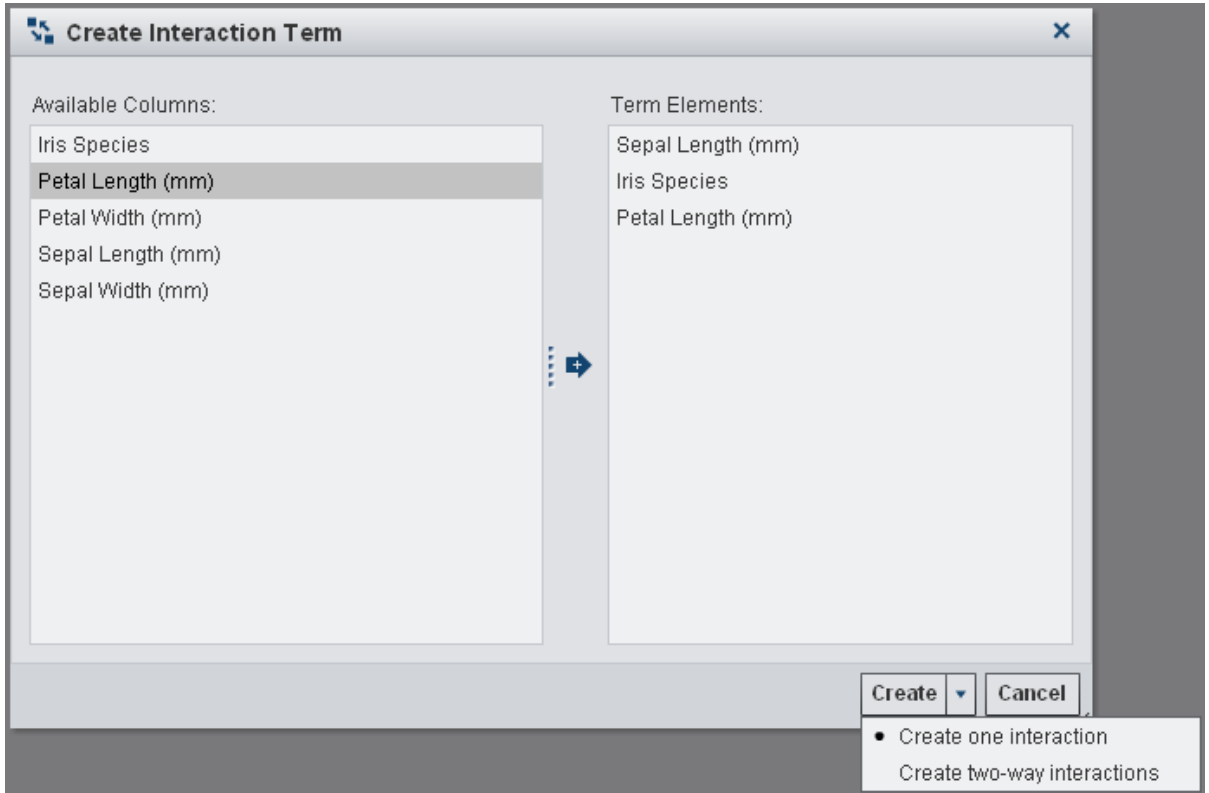
The **Data** pane enables you to access all of the variables in your data set. Variables are sorted into **Category** and **Measure** variable groups. Category variables have discrete levels. Measure variables are continuous. You can create variable interactions, which are available in the **Term** group.

When you enter something into the **Search data** field, only those variables that contain the search term are displayed. Search terms are not case sensitive.

To collapse the variable groups, click  to the left of the **Search data** field. To expand the variable groups, click  to the left of the **Search data** field.

The **Data** pane drop-down list icon,  , is located in the upper right corner of the **Data** pane. The following items are available:

- **Create Interaction** opens the Create Interaction Term window.



In the Create Interaction Term window, available variables are in the **Available Columns** area. Move the variables that you want to use to the **Term Elements** area by dragging and dropping the variables, double-clicking a variable, or using the arrows in the center of the window. After you have moved variables to the **Term Elements** area, click **Create** to create a single interaction. Alternatively, click the ▼ to specify whether you want to **Create one interaction** or **Create two-way interactions**.

When you choose **Create two-way interactions**, all of the possible pairs of interactions for the selected variables are created, except for square terms. For example, in the previous image, the two-way interactions are Sepal Length (mm) and Iris Species, Sepal Length (mm) and Petal Length (mm), and Iris Species and Petal Length (mm). To create a square term, select that variable in the **Data** pane, right-click the variable, and select **Create a Single Interaction**.

- **Data Properties** opens the Data Properties window. The Data Properties window displays the name, classification, data type, model type, and format for each variable in the data set.
- **Measure Details** opens the Measure Details window. The Measure Details window provides summary statistics and a histogram for each of the measure variables.
- **Show/Hide Items** opens the Show or Hide Items window. Variables in the **Visible items** area are displayed in the **Data** pane. Variables in the **Hidden items** area are not displayed.

To move variables from one area to the other, drag and drop the variable, double-click a variable, or use the arrows in the center of the window. You can move multiple variables by selecting them first, and then either dragging and dropping them or using the arrows.

Click **OK** to close the Show or Hide Items window and save your changes.

- **Sort Items** enables you to specify whether you want to sort the variables in ascending or descending order.

When you right-click a variable or interaction in the **Data** pane, a pop-up menu appears. The following items are available in this pop-up menu. A subset of items are available based on whether you right-clicked a category variable, measure variable, or term.

## Assign

enables you to assign the variable to one or more of the following roles:

- **Response** is available only when the variable type matches the model's response variable type.
- **Continuous Effect** is available only for a measure variable.
- **Classification Effect** is available only for a category variable.
- **Interactions** is available only for a term.
- **Weight** is available only for a measure variable.
- **Frequency** is available only for a measure variable.
- **Group By** is available only for a category variable.

## ■ Filter

### Create a Single Interaction

creates a single interaction for the selected variables. Creates a square interaction if only one variable is selected.

### Rename

enables you to specify a new display name for the selected variable.

### Hide

hides the selected variable.

### Delete

deletes the selected term. This is available only for items created in SAS Visual Statistics.

### Category

specifies whether the selected variable is a category variable.

### Measure




specifies whether the selected variable is a measure variable. A numeric variable can be assigned as either a category variable or measure variable.

### Properties

displays information about the selected variable or term.

## The Right Pane

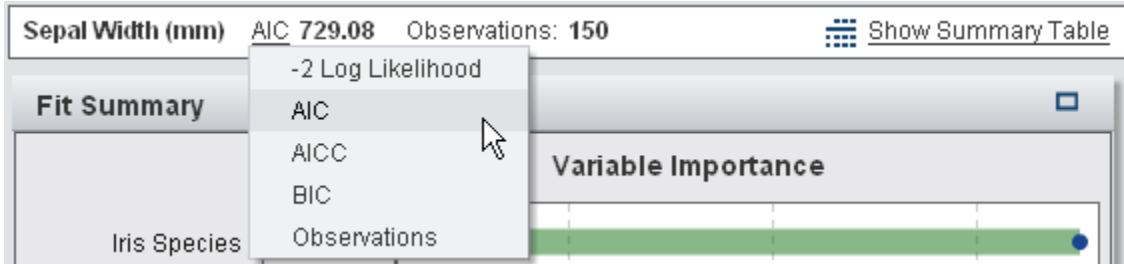
The right pane contains the **Roles**, **Properties**, and **Filters** tabs. Together, these three tabs define the modeling parameters. The **Roles** tab specifies what variables are used in the model and their purposes in the model. The **Properties** tab enables you to specify features that are unique to each model. The **Filters** tab enables you to subset the data that is modeled.

Tab	Description
Roles	Enables you to add variables to your model. From the <b>Data</b> pane, drag and drop a variable that you want to use to a role on the <b>Roles</b> tab. Alternatively, you can select several variables and drag and drop them onto the model pane. In this case, each variable is assigned to the first valid and available role. If there is no response variable, then the first valid variable is assigned to the <b>Response</b> role. This method never assigns <b>Group By</b> , <b>Frequency</b> , <b>Filter</b> , or <b>Weight</b> variables. You can use the  icon to add or remove variables from each individual field.
Properties	Enables you to specify features of the model. The available options vary based on the selected model.
Filters	Enables you to specify variables that are used to filter the data set. You can filter on category variables, measure variables, or both. To add a filter variable, drag and drop that variable from the <b>Data</b> pane to the <b>Filters</b> tab or use the  icon. To remove a filter variable, click  next to that variable's name.

## The Model Pane



The model pane contains the modeling results and plots. Because the windows available depend on the selected model, this section focuses on the common elements for all models. Specific information for each model is available in the chapter for that model.

The summary bar displays the response variable, model evaluation criterion (when available), and number of observations used in the model. To see all of the available model evaluation criteria, click the name of the current model evaluation criterion in the summary bar to open a pop-up menu.



On the right side of the summary bar, there is **Show Summary Table**. Click **Show Summary Table** to open the summary table at the bottom of the model pane. An example summary table from a decision tree model is shown below. The specific information in each summary table varies based on the model.

Node Statistics		Node Rules								
Node ID	Depth	Parent	Num Child	Type	NObs	Percent	NMiss	Gain Ratio	Predicted Value	Split
0	0	-1	2	Class	150	100.00%	0	0.35510569	4	
1	1	0	2	Class	100	66.67%	0	0.24064902	4	Virginica, Vers
2	1	0	2	Class	50	33.33%	0	0.5244616	5	Setosa
3	2	1	0	Leaf	35	23.33%	0	0.22460961	4	>= 64.6
4	2	1	2	Class	65	43.33%	0	0.25836628	2	< 64.6
5	2	2	0	Leaf	22	14.67%	0	0.15927894	7	>= 50.2

Except for the decision tree model, by default, all windows that are available are shown in the model pane. To maximize a window, click  in the upper right corner of the window. This hides all of the other windows in the model pane, but does not hide the summary table. To restore the default view, click .

Available windows are provided in the following table:

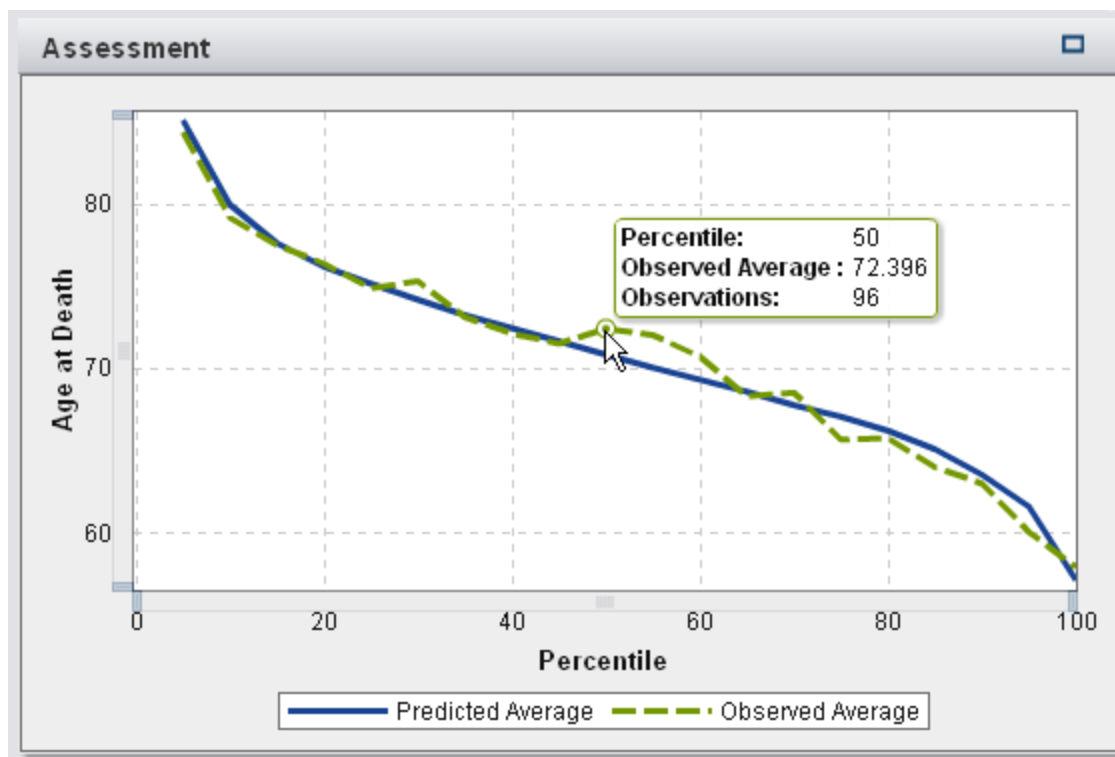
Window Name	Available In	Summary
Fit Summary	Linear, Logistic, Generalized Linear Model (GLM)	<p>Displays the <math>p</math>-value of each modeling variable on a log scale. The alpha value, plotted as <math>-\log(\alpha)</math>, is shown as a vertical line that you can click and drag to adjust. A histogram of the <math>p</math>-values is displayed at the bottom of the window.</p> <p>This window is divided when a Group By variable is used. The left side lists the groups and the right side condenses the <math>p</math>-values for each group into a single linear scatter plot. You can click on a group on the left side to change the Residual, Influence, and Assessment plots to show only the results for that group.</p>
Residual Plot	Linear, Logistic, GLM	<p>Displays various residual plots for the model. When the plot labels are buttons, you can select the values that are plotted on that axis. Each model has a unique set of plot combinations available.</p>
Assessment	Linear, Logistic, GLM, Decision Tree	<p>For measure target variables, Assessment plots the average predicted and observed values against the binned data set. For category target variables, it provides the Lift, ROC, and Misclassification plots.</p>



Window Name	Available In	Summary
Influence Plot	Linear, Logistic	Plots each observation against various computed statistics. The X axis label is a button that enables you to determine what is plotted. Each model has a unique set of plot combinations available.
Tree	Decision Tree	Displays the decision tree and the decision treemap. You can interactively train the decision tree from this window. Use the scroll wheel on your mouse to zoom in or out on the location of the mouse pointer.
Leaf Statistics	Decision Tree	Provides a stacked histogram of the response variable for each leaf node in the decision tree.
Cluster Matrix	Cluster	Displays the two-dimensional projection of every cluster for each pair of modeling variables. To view a larger plot of an individual projection, right-click in that cell, and select <b>Open</b> .
Parallel Coordinates	Cluster	Displays a color-coded strand for each observation, initially sorted by cluster membership. You can restrict the display to observations that match specific clusters or ranges of the modeling variables.

In every window, if you position your mouse pointer over an object, a tooltip provides specific information about that object. The information varies based on the plot that is

being displayed. For example, in the image below, the tooltip shows the percentile (in the bin), observed average value in that bin, and the number of observations in that bin.




Whenever a range of values is shown or selected, the interval is half-open. The minimum value is included in the interval. The maximum value is excluded from the interval. This affects heat maps, the parallel coordinates plot, and any other displayed or selected interval.

## Managing Projects and Models


### Projects


A SAS Visual Statistics project consists of one or more models and the associated data. Each project can contain only one data set. If you want to change the active data set, you must create a new project.

You can create a project using any of the following methods:

- Select **File** ► **New** ► **Project** from the main menu. Select a data source, and then click **Open**.
- Click the  icon in the toolbar.

**Note:** If a project is active, a window appears that asks whether you want to save your current project. Click **Save** to save your changes to the current project and open a new project, **Don't Save** to discard your changes to the current project and open a new project, or **Cancel** to return to the current project.

To save your current project, select **File** ► **Save As**, and then select a location and a name. Alternatively, click  in the toolbar to save your project.

To close your current project, select **File** ► **Close** from the main menu. Alternatively, you can click the  next to the project name in the application bar.

## Models

To create a model, select **File** ► **New** and a model type from the main menu. You can also create a model by clicking the icon for that model type in the toolbar.

To rename a model, select **Edit** ► **Rename** from the main menu. The Rename window appears. Enter a new name in the **New name** field, and click **OK**. This affects the current model in the model pane.

To duplicate a model, select **Edit** ► **Duplicate** from the main menu. This creates a model named Copy of <Model Type> with the same settings as the copied model. You might want to duplicate a model if you have a good model, but you think that some enhancements might improve it, and you do not want to risk losing the current model. This feature enables you to leave the original model intact while you adjust the duplicate model. This affects the current model in the model pane.

To delete a model, select **Edit** ► **Delete** from the main menu. This affects the current model in the model pane. Every time you delete a model, you are asked to confirm the action. Select **Don't show this message again** if you want to avoid this prompt in the future. You can reset this prompt in the Preferences window.

In addition to using the main menu, you can use the ▼ located next to a model name to rename, duplicate, or delete a model. The **Show Summary Table** option is also available from this icon..

---

## Specifying Your Preferences

To access the Preferences window, select **File ► Preferences** from the main menu. In the Preferences window, you are able to specify global preferences and local preferences. Global preferences persist across SAS web applications and include user locale information and the display theme. Local preferences apply only to SAS Visual Statistics and include the default model type, stepper delay time, and  $p$ -value precision.

All of your preferences persist between SAS Visual Statistics sessions.

---

## Integration with SAS Visual Analytics Explorer

SAS Visual Analytics Explorer (the explorer) provides a quick method for you to launch SAS Visual Statistics. After you have loaded your data in the explorer, select **File ► Extended Features ► View Data in SAS Visual Statistics**. This action launches SAS Visual Statistics with the data that you loaded in the explorer.

In addition, you can launch SAS Visual Statistics from a box plot, scatter plot, or correlation matrix in the explorer. In any of these visualizations, right-click inside the visualization, and select **Extended Features ► Model Responses In SAS Visual Statistics**.

For the scatter plot, you must specify two measures. The variable that is plotted on the vertical axis is specified as the response variable in SAS Visual Statistics. When available, the specified default model type is used. A linear regression model is used when the default model type is unavailable.

For the correlation matrix, you can transfer either a single cell or an entire row or column. When you select a single cell, two variables are transferred. The variable that is

plotted on the vertical axis is specified as the response variable in SAS Visual Statistics. When you select an entire row or column, the variable that defines the selected row or column is specified as the response variable. You cannot select nonadjacent cells and transfer their data to SAS Visual Statistics. Similarly, if you select adjacent cells that are not contained in a single row or column, then you cannot transfer their data to SAS Visual Statistics. When available, the specified default model type is used. A linear regression model is used when the default model type is unavailable.

For the box plot, you must specify at least one category variable and at least one measure variable. The variable in the **Category** field is always specified as the response variable in SAS Visual Statistics. Additional lattice variables are assigned as Group By variables. When available, the specified default model type is used. A logistic regression mode is used when the default model type is unavailable.

**Note:** When you transfer data from the explorer to SAS Visual Statistics, only the raw data is transferred. Renamed variables, classification changes, hidden variables, and other changes are not applied when the data is opened in SAS Visual Statistics.

Similarly, from SAS Visual Statistics, you are able to launch the explorer. After you have created a project in SAS Visual Statistics, select **File ► Extended Features ► View Data in SAS Visual Analytics Explorer**.





# Part 2

## Model Building

Chapter 3	
<i>Modeling Information</i> .....	25
Chapter 4	
<i>Linear Regression Model</i> .....	33
Chapter 5	
<i>Logistic Regression Model</i> .....	45
Chapter 6	
<i>Generalized Linear Model</i> .....	59
Chapter 7	
<i>Decision Tree</i> .....	71

Chapter 8	
<b>Cluster</b> .....	<b>85</b>
Chapter 9	
<b>Model Comparison</b> .....	<b>91</b>
Chapter 10	
<b>SAS Visual Statistics Example</b> .....	<b>97</b>



3

# Modeling Information

<i>Available Models</i> .....	25
<i>Variables and Interaction Terms</i> .....	26
Variables .....	26
Interaction Terms .....	27
<i>Variable Selection</i> .....	28
<i>Missing Values</i> .....	28
<i>Group By Variables</i> .....	28
<i>Filter Variables</i> .....	30
<i>Model Score Code</i> .....	31

## Available Models

The following models are available in SAS Visual Statistics:

- [Linear Regression on page 33](#) attempts to predict the value of an interval response as a linear function of one or more effect variables.
- [Logistic Regression on page 45](#) attempts to predict the probability that a binary or ordinal response will acquire the event of interest as a function of one or more effects.

- [Generalized Linear Model on page 59](#) is an extension of a traditional linear model that allows the population mean to depend on a linear predictor through a nonlinear link function.
- [Decision Tree on page 71](#) creates a hierarchical segmentation of the input data based on a series of rules applied to each observation.
- [Cluster on page 85](#) segments the input data into groups that share similar features.

---

## Variables and Interaction Terms

### Variables

#### Category Variables

Category variables are numeric or nonnumeric variables with discrete levels. The levels of a category variable are considered unordered by SAS Visual Statistics. Examples of category variables include drink size (small, medium, or large), number of cylinders in an engine (2, 4, 6, or 8), or whether a customer has made a purchase (yes or no).

You can create a category variable from a response variable by right-clicking the variable and selecting **Category**. In this case, each distinct value of the measure variable is turned into a level for the category variable.

Category variables can be used as response variables for classification models, classification effect variables, decision tree predictors, filter variables, and group by variables.

**Note:** To ensure proper performance and valid modeling results, the maximum number of distinct levels allowed for a category variable is limited based on the model type and variable role.

#### Measure Variables

Measure variables are continuous numeric variables that can assume an infinite number of possible values between two numbers. Even though some numeric variables are not

continuous, such as count variables, these variables can be treated as continuous values for the purpose of modeling. Examples of measure variables include the temperature of a drink, engine displacement amount, or a customer's total purchase amount.

Summary statistics and a histogram for each measure variable are obtained by right-clicking the variable in the **Data** pane, and selecting **Properties**. Use the **Name** drop-down menu to specify the variable that you want to view.

Measure variables can be used as response variables for continuous models, continuous effect variables, decision tree predictors, offset variables, frequency variables, weight variables, and filter variables.

## Interaction Terms

Two variables, A and B, *interact* if the effect of one variable on the model changes as the other variable changes. That is, the effects of variables A and B are not additive in the model.

SAS Visual Statistics enables you to create interactions between two or more input variables, including squared interactions. A squared interaction is the interaction of a variable with itself. You cannot create squared interactions for category variables.

For an example where interaction terms might be useful, consider a situation where you are modeling the fuel mileage (MPG) for several cars. Two of your input variables are engine displacement in liters and engine size (number of cylinders). You expect that as either value increases, fuel mileage will suffer. However, if you suspect that the effects on fuel mileage that are attributable to engine displacement are not constant across engine size, then you should consider creating the interaction term between those variables.

SAS Visual Statistics is not limited to creating just two-way interactions. You can create  $n$ -way interactions that include an arbitrary number of variables, but not more than the number of available input variables.

The number of distinct levels for an interaction term is the product of the number of levels for each variable in the term. Measure variables are treated as if they contain one

level. The number of levels in an interaction term counts against the maximum number of distinct levels allowed in regression models.

---

## Variable Selection

Variable selection is the process of reducing the number of input variables to include just the most significant variables. The Linear Regression and Logistic Regression models provide a property to automatically perform variable selection. When you use this property, SAS Visual Statistics performs backward selection on the input variables to determine the most significant variables. Modeling with just the most significant variables is intended to avoid creating a model that overfits the data. Automated variable selection can actually take longer to run than not performing variable selection.

---

## Missing Values

By default, SAS Visual Statistics handles missing values by dropping all observations that contain a missing value in any assigned role variable. However, the Linear Regression, Logistic Regression, and GLM models provide the **Informative missingness** property. In some cases, the fact that an observation contains a missing value provides relevant modeling information. Selecting this property explicitly models missing values of variables as a separate variable. For measure variables, missing values are imputed with the observed mean, and an indicator variable is created to denote missingness. For category variables, missing values are considered a distinct level.

---

## Group By Variables

A group by variable enables you to fit a model for each data segment defined by one or more category variables. Each unique combination of levels across all of the group by variables is a specific data segment. For example, if you have one group by variable

with three levels, then there are three data segments. But, if you have two group by variables, one with three levels and the other with four levels, then there are at most 12 data segments. A data segment is not created when there are no observations in a combination of classification levels.

SAS Visual Statistics enforces a maximum number of BY groups, except when you use the Advanced Group By feature. By default, the maximum number of BY groups allowed is 1024. Empty data segments count against the maximum number of BY groups allowed in a model.

When you specify two or more group by variables, the results are grouped in the order in which the variables appear in the **Group By** field.

In the Fit Summary window, when you select a specific data segment, the Residual Plot and Influence Plot windows are updated to include only the observations in the specified data segment.

The Advanced Group By window provides more control over variable grouping. To access the Advanced Group By window, click **Advanced** next to **Group By** in the right pane.

**Advanced Group By**

Group By: Make

☒ Use advanced features

Measure: Engine Size (L)

Aggregation: Sum

Count: Top

100

Results:

Name	Value
Mercedes-Benz	101.5
Chevrolet	100.8
Ford	81.7
Toyota	75.1
BMW	62.5
Audi	58.1

OK Cancel

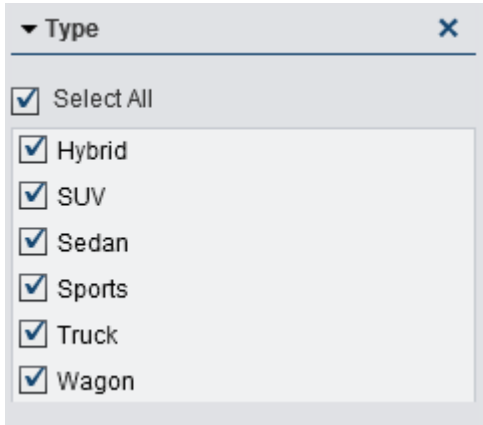
The **Group By** field enables you to select the variable that is used for grouping. Select the **Use advanced features** option to display aggregation statistics for a specified measure variable. Specify the measure variable in the **Measure** field. The **Aggregation** field specifies whether the **Average** or **Sum** is computed. Use the **Count** field to specify whether you want the **Top** or **Bottom**  $n$  values. The field below **Count** enables you to specify the value of  $n$ .

---

## Filter Variables

Filter variables are used to subset the modeling data. You can filter on any variable included in the data, not just on variables used in the model. Filter variables are applied only to the current model.

When you filter on a category variable, you are presented with a list of the levels for that variable. Select only values that you want to include in the model. In the following image, all levels are available.



When you filter on a measure variable, a slider lets you specify a range of values. Use the triangles to specify the lower and upper limits of the filter variable.



---

## Model Score Code

Model scoring refers to the process of generating predicted values for a data set that might contain the response variable of interest. Score code is exported as a SAS DATA step that can be executed on new data sets in any SAS environment. All variables used by the model in any capacity are included in the score code. This includes interaction terms, group by variables, frequency variables, and weight variables. Score code is not available for interactive decision trees.

To generate model score code, select **File** ► **Export** ► **Model Score Code** from the main menu. In the Export Model Score Code window, select the model that you want to

export, and click **OK**. In the Save As window, navigate to where you want to save the code, and click **Save**.

Score code is saved as a .sas file and can be viewed in any word processing program.



4

# Linear Regression Model

- Overview of the Linear Regression Model* ..... 33
- Linear Regression Model Properties* ..... 34
- Linear Regression Model Results Windows* ..... 35
  - The Fit Summary Window ..... 35
  - Residual Plot ..... 37
  - Assessment ..... 40
  - Influence Plot ..... 41
  - Fit Statistics ..... 42
  - Summary Table ..... 43

## Overview of the Linear Regression Model

A linear regression attempts to predict the value of a measure response variable as a linear function of one or more effects. The linear regression model uses the least squares method to determine the model. The least squares method creates a line of best fit by minimizing the residual sum of squares for every observation in the input data set. The residual sum of squares is the vertical distance between an observation and the line of best fit. The least squares method requires no assumptions about the distribution of the input data.

The linear regression model requires a measure response variable and at least one effect variable or interaction term.

---

## Linear Regression Model Properties

The following properties are available for the linear regression model:

### Name

enables you to specify the name for this model.

### Informative missingness

specifies whether the informative missingness algorithm is used. For more information, see [Missing Values on page 28](#).

### Use variable selection

specifies whether variable selection is performed. For more information, see [Variable Selection on page 28](#).

### Significance level

specifies the significance level that is required in order for variables to be considered for the model. This property is available only when **Use variable selection** is selected.

### Assessment

- **Use default number of bins** specifies whether you want to use the default number of bins or to set your own value. By default, measure variables are grouped into 20 bins.
- **Number** specifies the number of bins to use when the **Use default number of bins** property is not selected. You must specify an integer value between 5 and 100.
- **Tolerance** specifies the tolerance value that is used to determine the convergence of the iterative algorithm that estimates the percentiles. Specify a smaller value to increase the algorithmic precision.

**Show diagnostic plots**

specifies whether the Residual Plot, Assessment, and Influence Plot windows appear in the model pane.

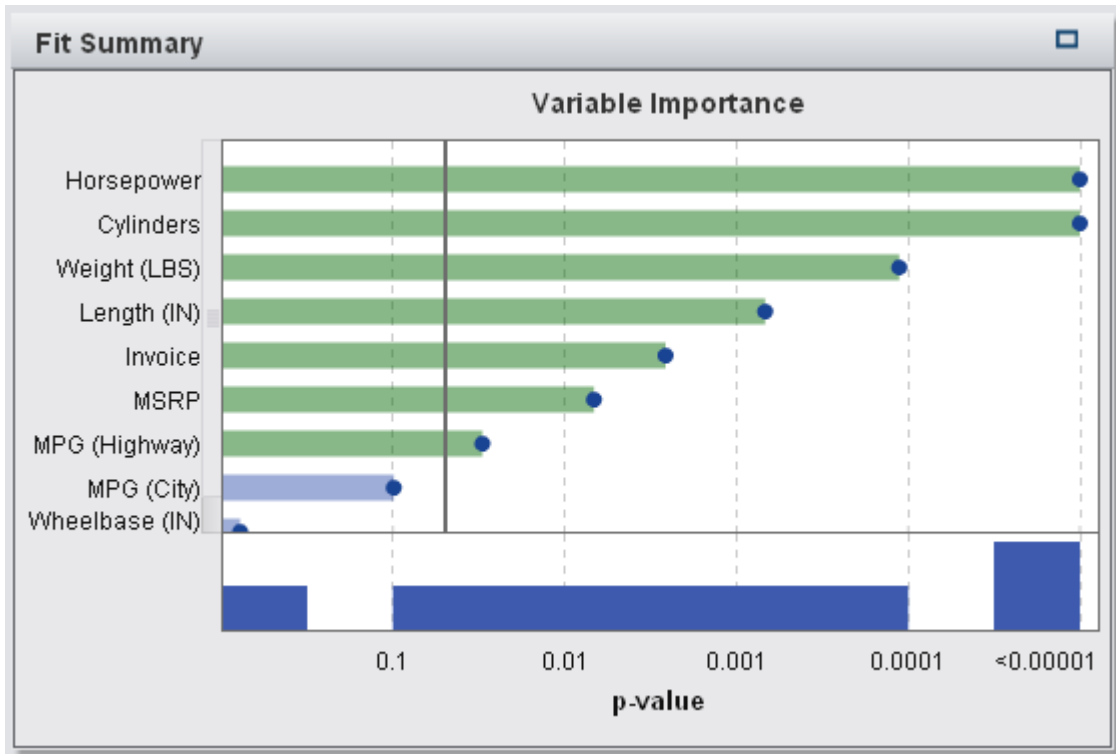
---

## Linear Regression Model Results Windows

### The Fit Summary Window

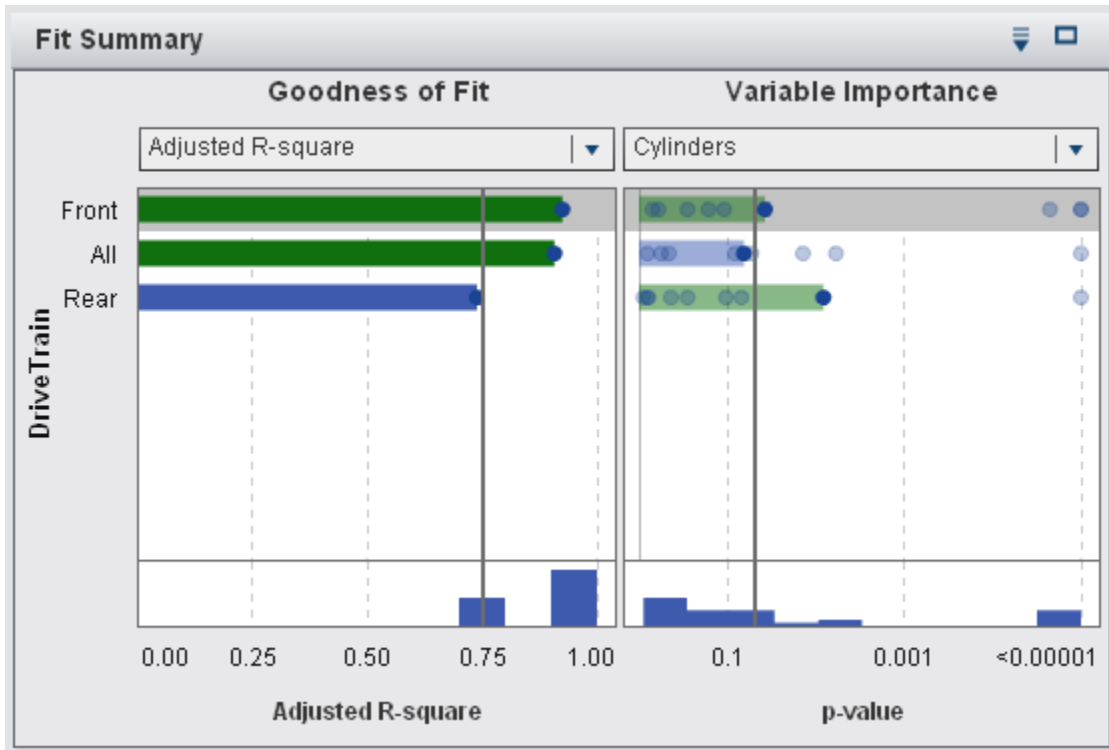
#### Without a Group By Variable

The Fit Summary window plots the relative importance of each variable as measured by its  $p$ -value. The  $p$ -value is plotted on a log scale and the alpha value (plotted as  $-\log(\alpha)$ ), is shown as a vertical line. To adjust the alpha value, click, drag, and drop the vertical line. A histogram of the  $p$ -values is displayed at the bottom of the window. A typical Fit Summary window is shown below.



### With a Group By Variable

When your analysis includes a group by variable, the Fit Summary window displays a different set of plots.



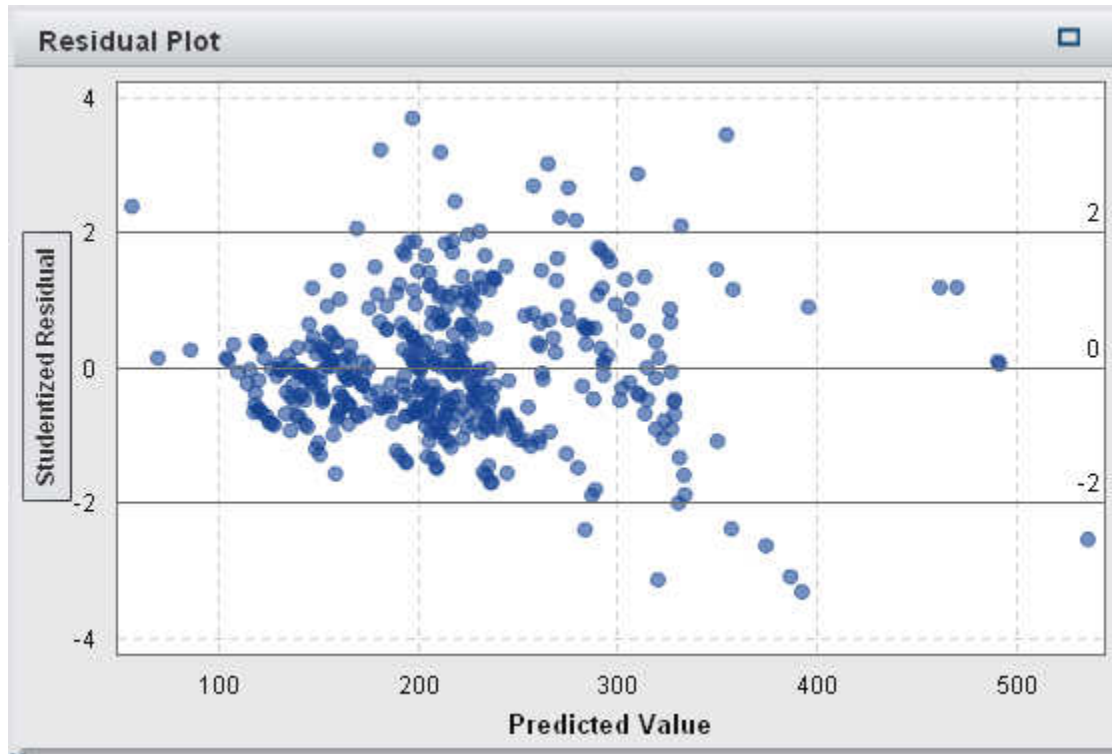
First, notice that the **Variable Importance** plot displays only a single variable. This is because the variable importance for every variable is computed within each level of the group by variable. Use the drop-down menu to view the variable importance for a different effect. Second, notice the **Goodness of Fit** plot, which is not available when there is no group by variable. This plot displays how well the model predicts the response variable within each level of the group by variable. Use this plot to determine whether your model has a significantly different fit within different levels.

Use the  icon to specify how the plot is sorted.

## Residual Plot

### Scatter Plot

The residual of an observation is the difference between the predicted response value and the actual response value. By default, the Residual Plot displays a scatter plot of the residuals against the predicted value, as shown below.



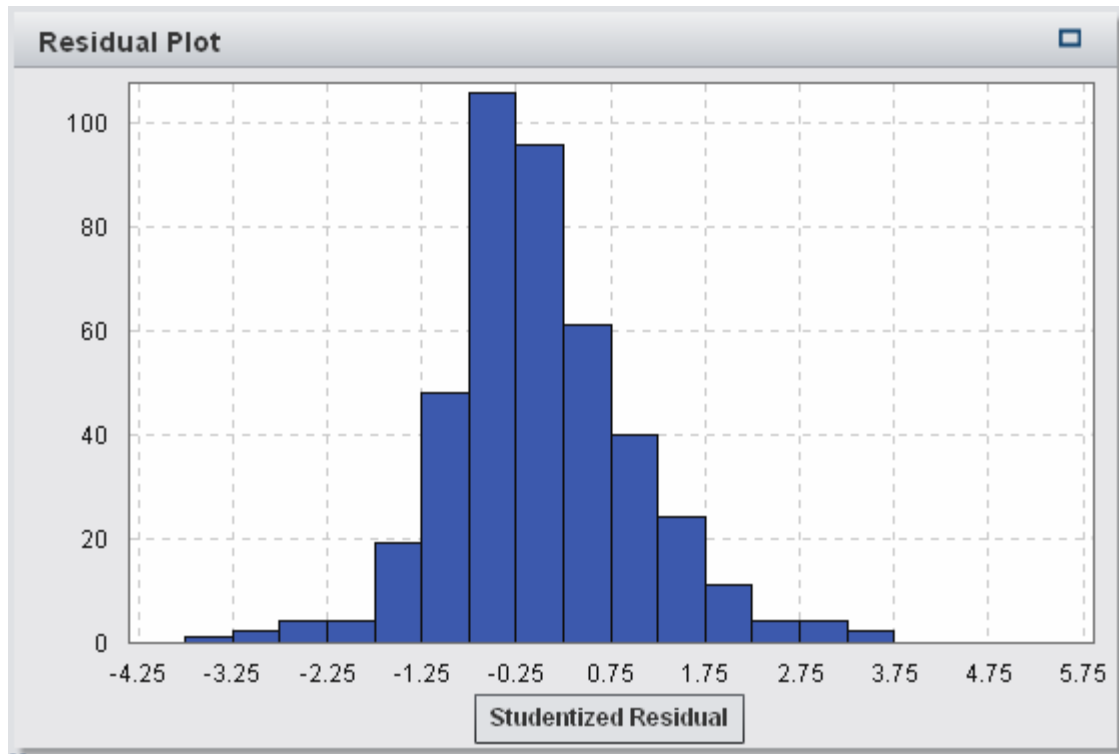
Notice that the label on the Y axis is a button. You can click this button to change the value that is plotted on the Y axis. For the Y axis, you can select from the Studentized Deleted Residual, Residual, Studentized Residual, and PRESS statistic.

Residual plots have several uses when examining your model. First, obvious patterns in the residual plot indicate that the model might not fit the data. Second, residual plots can detect nonconstant variance in the input data when you plot the residuals against the predicted value. Nonconstant variance is evident when the relative spread of the residual values changes as the predicted values change. Third, in combination with other methods, the residual plot can help identify outliers in your data.

When using very large data sets, the residual plots are displayed as heat maps instead of actual plots. In a heat map, the actual observations are binned, and the color of each point indicates the relative number of observations in that bin.

## Histogram

To view the data in the residual plot as a histogram, right-click in the Residual Plot window, and select **Use Histogram**. Each of the four values that were available on the Y axis of the residual plot are available as a histogram.



Click the label on the X axis to change the value that is plotted. You can select the Studentized Deleted Residual, Residual, Studentized Residual, and PRESS statistic.

From the histogram, it is fairly easy to determine whether the distribution of the residuals is approximately normal or skewed. For example, in the previous image, the residuals are skewed right. A non-normal residual histogram can indicate that the model does not fit the data.

## Box Plot

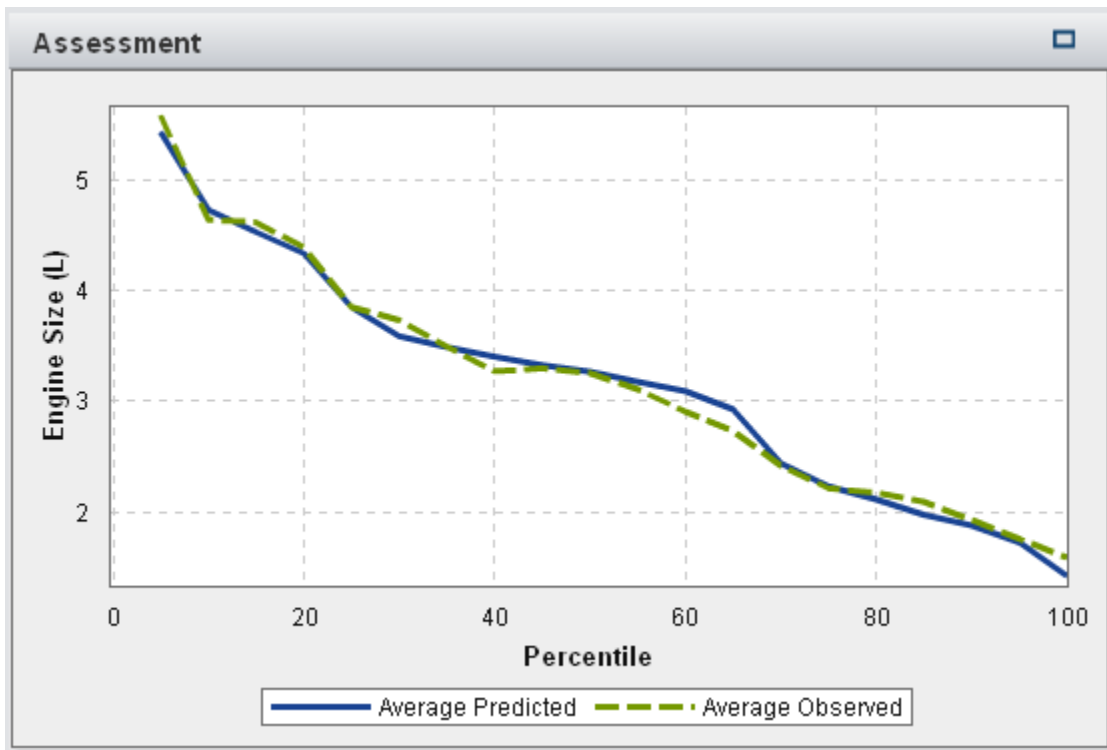
To view a box plot of the currently plotted residual, right-click in the Residual Plot window, and select **Plot By**. Then, select a category variable, which is used to group the residuals when the box plots are created. All category variables are available,

regardless of whether they are included in the model. For variables not included in the model, you can right-click in the Box Plot window, and assign each of those variables as either a classification effect or a group by variable. The **Assign to Classification** and **Assign to Group By** menu options are unavailable for variables that are already included in the model.

Outliers are hidden by default. To show outliers, right-click in the Box Plot window, and select **Show Outliers**.

## Assessment

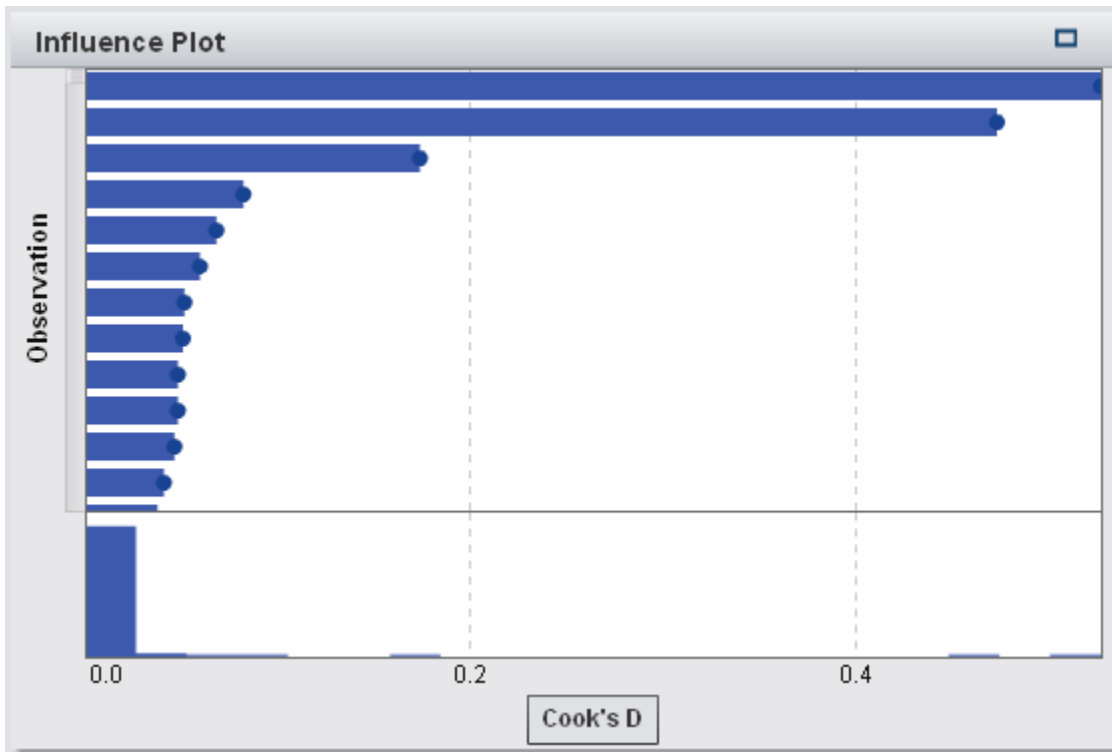
For a linear regression, the Assessment window plots the average predicted and average observed response values against the binned data. Use this plot to determine how well the model fits the data.





## Influence Plot

The Influence Plot displays several measurements that are computed for each observation. The histogram below the measurements is based only on the displayed observations. When the input data contains a large number of observations, the observations are binned. By default, the **Cook's D** value is plotted on the X axis.



Other available values are Covariance Ratio, DFFITS, Leverage, and Likelihood Displacement. You can view a histogram of these measurements by right-clicking in the Influence Plot window, and selecting **Use Histogram**. This histogram uses the entire data set and applied filters.

Use these values to help identify outliers and other points that greatly affect the predicted regression model.

## Fit Statistics

The linear regression model computes several assessment measures to help you evaluate how well the model fits the data. These assessment measures are available at the top of the model pane. Click the currently displayed assessment measure to see all of the available assessment measures.

### Adjusted R-square

The Adjusted R-squared value attempts to account for the addition of more effect variables. Values can range from 0 to 1. Values closer to 1 are preferred.

### AIC

Akaike's Information Criterion. Smaller values indicate better models, and AIC values can become negative. AIC is based on the Kullback-Leibler information measure of discrepancy between the true distribution of the response variable and the distribution specified by the model.

### AICC

Corrected Akaike's Information Criterion. This version of AIC adjusts the value to account for sample size. The result is that extra effects penalize AICC more than AIC. As the sample size increases, AICC and AIC converge.

### Average Squared Error

The average squared error (ASE) is the sum of squared errors (SSE) divided by the number of observations. Smaller values are preferred.

### F Value for Model

The value of the F test in a one-way ANOVA after the variances are normalized by the degrees of freedom. Larger values are better, but can indicate overfitting.

### Mean Square Error

The mean squared error (MSE) is the SSE divided by the degrees of freedom for error. The degrees of freedom for error is the number of cases minus the number of weights in the model. This process yields an unbiased estimate of the population noise variance under the usual assumptions. Smaller values are preferred.

### Observations

The number of observations used in the model.

**Pr > F**

The  $p$ -value associated with the corresponding F statistic. Smaller values are preferred.

**R-Square**

The R-squared value is an indicator of how well the model fits the data. R-squared values can range from 0 to 1. Values closer to 1 are preferred.

**Root MSE**

Square root of the MSE.

**SBC**

The Schwarz's Bayesian Criterion (SBC), also known as the Bayesian Information Criterion (BIC), is an increasing function of the model's residual sum of squares and the number of effects. Unexplained variations in the response variable and the number of effects increase the value of the SBC. As a result, a lower SBC implies either fewer explanatory variables, better fit, or both. SBC penalizes free parameters more strongly than AIC.

**Summary Table**

When you click **Show Summary Table** at the top of the model pane, the summary panel is displayed at the bottom of the model pane. The summary table contains the following information:

**Overall ANOVA**

The analysis of variance results for the model, error, and corrected total.

**Dimensions**

An overview of the effect variables used in the model. This tab identifies how many measures and classification effects were chosen for the model, the rank of the cross-product matrix, how many observations were read, and how many observations were used in the model.

**Fit Statistics**

Lists all of the fit statistics described in the previous section.

## **Model ANOVA**

The analysis of variance results for the model.

## **Type III Test**

Provides details for the Type III test. A Type III test examines the significance of each partial effect with all other effects in the model. For more information, see the chapter “The Four Types of Estimable Functions,” in the *SAS/STAT User’s Guide*.

## **Parameter Estimates**

Gives the estimated values for the model parameters.

5

# Logistic Regression Model

- Overview of the Logistic Regression Model* ..... 45
- Logistic Regression Model Properties* ..... 46
- Logistic Regression Model Results Windows* ..... 47
  - The Fit Summary Window ..... 47
  - Residual Plot ..... 49
  - Assessment ..... 52
  - Influence Plot ..... 55
  - Fit Statistics ..... 56
  - Summary Table ..... 57

## Overview of the Logistic Regression Model

A logistic regression attempts to predict the value of a binary response variable. A logistic regression analysis models the natural logarithm of the odds ratio as a linear combination of the explanatory variables. This approach enables the logistic regression model to approximate the probability that an individual observation belongs to the level of interest.

The logistic regression model requires a category response variable and at least one effect variable or interaction term. When your category response variable contains more than two levels, SAS Visual Statistics prompts you to select the level of interest. That is,

SAS Visual Statistics treats all observations in the level of interest as an event and all other observations as nonevents.

---

## Logistic Regression Model Properties

The following properties are available for the logistic regression model:

### Name

enables you to specify the name for this model.

### Informative missingness

specifies whether the informative missingness algorithm is used. For more information, see [Missing Values on page 28](#).

### Use variable selection

specifies whether variable selection is performed. For more information, see [Variable Selection on page 28](#).

### Significance level

specifies the significance level that is required in order for variables to be considered for the model. This property is available only when **Use variable selection** is selected.

### Convergence

- **Override function convergence** enables you to manually specify the function convergence value.
- **Value** specifies the function convergence value when **Override function convergence** is selected. When you specify a larger value, the model will converge sooner. This reduces the amount of time spent training the model, but it can create a suboptimal model.
- **Override gradient convergence** enable you to manually specify the gradient convergence value.
- **Value** specifies the gradient convergence value when **Override gradient convergence** is selected. When you specify a larger value, the model will

converge sooner. This reduces the amount of time spent training the model, but it can create a suboptimal model.

- **Maximum iterations** specifies the maximum number of iterations performed during model training. If you specify a relatively small value, you reduce the amount of time spent training the model, but it can create a suboptimal model.

### Assessment

- **Use default number of bins** specifies whether you want to use the default number of bins or to set your own value. By default, measure variables are grouped into 20 bins.
- **Number** specifies the number of bins to use when the **Use default number of bins** property is not selected. You must specify an integer value between 5 and 100.
- **Prediction cutoff** specifies the value at which a computed probability is considered an event.
- **Tolerance** specifies the tolerance value that is used to determine the convergence of the iterative algorithm that estimates the percentiles. Specify a smaller value to increase the algorithmic precision.

### Show diagnostic plots

specifies whether the Residual Plot, Assessment, and Influence Plot windows appear in the model pane.

---

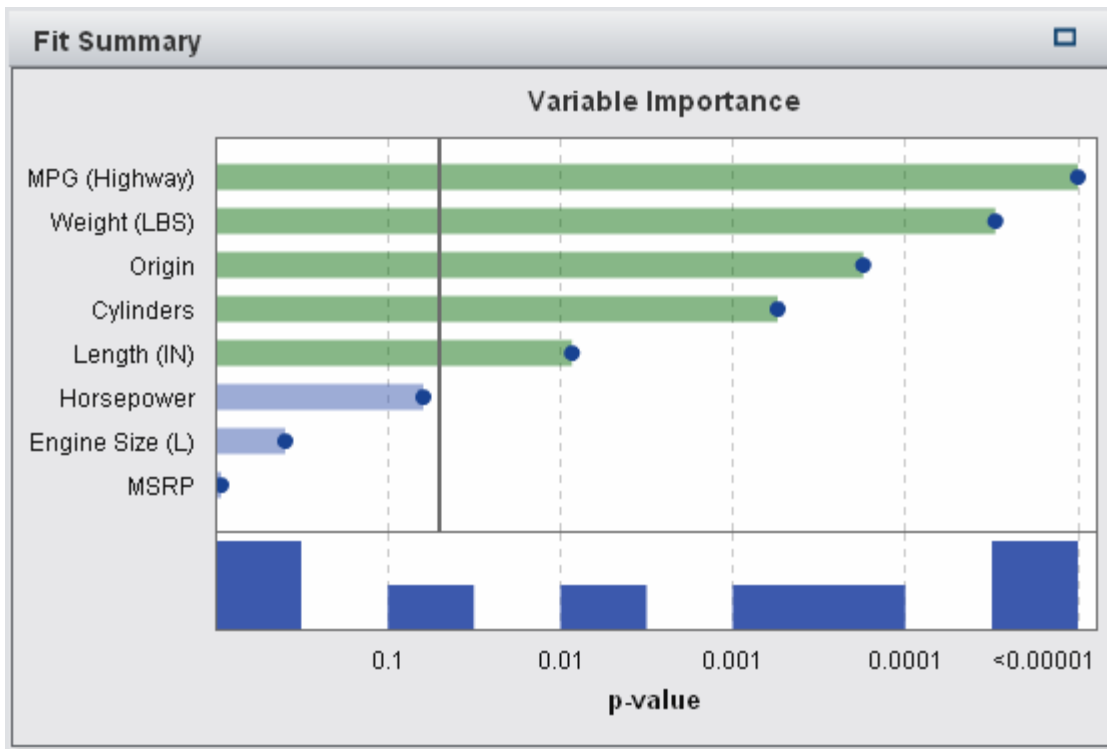
## Logistic Regression Model Results Windows

### The Fit Summary Window

#### Without a Group By Variable

The Fit Summary window plots the relative importance of each variable as measured by its  $p$ -value. The  $p$ -value is plotted on a log scale and the alpha value (plotted as  $-\log(\alpha)$ ) is shown as a vertical line. To adjust the alpha value, click, drag, and drop

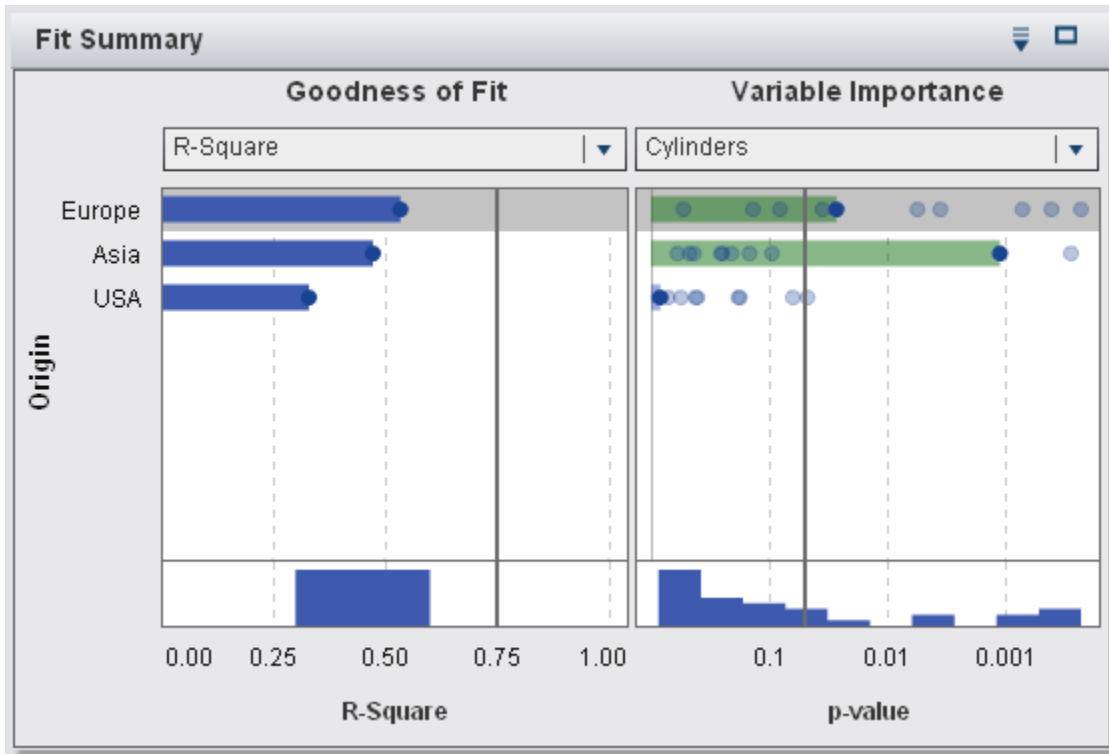
the vertical line. A histogram of the  $p$ -values is displayed at the bottom of the window. A typical Fit Summary window is shown below.



### With a Group By Variable

When your analysis includes a group by variable, the Fit Summary window displays a different set of plots.





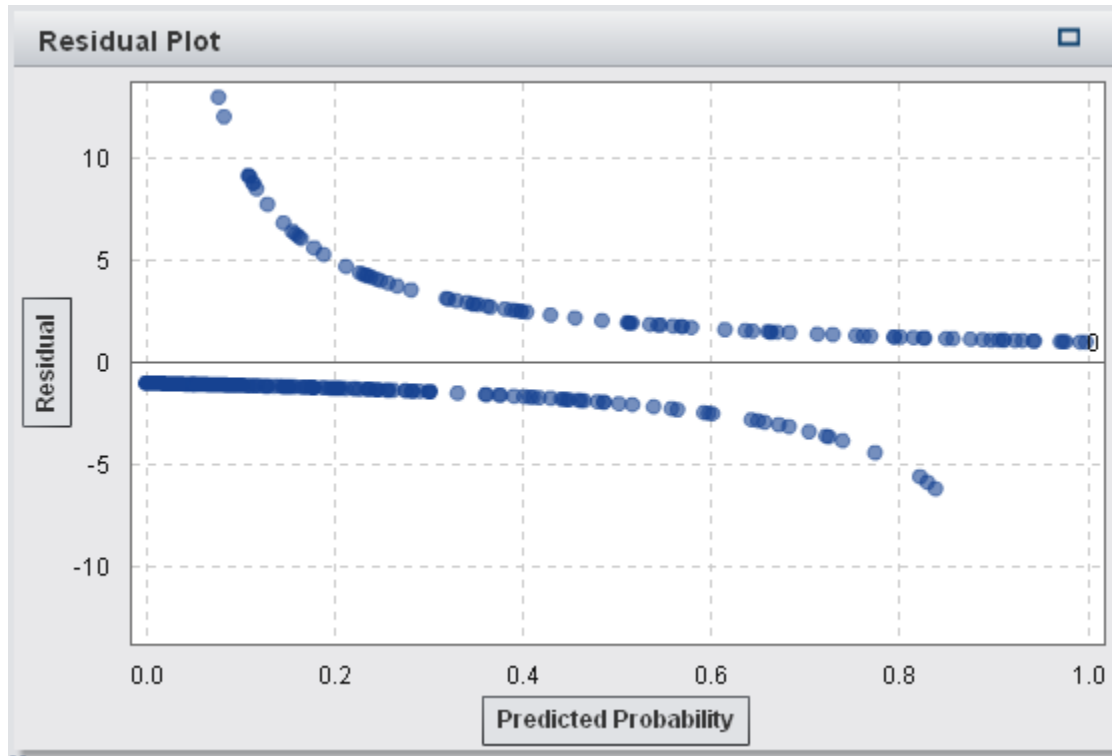
First, notice that the **Variable Importance** plot displays only a single variable. This is because the variable importance for every variable is computed within each level of the group by variable. Use the drop-down menu to view the variable importance for a different effect. Second, notice the **Goodness of Fit** plot, which is not available when there is no group by variable. This plot displays how well the model predicts the response variable within each level of the group by variable. Use this plot to determine whether your model has a significantly different fit within different levels.

Use the ▾ icon to specify how the plot is sorted.

## Residual Plot

### Scatter Plot

The residual of an observation is the difference between the predicted response value and the actual response value. By default, the Residual Plot displays a scatter plot of the residuals against the predicted probability, as shown below.



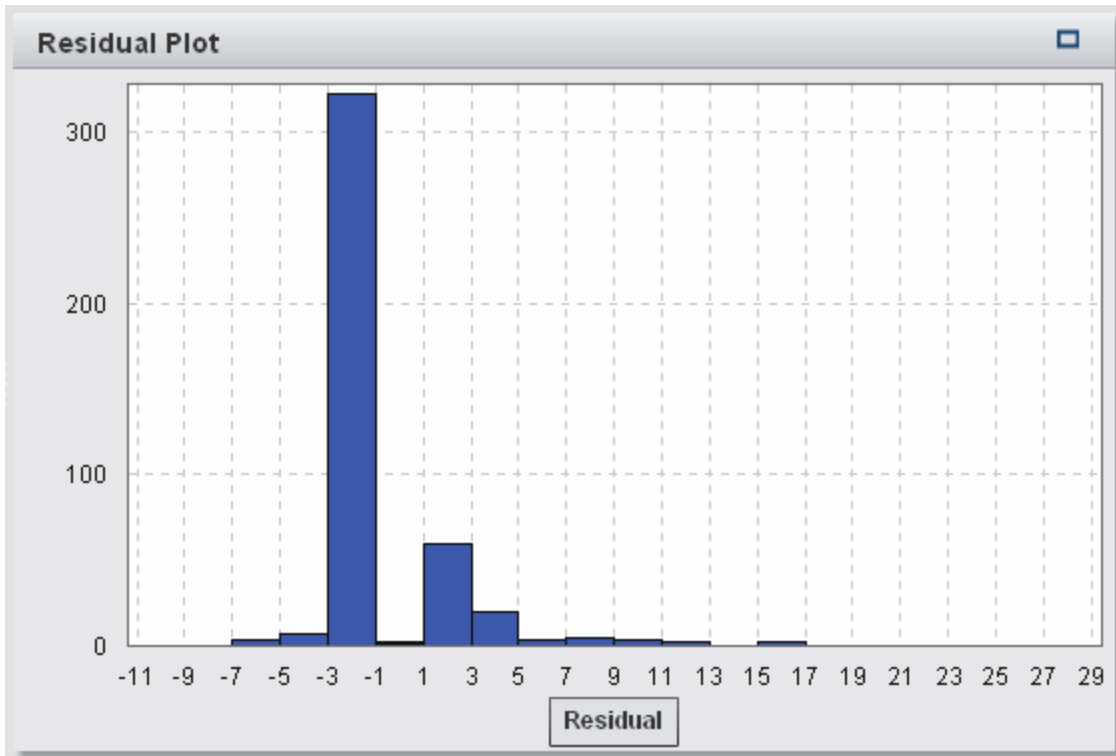
Notice that the labels on both the Y axis and X axis are buttons. You can click either of these buttons to change the value that is plotted on that axis. For the Y axis, you can plot the Residual, Pearson Residual, Deviance Residual, and the Standardized Pearson Residual. The X axis enables you to plot against the predicted probability and the linear predictor. The Standardized Pearson Residual is the individual contribution to the Pearson chi-square test.

Residual plots have several uses when examining your model. First, obvious patterns in the residual plot indicate that the model might not fit the data. Second, residual plots can detect nonconstant variance in the input data when you plot the residuals against the predicted probabilities. Nonconstant variance is evident when the relative spread of the residual values changes as the predicted probability changes. Third, in combination with other methods, the residual plot can help identify outliers in your data.

When using very large data sets, the residual plots are displayed as heat maps instead of actual plots. In a heat map, the actual observations are binned, and the color of each point indicates the relative number of observations in that bin.

## Histogram

To view the data in the residual plot as a histogram, right-click in the Residual Plot window, and select **Use Histogram**. Each of the four values that were available on the Y axis of the residual plot are available as a histogram.



Click the label on the X axis to change the value that is plotted. You can select the Residual, Pearson Residual, Deviance Residual, or Standardized Pearson Residual.

From the histogram, it is fairly easy to determine whether the distribution of the residuals is approximately normal or skewed. For example, in the previous image, the residuals are skewed right. A non-normal residual histogram can indicate that the model does not fit the data.

## Box Plot

To view a box plot of the currently plotted residual, right-click in the Residual Plot window, and select **Plot By**. Then, select a category variable, which is used to group the residuals when the box plots are created. All category variables are available,

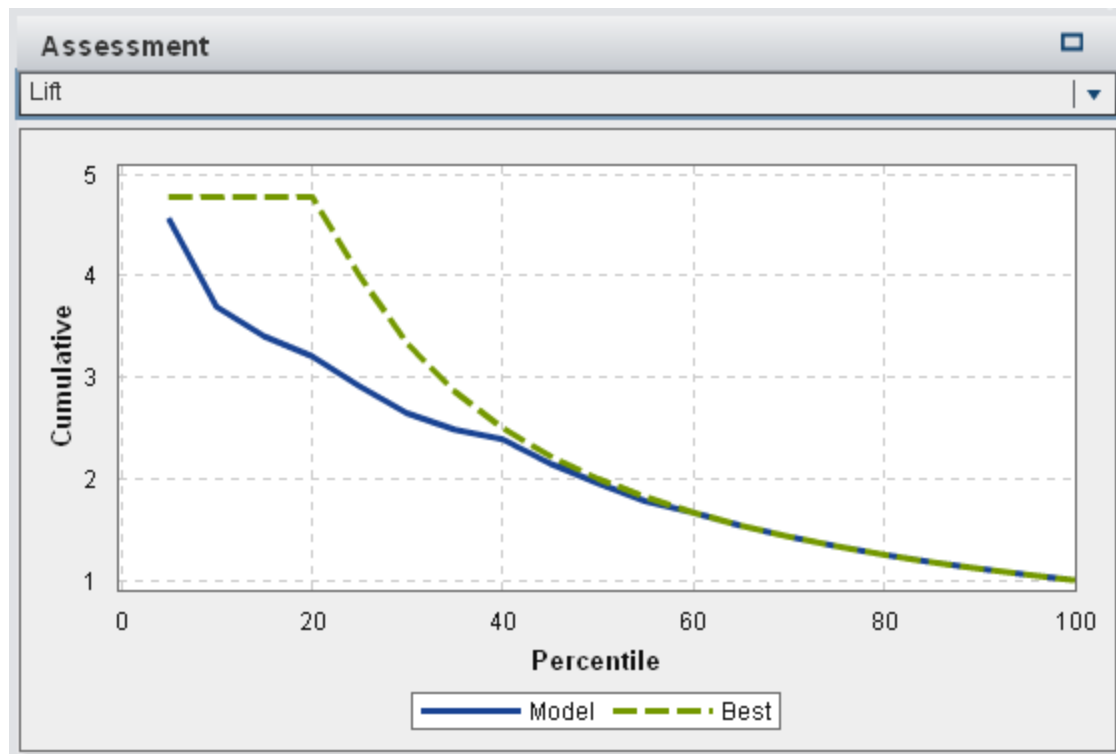
regardless of whether they are included in the model. For variables not included in the model, you can right-click in the Box Plot window, and assign each of those variables as either a classification effect or a group by variable. The **Assign to Classification** and **Assign to Group By** menu options are unavailable for variables that are already included in the model.

Outliers are hidden by default. To show outliers, right-click in the Box Plot window, and select **Show Outliers**.

## Assessment

### Lift

*Lift* is the ratio of the percent of captured responses within each percentile bin to the average percent of responses for the model. Similarly, *cumulative lift* is calculated by using all of the data up to and including the current percentile bin. The default lift chart displays the cumulative lift of the model. Right-click in the lift chart, and select **Switch Lift Type** to alternate between cumulative and noncumulative lift.

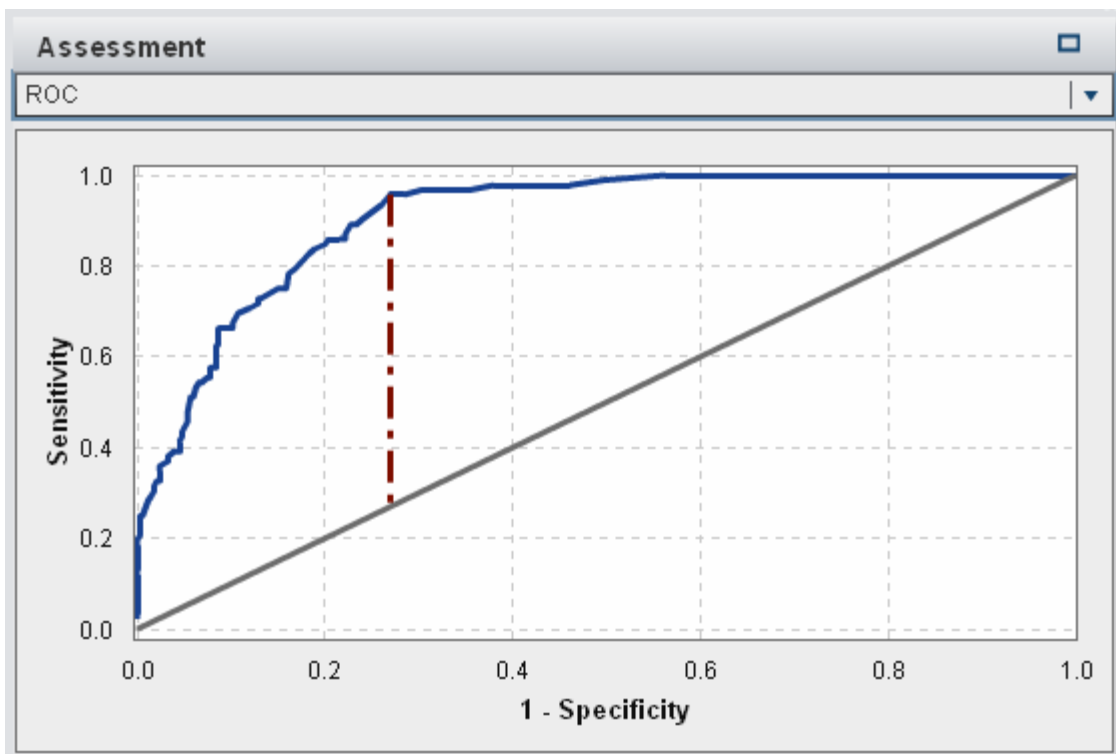


For comparison, the lift chart plots a best model based on complete knowledge of the input data.

## ROC

A receiver operating characteristic (ROC) chart displays the ability of a model to avoid false positive and false negative classifications. A false positive classification means that an observation has been identified as an event when it is actually a nonevent (also referred to as a Type I error). A false negative classification means that an observation has been identified as a nonevent when it is actually an event (also referred to as a Type II error).

The *specificity* of a model is the true negative rate. To derive the false positive rate, subtract the specificity from 1. The false positive rate, labeled **1 – Specificity**, is the X axis of the ROC chart. The *sensitivity* of a model is the true positive rate. This is the Y axis of the ROC chart. Therefore, the ROC chart plots how the true positive rate changes as the false positive rate changes.

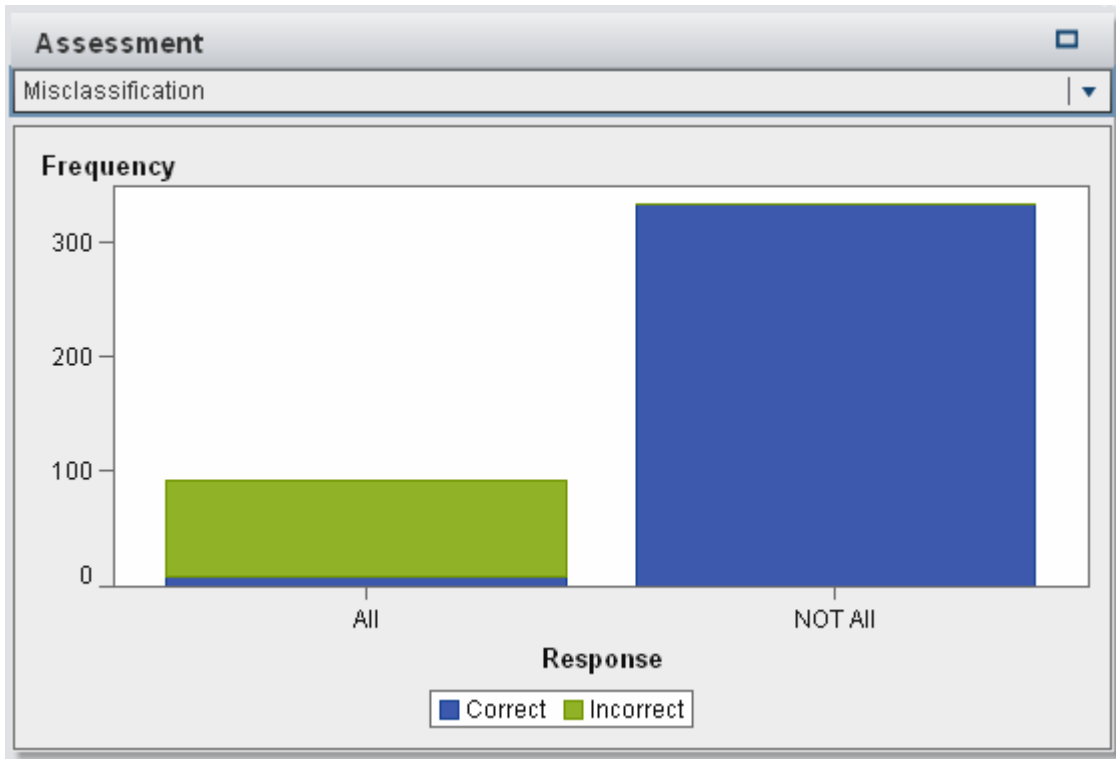


A good ROC chart has a very steep initial slope and levels off quickly. That is, for each misclassification of an observation, significantly more observations are correctly classified. For a perfect model, one with no false positives and no false negatives, the ROC chart would start at (0,0), continue vertically to (0,1), and then horizontally to (1,1). In this instance, the model would correctly classify every observation before a single misclassification could occur.

The ROC chart includes two lines to help you interpret the ROC chart. The first line is a baseline model that has a slope of 1. This line mimics a model that correctly classifies observations at the same rate it incorrectly classifies them. An ideal ROC chart maximizes the distance between the baseline model and the ROC chart. A model that classifies more observations incorrectly than correctly would fall below the baseline model. The second line is a vertical line at the false positive rate where the difference between the Kolmogorov-Smirnov values for the ROC chart and baseline models is maximized.

## **Misclassification**

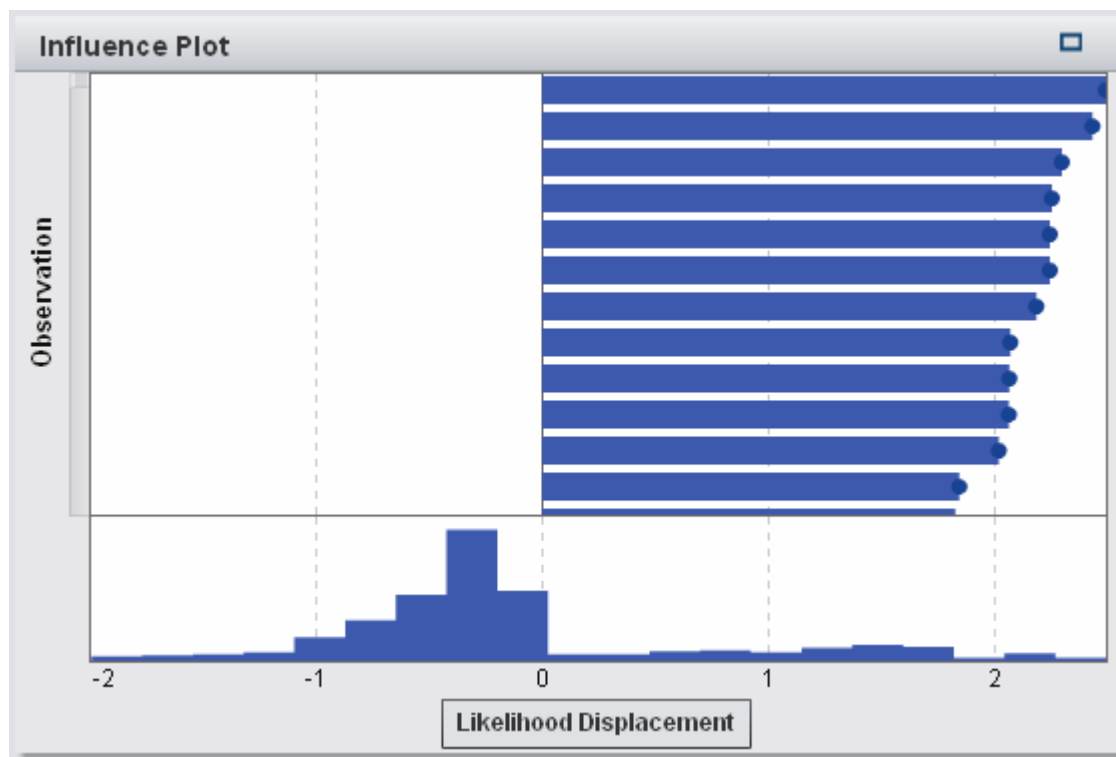
The misclassification plot displays how many observations were correctly and incorrectly classified for each value of the response variable. When the response variable is not binary, as shown below, the logistic regression model considers all levels that are not events as equal.



A significant number of misclassifications might indicate that the model does not fit the data.

## Influence Plot

The Influence Plot displays several measurements that are computed for each observation. When the input data contains a large number of observations, the observations are binned. By default, the **Likelihood Displacement** value is plotted on the X axis.



Other available values are CBAR, Deviance Change, and Pearson Change.

Use these values to help identify outliers and other points that greatly affect the predicted regression model.

## Fit Statistics

The logistic regression model computes several assessment measures to help you evaluate how well the model fits the data. These assessment measures are available at the top of the model pane. Click the currently displayed assessment measure to see all of the available assessment measures.

### -2 Log Likelihood

The likelihood function estimates the probability of an observed sample given all possible parameter values. The log likelihood is simply the logarithm of the likelihood function. The likelihood function value is -2 times the log likelihood. Smaller values are preferred.



**AIC**

Akaike's Information Criterion. Smaller values indicate better models, and AIC values can become negative. AIC is based on the Kullback-Leibler information measure of discrepancy between the true distribution of the response variable and the distribution specified by the model.

**AICC**

Corrected Akaike's Information Criterion. This version of AIC adjusts the value to account for sample size. The result is that extra effects penalize AICC more than AIC. As the sample size increases, AICC and AIC converge.

**BIC**

The Bayesian Information Criterion (BIC), also known as Schwarz's Bayesian Criterion (SBC), is an increasing function of the model's residual sum of squares and the number of effects. Unexplained variations in the response variable and the number of effects increase the value of the BIC. As a result, a lower BIC implies either fewer explanatory variables, better fit, or both. BIC penalizes free parameters more strongly than AIC.

**Max-rescaled R-Square**

The observed R-squared value divided by the maximum attainable R-squared value. This value is useful when there are multiple independent category variables. Values can range from 0 to 1. Values closer to 1 are preferred.

**Observations**

The number of observations used in the model.

**R-Square**

The R-squared value is an indicator of how well the model fits the data. R-squared values can range from 0 to 1. Values closer to 1 are preferred.

**Summary Table**

When you click **Show Summary Table** at the top of the model pane, the summary panel is displayed at the bottom of the model pane. The summary table contains the following information:

## **Dimensions**

An overview of the effect variables used in the model. This tab identifies how many measures and classification effects were chosen for the model, the rank of the cross-product matrix, how many observations were read, and how many observations were used in the model.

## **Iteration History**

The function and gradient convergence results. This tab shows at which iteration the function and gradient converged.

## **Convergence**

Provides the reason for convergence.

## **Fit Statistics**

Lists all of the fit statistics described in the previous section.

## **Type III Test**

Provides details for the Type III test. A Type III test examines the significance of each partial effect with all other effects in the model. For more information, see the chapter “The Four Types of Estimable Functions,” in the *SAS/STAT User’s Guide*.

## **Parameter Estimates**

Gives the estimated values for the model parameters.

## **Response Profile**

Displays the event and nonevent counts.

6

# Generalized Linear Model

- Overview of the Generalized Linear Model* ..... 59
- Generalized Linear Model Properties* ..... 60
- Generalized Linear Model Results Windows* ..... 63
  - The Fit Summary Window ..... 63
  - Residual Plot ..... 65
  - Assessment ..... 67
  - Fit Statistics ..... 68
  - Summary Table ..... 69

## Overview of the Generalized Linear Model

A generalized linear model (GLM) is an extension of a traditional linear model that allows the population mean to depend on a linear predictor through a nonlinear link function. A GLM requires that you specify a distribution and a link function. The distribution should match the distribution of the response variable. The link function is used to relate the response variable to the effect variables.

The GLM requires a measure response variable and at least one effect variable or interaction term. The distribution imposes range requirements on the measure response variable. These requirements are provided in the following table:

Distribution	Range Requirements
Beta	Values must be between 0 and 1, exclusive
Binary	Two distinct values
Exponential	Nonnegative real values
Gamma	Nonnegative real values
Geometric	Positive integers
Inverse Gaussian	Positive real values
Negative Binomial	Nonnegative integers
Normal	Real values
Poisson	Nonnegative integers

# Generalized Linear Model Properties

The following properties are available for the GLM:

**Name**

enables you to specify the name for this model.

**Informative missingness**

specifies whether the informative missingness algorithm is used. For more information, see [Missing Values on page 28](#)

**Distribution**

specifies the distribution used to model the response variable.

## Link function

specifies the link function used to relate the linear model to the distribution of the response variable. Available link functions are different for each distribution and are shown in the following table:

Distribution	Available Link Functions
Beta	Logit, Probit, Log-log, C-log-log
Binary	Logit, Probit, Log-log, C-log-log
Exponential	Log, Identity
Gamma	Log, Identity, Recip
Geometric	Log, Identity
Inverse Gaussian	Power(-2), Log, Identity
Negative Binomial	Log, Identity
Normal	Log, Identity
Poisson	Logarithmic, Identity

## Convergence

- **Override function convergence** enables you to manually specify the function convergence value.
- **Value** specifies the function convergence value when **Override function convergence** is selected. When you specify a larger value, the model will converge sooner. This reduces the amount of time spent training the model, but it can create a suboptimal model.
- **Override gradient convergence** enables you to manually specify the gradient convergence value.
- **Value** specifies the gradient convergence value when **Override gradient convergence** is selected. When you specify a larger value, the model will

converge sooner. This reduces the amount of time spent training the model, but it can create a suboptimal model.

- **Maximum iterations** specifies the maximum number of iterations performed during model training. If you specify a relatively small value, you reduce the amount of time spent training the model, but it can create a suboptimal model.

**Note:** When you specify a gradient convergence or function convergence criterion, it is possible for the model to converge based on an internal convergence criterion before your criterion is reached. The reason for convergence can be found on the **Convergence** tab of the summary table.

### Assessment

- **Use default number of bins** specifies whether you want to use the default number of bins or to set your own value. By default, measure variables are grouped into 20 bins.
- **Number** specifies the number of bins to use when the **Use default number of bins** property is not selected. You must specify an integer value between 5 and 100.
- **Tolerance** specifies the tolerance value that is used to determine the convergence of the iterative algorithm that estimates the percentiles. Specify a smaller value to increase the algorithmic precision.

### Show diagnostic plots

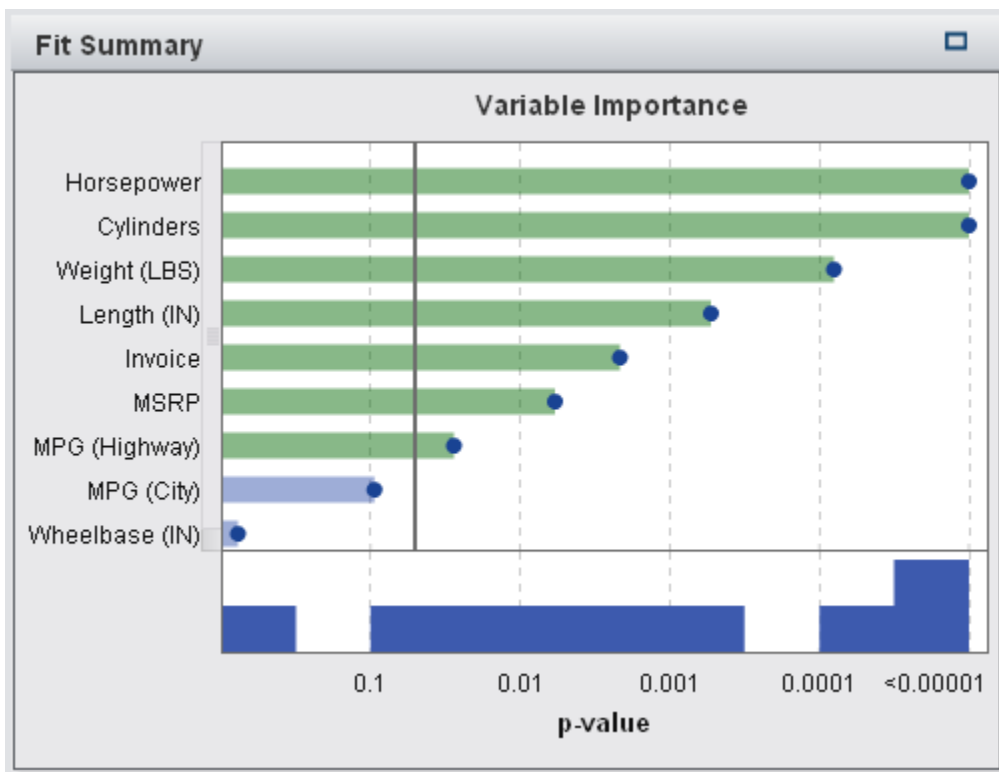
specifies whether the Residual Plot and Assessment windows appear in the model pane.

# Generalized Linear Model Results Windows

## The Fit Summary Window

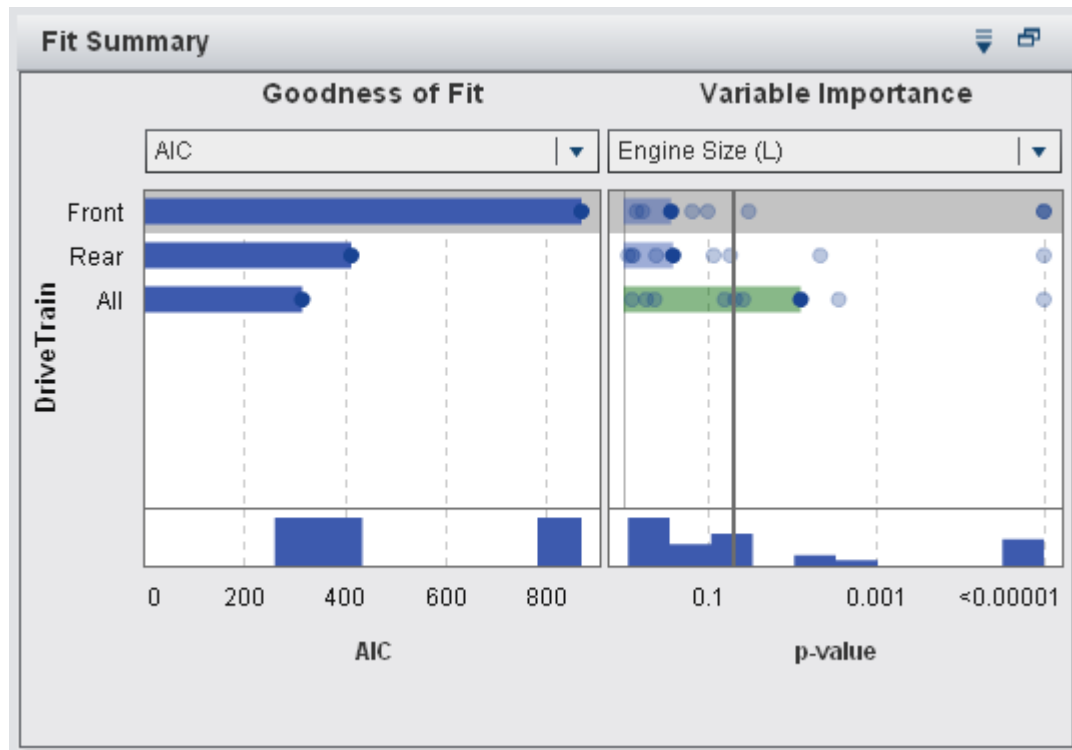
### Without a Group By Variable

The Fit Summary window plots the relative importance of each variable as measured by its  $p$ -value. The  $p$ -value is plotted on a log scale and the alpha value (plotted as  $-\log(\alpha)$ ) is shown as a vertical line. To adjust the alpha value, click and drag the vertical line. A histogram of the  $p$ -values is displayed at the bottom of the window. A typical Fit Summary window is shown below.



## With a Group By Variable

When your analysis includes a group by variable, the Fit Summary window displays a different set of plots.



First, notice that the **Variable Importance** plot displays only a single variable. This is because the variable importance for every variable is computed within each level of the group by variable. Use the drop-down menu to view the variable importance for a different effect. Second, notice the **Goodness of Fit** plot, which is not available when there is no group by variable. This plot displays how well the model predicts the response variable within each level of the group by variable. Use this plot to determine whether your model has a significantly different fit within different levels.

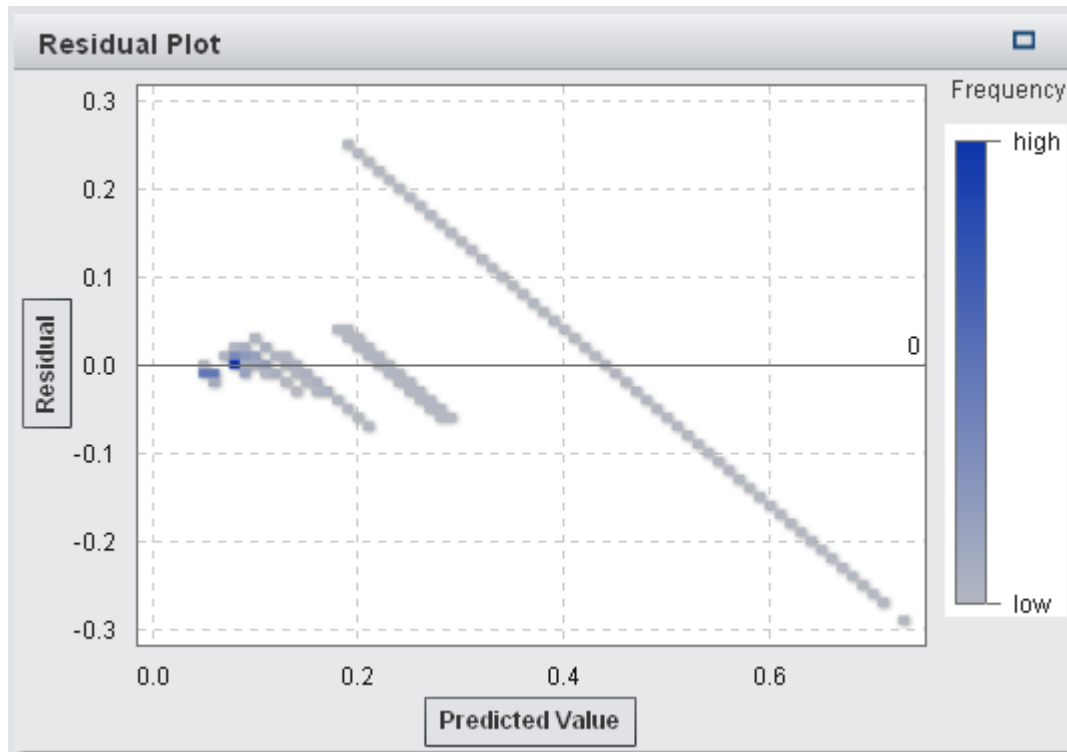
Use the  icon to specify how the plot is sorted.



## Residual Plot

### Scatter Plot

The residual of an observation is the difference between the predicted response value and the actual response value. By default, the Residual Plot displays a scatter plot of the residuals against the predicted value, as shown below.



Notice that the labels on both the Y axis and X axis are buttons. You can click either of these buttons to change the value that is plotted on that axis. For the Y axis, you can plot the Residual or Standardized Pearson Residual. The Standardized Pearson Residual is the individual contribution to the Pearson chi-square test. For the X axis, you can plot the predicted value or the linear predictor.

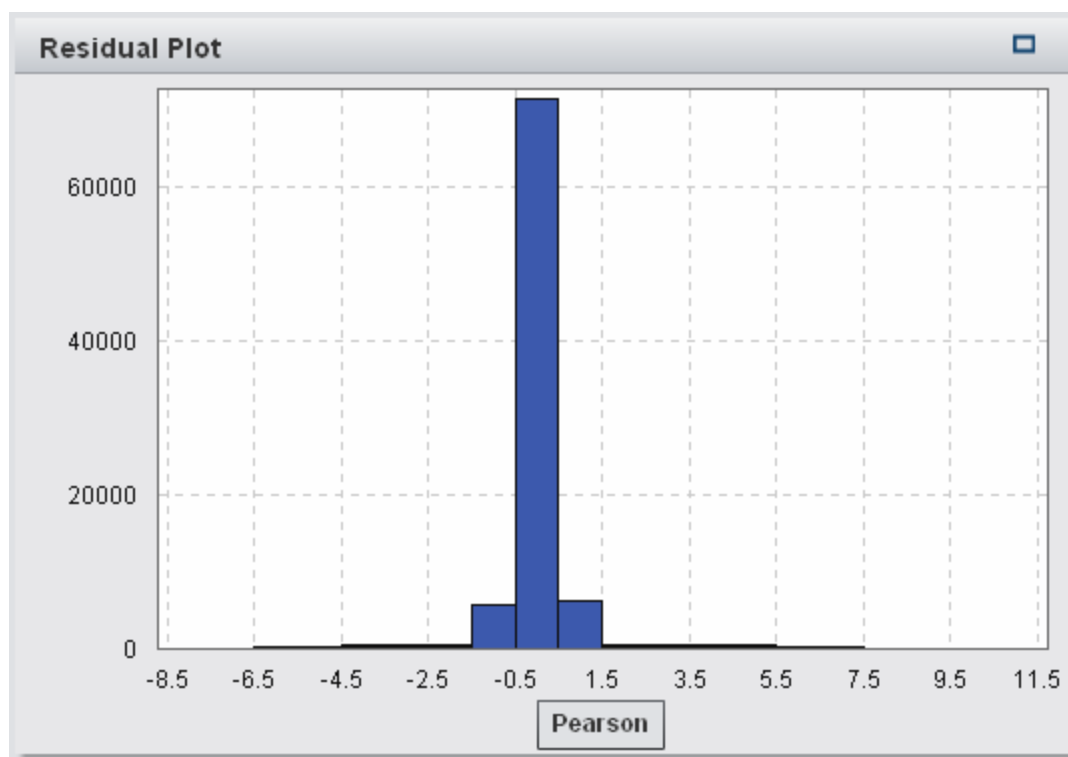
Residual plots have several uses when examining your model. First, obvious patterns in the residual plot indicate that the model might not fit the data. Second, residual plots can detect nonconstant variance in the input data when you plot the residuals against

the predicted values. Nonconstant variance is evident when the relative spread of the residual values changes as the predicted values change. Third, in combination with other methods, the residual plot can help identify outliers in your data.

When using very large data sets, the residual plots are displayed as heat maps instead of actual plots. In a heat map, the actual observations are binned, and the color of each point indicates the relative number of observations in that bin. A heat map is shown above.

## Histogram

To view the data in the residual plot as a histogram, right-click in the Residual Plot window, and select **Use Histogram**. Each of the values that were available on the Y axis of the residual plot are available as a histogram.



Click the label on the X axis to change the value that is plotted. You can select the Residual or Standardized Pearson Residual.

From the histogram, it is fairly easy to determine whether the distribution of the residuals is approximately normal or skewed. For example, in the previous image, the standardized Pearson residuals are symmetric. A non-normal residual histogram can indicate that the model does not fit the data.

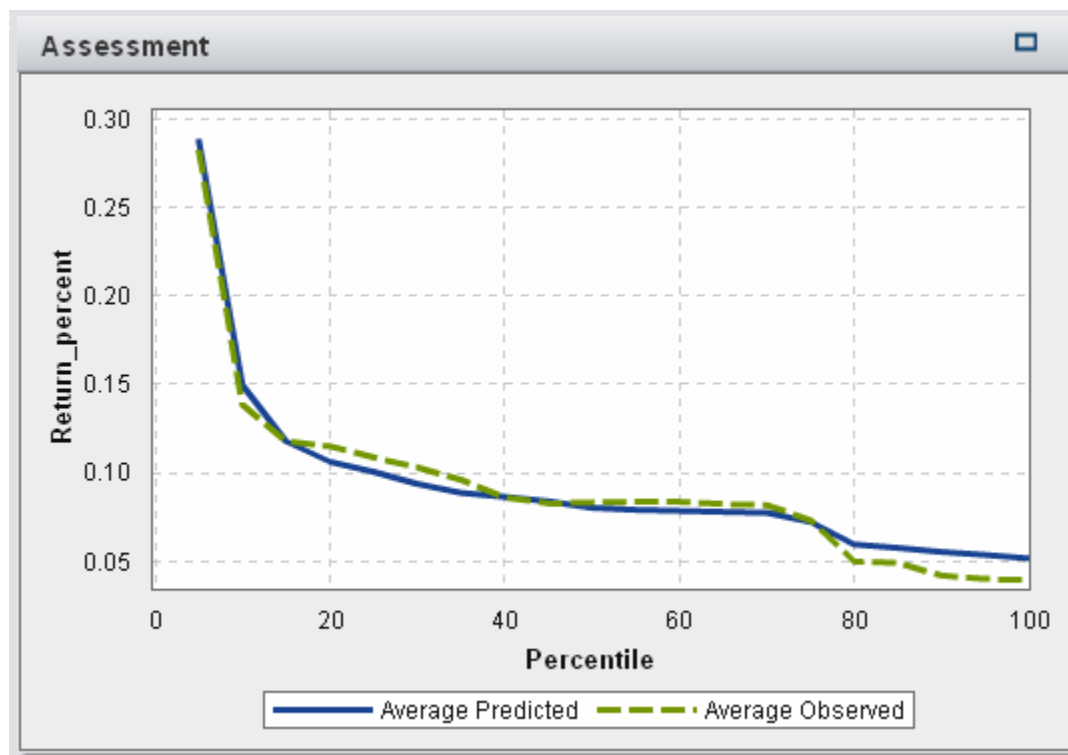
## Box Plot

To view a box plot of the currently plotted residual, right-click in the Residual Plot window, and select **Plot By**. Then, select a category variable, which is used to group the residuals when the box plots are created. All category variables are available, regardless of whether they are included in the model. For variables not included in the model, you can right-click in the Box Plot window, and assign each of those variables as either a classification effect or a group by variable. The **Assign to Classification** and **Assign to Group By** menu items are unavailable for variables that are already included in the model.

Outliers are hidden by default. To show outliers, right-click in the Box Plot window, and select **Show Outliers**.

## Assessment

For a GLM, the Assessment window plots the average predicted and average observed response values against the binned data. Use this plot to reveal any strong biases in your model. Large differences in the average predicted and average observed values can indicate a bias.



## Fit Statistics

The GLM computes several assessment measures to help you evaluate how well your model fits the data. These assessment measures are available at the top of the Model Pane. Click the currently displayed assessment measure to see all available assessment measures. The available assessment measures are as follows:

### -2 Log Likelihood

The likelihood function estimates the probability of an observed sample given all possible parameter values. The log likelihood is simply the logarithm of the likelihood function. This value is -2 times the log likelihood. Smaller values are preferred.

### AIC

Akaike's Information Criterion. Smaller values indicate better models. AIC values should be compared only when two models have an approximately equal number of observations. AIC values can become negative. AIC is based on the Kullback-

Leibler information measure of discrepancy between the true distribution of the response variable and the distribution specified by the model.

### **AICC**

Corrected Akaike's Information Criterion. This version of AIC adjusts the value to account for a relatively small sample size. The result is that extra effects penalize AICC more than AIC. As the sample size increases, AICC and AIC converge.

### **BIC**

The Bayesian Information Criterion (BIC), also known as Schwarz's Bayesian Criterion (SBC), is an increasing function of the model's residual sum of squares and the number of effects. Unexplained variations in the response variable and the number of effects increase the value of the BIC. As a result, a lower BIC implies either fewer explanatory variables, better fit, or both. BIC penalizes free parameters more strongly than AIC.

### **Observations**

The number of observations used in the model.

## **Summary Table**

When you click **Show Summary Table** at the top of the model pane, the summary panel is displayed at the bottom of the model pane. The summary table contains the following information:

### **Dimensions**

An overview of the effect variables used in the model. This tab identifies how many measures and classification effects were chosen for the model, the rank of the cross-product matrix, how many observations were read, and how many observations were used in the model.

### **Iteration History**

The function and gradient iteration results. This tab shows the value of the objective (likelihood) function, its change in value, and its maximum gradient.

### **Convergence**

Provides the reason for convergence.

### **Fit Statistics**

Lists all of the fit statistics described in the previous section.

### **Type III Test**

Provides details for the Type III test. A Type III test examines the significance of each partial effect with all other effects in the model. For more information, see the chapter “The Four Types of Estimable Functions,” in the *SAS/STAT User’s Guide*.

### **Parameter Estimates**

Gives the estimated values for the model parameters.

7

# Decision Tree

- Overview of the Decision Tree* ..... 71
- Decision Tree Properties* ..... 72
- Information Gain and Gain Ratio Calculations* ..... 74
- Decision Tree Results Windows* ..... 76
  - Tree ..... 76
  - Leaf Statistics ..... 79
  - Assessment ..... 80
  - Summary Table ..... 83

## Overview of the Decision Tree

A decision tree creates a hierarchical segmentation of the input data based on a series of rules applied to each observation. Each rule assigns an observation to a segment based on the value of one effect. Rules are applied sequentially, which results in a hierarchy of segments within segments. The hierarchy is called a tree, and each segment is called a *node*. The original segment contains the entire data set and is called the *root node*. A node and all of its successors form a *branch*. The final nodes are called *leaves*. For each leaf, a decision is made about the response variable and applied to all observations in that leaf. The exact decision depends on the response variable.

The decision tree requires a measure response variable or category response variable and at least one predictor. A predictor can be a category or measure variable, but not an interaction term.

The decision tree enables you to manually train and prune a decision tree by entering interactive mode. In interactive mode, you are unable to modify the response variable or predictors that are being used. You cannot export model score code. To enter interactive mode, you can either start making changes to the decision tree in the Tree window or you can click **Use Interactive Mode** on the **Roles** tab in the right pane. To leave interactive mode, click **Use Non-Interactive Mode** on the **Roles** tab.

**Note:** When you leave interactive mode, you lose all of your changes.

---

## Decision Tree Properties

The following properties are available for the decision tree:

### **Name**

enables you to specify the name for this model.

### **Maximum branches**

specifies the maximum number of branches allowed when splitting a node.

### **Maximum levels**

specifies the maximum depth of the decision tree.

### **Leaf size**

specifies the minimum number of observations allowed in a leaf node.

### **Response bins**

specifies the number of bins used to categorize a measure response variable.

### **Predictor bins**

specifies the number of bins used to categorize a predictor that is a measure variable.



**Pruning**

specifies the aggressiveness of the tree pruning algorithm. A more aggressive algorithm creates a smaller decision tree. Larger values are more aggressive.

**Rapid growth**

enables you to use the information gain ratio and k-means fast search methods for decision tree growth. When disabled, the information gain and greedy search methods are used, which generally produce a larger tree and require more time to create.

**Include missing**

enables you to include observations with missing values. For category variables, a missing value is assigned to its own level. For measure variables, a missing value is assigned to the smallest available machine value (negative infinity).

**Reuse predictors**

allows more than one split in the same branch based on a predictor.

**Frequency**

specifies whether nodes report how many observations they contain or what percentage of the observations they contain.



**Assessment**

- **Use default number of bins** specifies whether you want to use the default number of bins or to set your own value. By default, measure variables are grouped into 20 bins.
- **Number** specifies the number of bins to use when the **Use default number of bins** property is not selected. You must specify an integer value between 5 and 100.
- **Prediction cutoff** specifies the value at which a computed probability is considered an event.
- **Tolerance** specifies the tolerance value that is used to determine the convergence of the iterative algorithm that estimates the percentiles. Specify a smaller value to increase the algorithmic precision.

**Show diagnostic plots**

specifies whether the Leaf Statistics and Assessment windows appear in the model pane.

**Show tree overview**

displays the tree overview. The tree overview enables quick navigation of large decision trees. When you zoom in to view a specific area of the decision tree, the tree overview shows the entire decision tree and highlights the area that you are viewing. You can click and drag the highlighted area to change the display of the decision tree. Click the  icon in the upper left corner of the tree overview to view the entire decision tree. Click the  icon in the upper left corner of the tree overview to minimize the tree overview.

---

## Information Gain and Gain Ratio Calculations

When the **Rapid growth** property is enabled, node splits are determined in part by information gain ratio instead of information gain. The information gain and information gain ratio calculations and their benefits and drawbacks are explained in this section. In these explanations, an attribute is considered any specific measurement level of a classification variable or bin of a measure variable.

The information gain method chooses a split based on which attribute provides the greatest information gain. The gain is measured in bits. Although this method provides good results, it favors splitting on variables that have a large number of attributes. The information gain ratio method incorporates the value of a split to determine what proportion of the information gain is actually valuable for that split. The split with the greatest information gain ratio is chosen.

The information gain calculation starts by determining the information of the training data. The information in a response value,  $r$ , is calculated in the following expression:

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

$T$  represents the training data and  $|T|$  is the number of observations. To determine the expected information of the training data, sum this expression for every possible response value:

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

Here,  $n$  is the total number of response values. This value is also referred to as the *entropy* of the training data.

Next, consider a split  $S$  on a variable  $X$  with  $m$  possible attributes. The expected information provided by that split is calculated by the following equation:

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

In this equation,  $T_j$  represents the observations that contain the  $j^{\text{th}}$  attribute.

The information gain of split  $S$  is calculated by the following equation:

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

Information gain ratio attempts to correct the information gain calculation by introducing a split information value. The split information is calculated by the following equation:

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

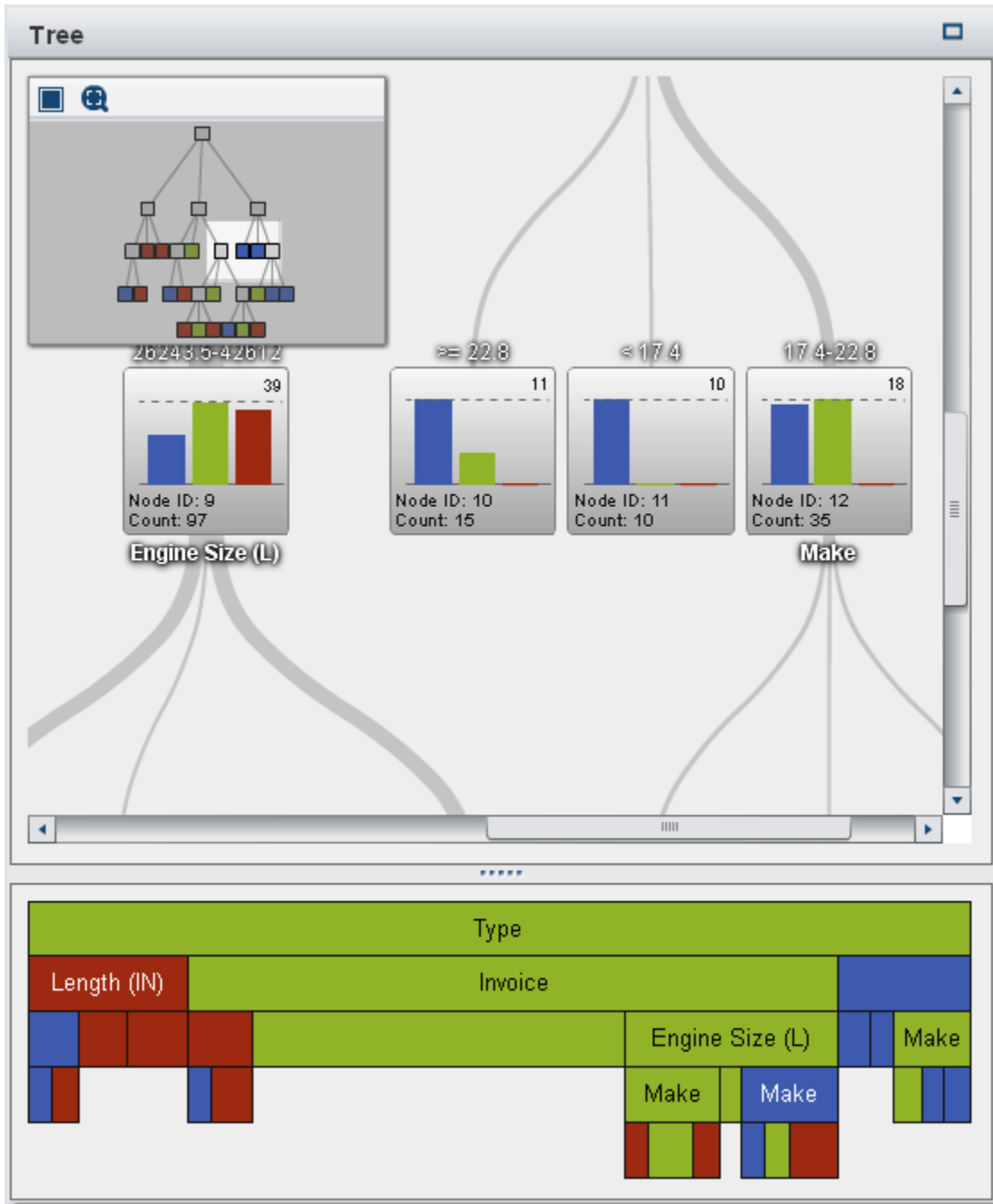
As its name suggests, the information gain ratio is the ratio of the information gain to the split information:

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

## Decision Tree Results Windows

### Tree

The Tree window contains the decision tree, tree overview, and treemap (also called an icicle plot).



In this decision tree, the response variable is a category variable with three levels. The tree overview shows that you are zoomed in to the middle right side of the decision tree.

**TIP** To navigate the decision tree, you can use the mouse and keyboard. Hold down the **Shift** key and click anywhere in the Tree window to move the decision tree within the window. Use your mouse's scroll wheel to zoom in and out of the decision tree. Scroll up to zoom in, and scroll down to zoom out. The zoom is centered on the position of your cursor.

The color of the node in the treemap indicates the predicted level for that node. When you select a node in either the decision tree or the treemap, the corresponding node is selected in the other location. When you select a leaf node, that node is selected in the Leaf Statistics window. A legend is available at the bottom of the model pane.

When the response variable is a measure variable, a gradient is used to denote the predicted bin. Darker colors represent larger values.

Right-click outside of a node in the Tree window to open a pop-up menu. The first item in this menu is **Derive a Leaf ID Variable**. When you click this item, SAS Visual Statistics creates a category variable that contains the leaf ID for each observation. You can use this variable as an effect in other models.

Right-click inside a node to open a different pop-up menu. The available menu options depend on whether you clicked a leaf node.

For leaf nodes, you can select from the following menu options:

### Split

opens the Split Decision Tree window. Use this window to select the variable that is used to split the node. Click **OK** to split the node based on the selected variable. Click **Cancel** to not split the node. Variables are sorted in descending order by their log worth.

### Split Best

splits the node based on the variable with the best information gain ratio when **Rapid growth** is enabled. In addition, splits the node based on the variable with the best information gain when **Rapid growth** is disabled.

### Train

opens the Train Decision Tree window. Use this window to train more than one level beyond the leaf node. First, select every variable that you want to be available for training. Only those variables selected in the Train Decision Tree window are

available for training. Specify the maximum depth of training in the **Maximum depth of subtree** property. Click **OK** to train the decision tree.

For other nodes, select **Prune** to remove all nodes that follow the selected node. This turns the selected node into a leaf node. After pruning a node, you can select **Restore** to undo the prune.

## Leaf Statistics

The Leaf Statistics window plots the percentage of each observation in each leaf node. The most common level in a node is the predicted value assigned to that node. Leaf nodes that contain approximately equal amounts of more than one level might benefit from additional training.

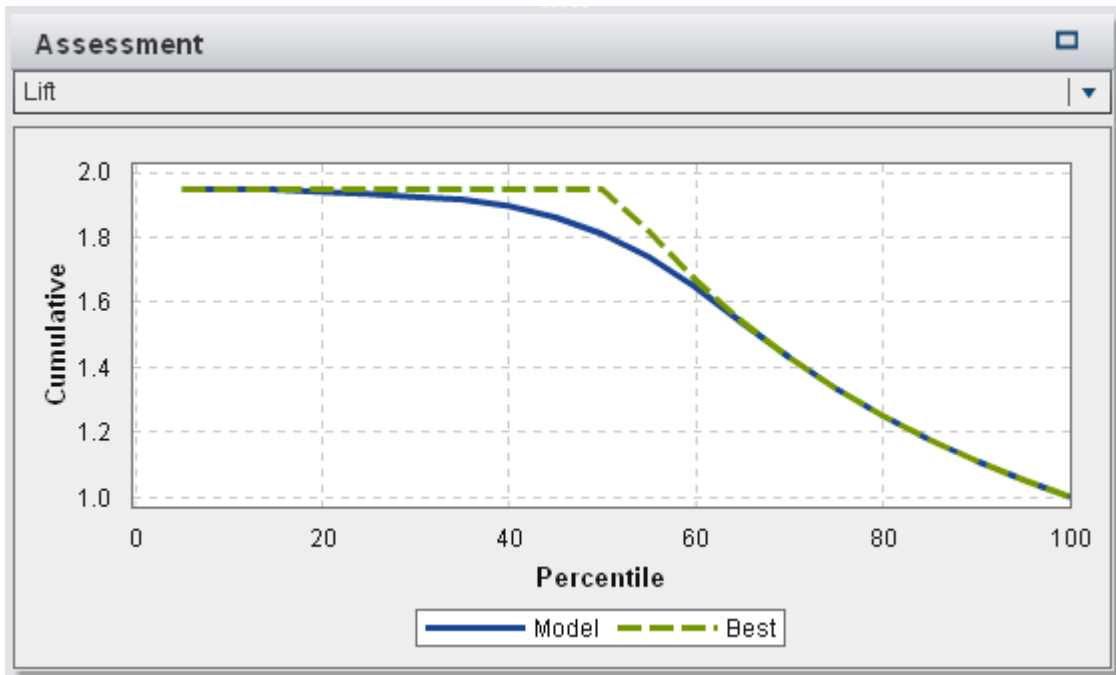


When you select a column in the Leaf Statistics window, the corresponding leaf is selected in the Tree window.

## Assessment

### Lift

*Lift* is the ratio of the percent of captured responses within each percentile bin to the average percent of responses for the model. Similarly, *cumulative lift* is calculated by using all of the data up to and including the current percentile bin.



The lift chart indicates a good fit when the **Model** line is greater than one. For approximately the first 10 to 20 percentiles, your model should provide significant lift when compared to the average response.

The location where the lift chart begins to decrease indicates where the predictive value of the model diminishes. In a decision tree, this location indicates that the next leaf or group of leaves are significantly less powerful.

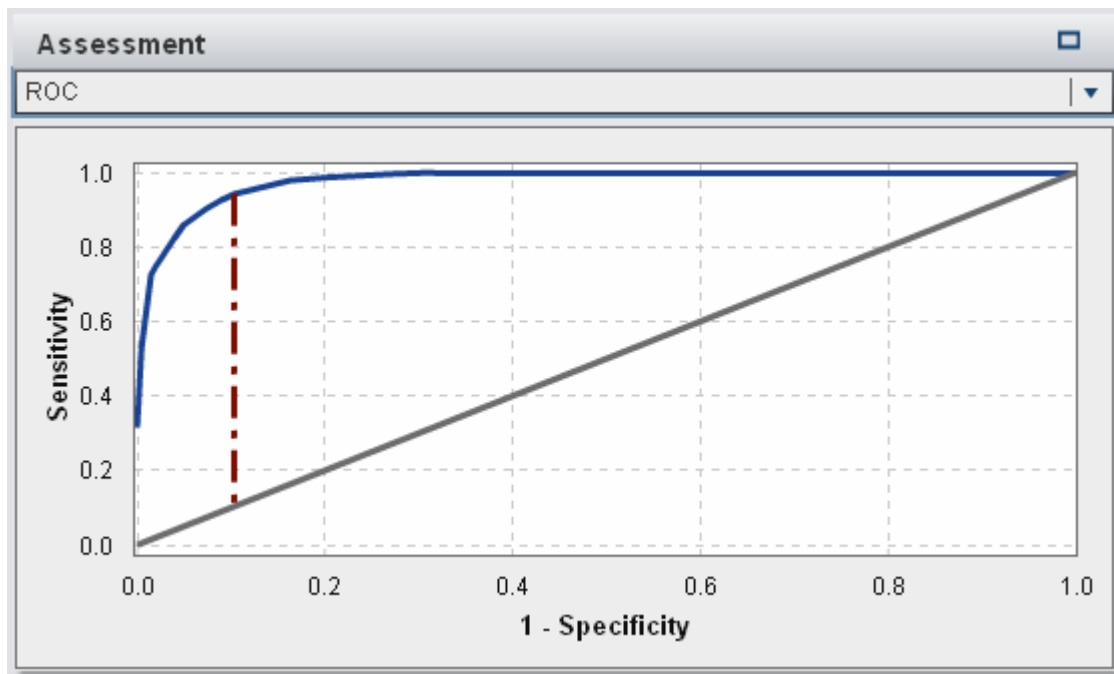
For comparison, the lift chart plots a best model based on complete knowledge of the input data.



## ROC

A receiver operating characteristic (ROC) chart displays the ability of a model to avoid false positive and false negative classifications. A false positive classification means that an observation has been identified as an event when it is actually a nonevent (also referred to as Type I error). A false negative classification means that an observation has been identified as a nonevent when it is actually an event (also referred to as Type II error).

The *specificity* of a model is the true negative rate. To derive the false positive rate, subtract the specificity from 1. The false positive rate, labeled **1 – Specificity**, is the X axis of the ROC chart. The *sensitivity* of a model is the true positive rate. This is the Y axis of the ROC chart. Therefore, the ROC chart plots how the true positive rate changes as the false positive rate changes.



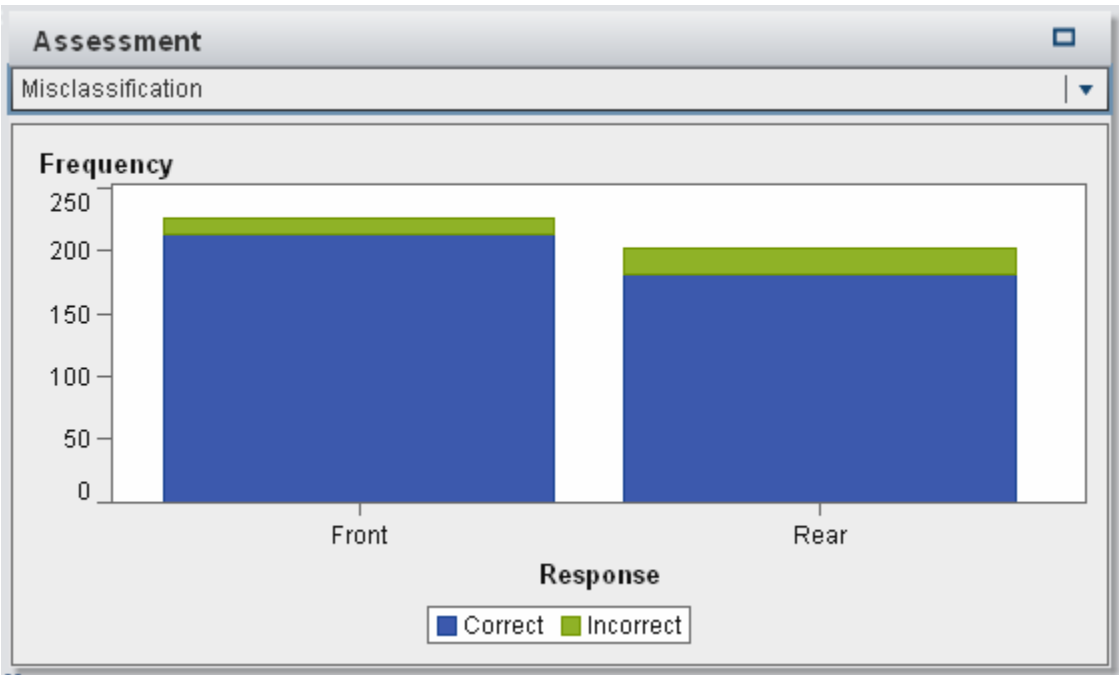
A good ROC chart has a very steep initial slope and levels off quickly. That is, for each misclassification of an observation, significantly more observations are correctly classified. For a perfect model, one with no false positives and no false negatives, the ROC chart would start at (0,0), continue vertically to (0,1), and then horizontally to (1,1).

In this instance, the model would correctly classify every observation before a single misclassification could occur.

The ROC chart includes two lines to help you interpret the ROC chart. The first line is a baseline model that has a slope of 1. This line mimics a model that correctly classifies observations at the same rate it incorrectly classifies them. An ideal ROC chart maximizes the distance between the baseline model and the ROC chart. A model that classifies more observations incorrectly than correctly would fall below the baseline model. The second line is a vertical line at the false positive rate where the difference between the Kolmogorov-Smirnov values for the ROC chart and baseline models is maximized.

**Misclassification**

The misclassification plot displays how many observations were correctly and incorrectly classified.

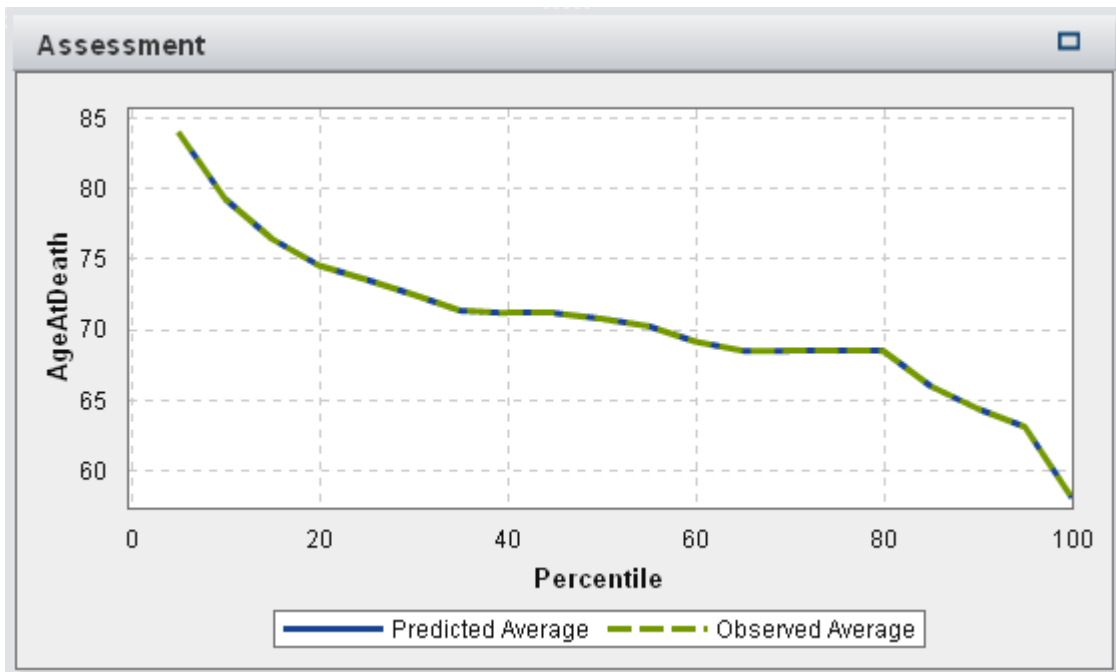


A significant number of misclassifications might indicate that the model does not fit the data.

When the ratio of events to non-events in your data is relatively large, the misclassification plot might show a large number of true positives and false positives. In this case, your model predicts most observations as events and is correct more often than not.

## Predicted Average versus Observed Average

When the number of **Response bins** is set to more than 10, the Assessment window plots the predicted average and observed average values.



## Summary Table

When you click **Show Summary Table** at the top of the model pane, the summary panel is displayed at the bottom of the model pane. The summary table contains the following information:

### Node Statistics

provides summary statistics for each node in the decision tree. Available statistics include **Depth**, **Parent ID**, **N Children**, **Type**, **Observations**, **% Observations**, **N**

**Missing, Gain, Predicted Value, Split**, and the number and percentage of observations in each bin.

**Node Rules**

provides the sorting rule used for each node in the decision tree. Every available variable is listed as a column in the table. If a rule was applied for a variable in a node or any of its parent nodes, then it is listed in the table. Otherwise, the entry is blank.

8

# Cluster

<i>Overview of the Cluster Tool</i> .....	85
<i>Cluster Properties</i> .....	86
<i>Cluster Results Windows</i> .....	87
Cluster Matrix .....	87
Parallel Coordinates .....	89
Summary Table .....	90

## Overview of the Cluster Tool

Clustering is a method of data segmentation that puts observations into groups that are suggested by the data. The observations in each cluster tend to be similar in some measurable way, and observations in different clusters tend to be dissimilar. Observations are assigned to at most one cluster. From the clustering analysis, you can generate a cluster ID variable to use in other tools.

The cluster tool requires at least two measure variables as input. You cannot specify an interaction term or category variable.

---

## Cluster Properties

The following properties are available for the cluster tool:

### Name

enables you to specify the name for this model.

### Cluster Matrix

- **Number of clusters** specifies the number of clusters that are generated.
- **Seed** specifies the seed value of the random number generator that is used during initial cluster assignments.
- **Initial assignment** specifies the method that is used to create the initial cluster assignments. The available methods are:
  - ☐ **Forgy** specifies that  $k$  data points are selected at random to use as the centroids of the  $k$  clusters.
  - ☐ **Random** assigns observations to a cluster at random.
- **Visible roles** determines how many effects are shown in the Cluster Matrix. Valid values are integers between 2 and 6, inclusive.

When you specify a value  $n$ , the first  $n$  effects listed in the **Variables** table on the **Roles** tab are displayed. To change the effect pairs that are plotted in the Cluster Matrix, you can remove an effect from the analysis, and then immediately add it back in. The clustering results remain unchanged because you are using the same input data. However, the **Variables** table adds new effects to the bottom of the list.

- **Variable standardization** transforms the effect variables so that they have a mean of zero and a standard deviation of 1. This property is enabled by default and affects the results displayed in the summary table. The Cluster Matrix window and the Parallel Coordinates window display the original variables.

## Parallel Coordinates

- **Number of bins** specifies the number of bins used when generating the parallel coordinate polyline plots.
- **Maximum polylines** specifies the maximum number of polylines generated by the parallel coordinate algorithm.

## Show ellipses

enables you to display the cluster projection ellipses in the Cluster Matrix.

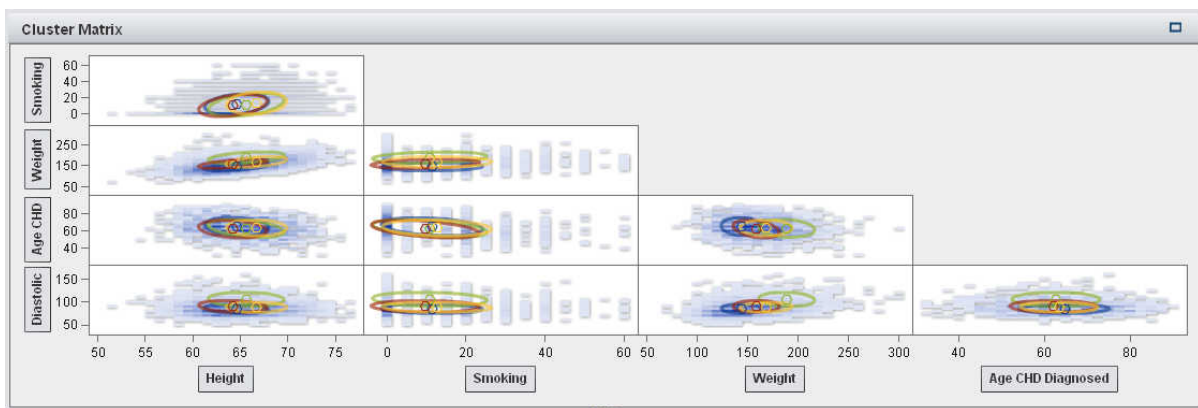
## Show centroids

enables you to display the centroids in the Cluster Matrix.

# Cluster Results Windows

## Cluster Matrix

The Cluster Matrix displays a two-dimensional projection of each cluster onto a specified number of effect pairs. These projections are useful for spotting cluster similarities and differences within the plotted effect pairs.



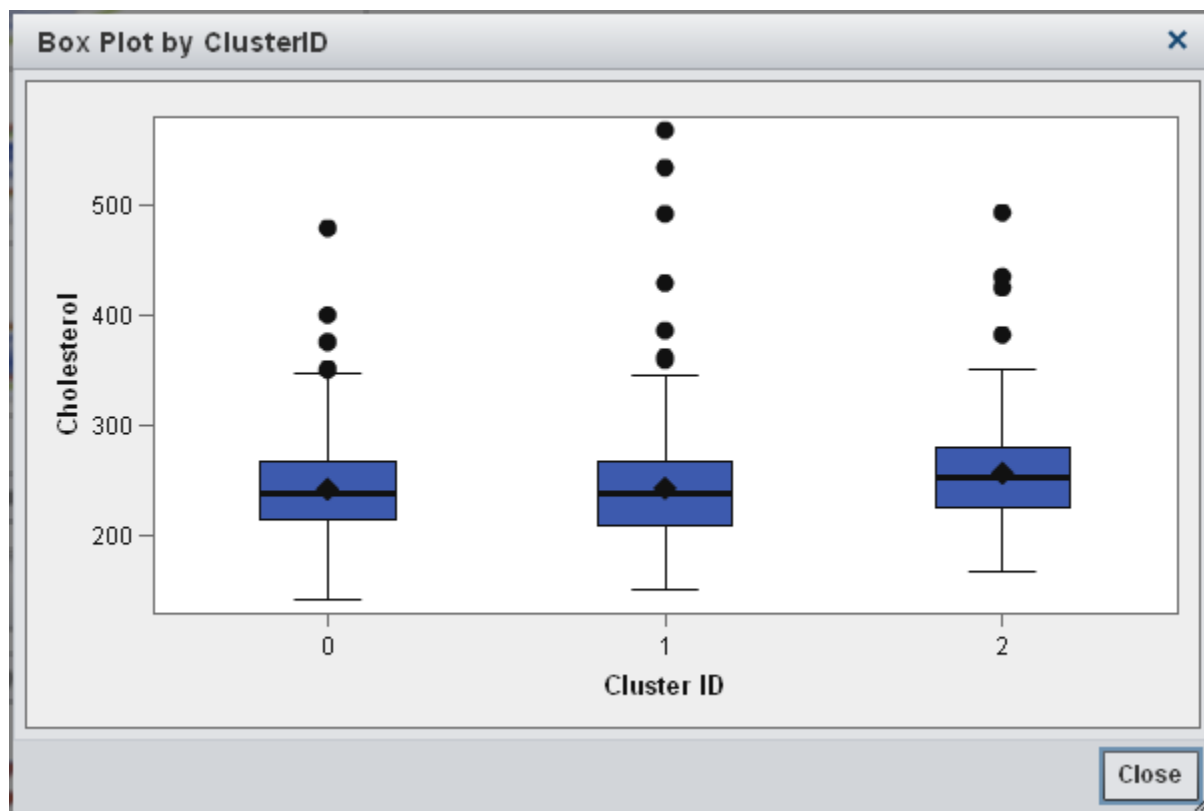
Each cluster is assigned a unique color. Although each cluster is unique in  $n$ -space, the two-dimensional projections will overlap. When a heat map is not used, individual observations are color-coded to indicate cluster membership.

To view a larger plot for an effect pair, right-click inside that plot, and click **Explore**. In the Explore window, it is easier to view and select observations. When you select an observation, the clusters that overlap the selected observation are also selected.

It is important to note that every observation can belong to at most one cluster. However, because the Cluster Matrix displays a projection in just two dimensions, multiple clusters can overlap an observation.

Right-click on a Cluster Matrix plot to open a pop-up menu. The last item in this menu is **Derive a Cluster ID Variable**. When you select this item, SAS Visual Statistics creates a category variable that contains the cluster ID for each observation. You can use this variable as an effect in other models.

You can view a box plot for a variable that segments the observations by cluster. Right-click inside a plot that contains the variable of interest, and select **Plot *variable\_name* by Cluster ID**. Each variable in the selected plot has a menu item.

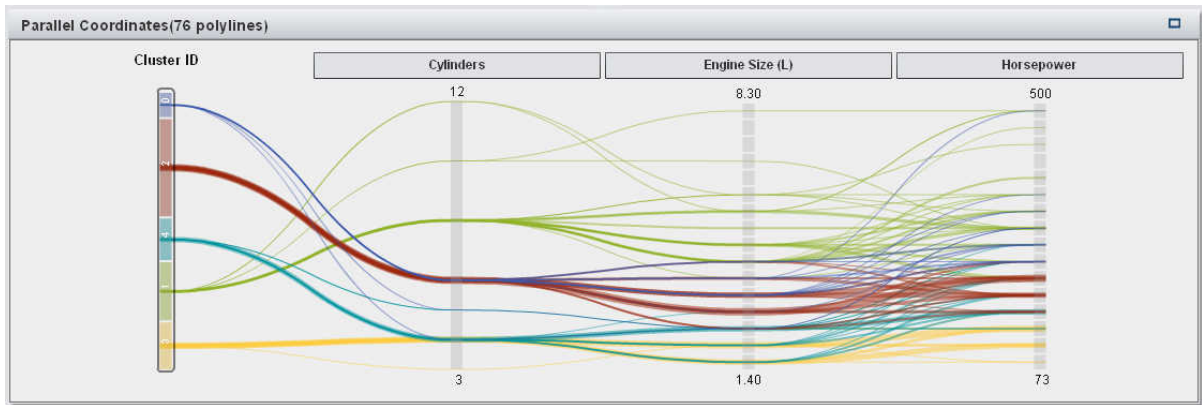




The box plot is used to determine how similar the clusters are for a variable.

## Parallel Coordinates

The Parallel Coordinates plot shows patterns in the data and clusters. In this plot, the cluster ID is on the far left, and each variable is a column with its binned range of values displayed vertically. Color-coded polylines are drawn from each cluster and show which range of values the cluster contains for every variable.



Although it is initially confusing, you can use the Parallel Coordinates plot to make several inferences about the data. You can adjust the plot to explore the data based on cluster membership, a specified range for one or more variables, or both.

When there are several clusters, it can be confusing to determine how each cluster classifies the data. To view just the polylines for a single cluster, select that cluster ID on the far left. Notice that the polylines for all other clusters are grayed out. This enables you to focus on one cluster. Hold down the **Control** key and click on multiple clusters to show only those clusters.

Click on or near a variable name to select that variable. This action changes the color gradient of the polylines so that larger values are darker than smaller values. You can click and drag from the top or bottom of a variable range to adjust the range of values that is shown. You can repeat this step for multiple variables.

Combining these two features, you can restrict the display to specific clusters and variable ranges that interest you.

## Summary Table

When you click **Show Summary Table** at the top of the model pane, the summary panel is displayed at the bottom of the model pane. The summary table contains the following information:

- **Cluster Summary** provides summary statistics for each cluster. Available statistics include **Observations**, **RMS of STD**, **Within-Cluster SS**, **Min centroid-to-observation**, **Max centroid-to-observation**, **Nearest Cluster**, and **Centroid Distance**.

9

# Model Comparison

<i>Overview of Model Comparison</i> .....	91
<i>Model Comparison Usage</i> .....	92
<i>Model Comparison Properties</i> .....	93
<i>Model Comparison Results Windows</i> .....	93
Assessment .....	93
Fit Statistic .....	94
Summary Table .....	95

---

## Overview of Model Comparison

The model comparison tool enables you to compare the performance of competing models using various benchmarking criteria. The comparison criteria available depends on the models and response variable used in your analysis. A model comparison requires that at least one other model is trained before you can perform a comparison.

**Note:** Model comparisons are not saved in SAS Visual Statistics. After closing and reopening a project, if you want to revisit a model comparison, then you must re-create that comparison.


Before performing a model comparison, ensure that all models are initialized and updated. If the **Auto-update model** property is disabled for a model, you must manually

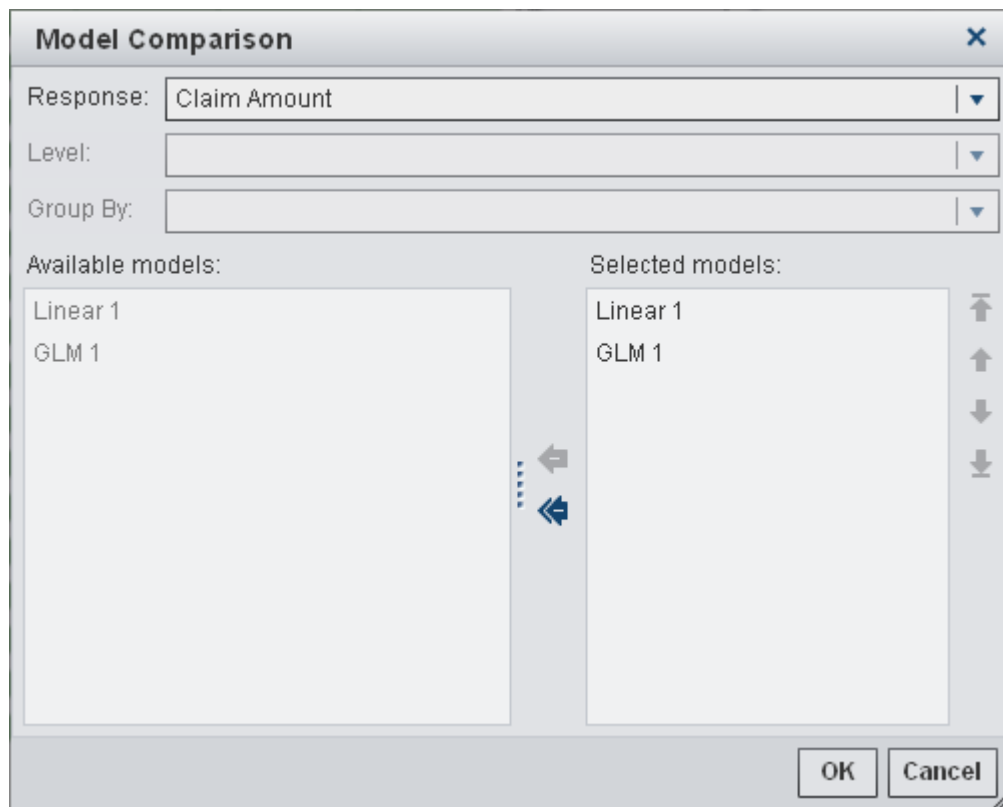
update it before you can compare it to another model. A model is not considered initialized until it has been trained.

When you change a model after a comparison has been created, changes are not carried over to the model comparison.

---

## Model Comparison Usage

When you click the  icon in the toolbar, the Model Comparison window appears.



The Model Comparison window enables you to specify the response variable of interest, the level of interest, a group by variable, and the models for comparison. You must specify a response variable and at least two models.

**Note:** You are able to compare two or more models only when the response variable, level of interest, and group by variable are identical.

---

## Model Comparison Properties

The following properties are available for model comparison:

### **Name**

enables you to specify the name for this comparison.

### **Fit statistic**

specifies the comparison criterion that is plotted in the Fit Statistic window and used to determine the champion model. The fit statistics available depend on the models being compared.

For the error sum of squares (SSE) fit statistic, the linear regression model and logistic regression model use the weighted SSE. The generalized linear model uses the unweighted SSE.

### **Prediction Cutoff**

specifies the cutoff probability that determines whether an observation is a modeled event.

### **Percentile**

when available, specifies the percentile at which the specified fit statistic is plotted.

---

## Model Comparison Results Windows

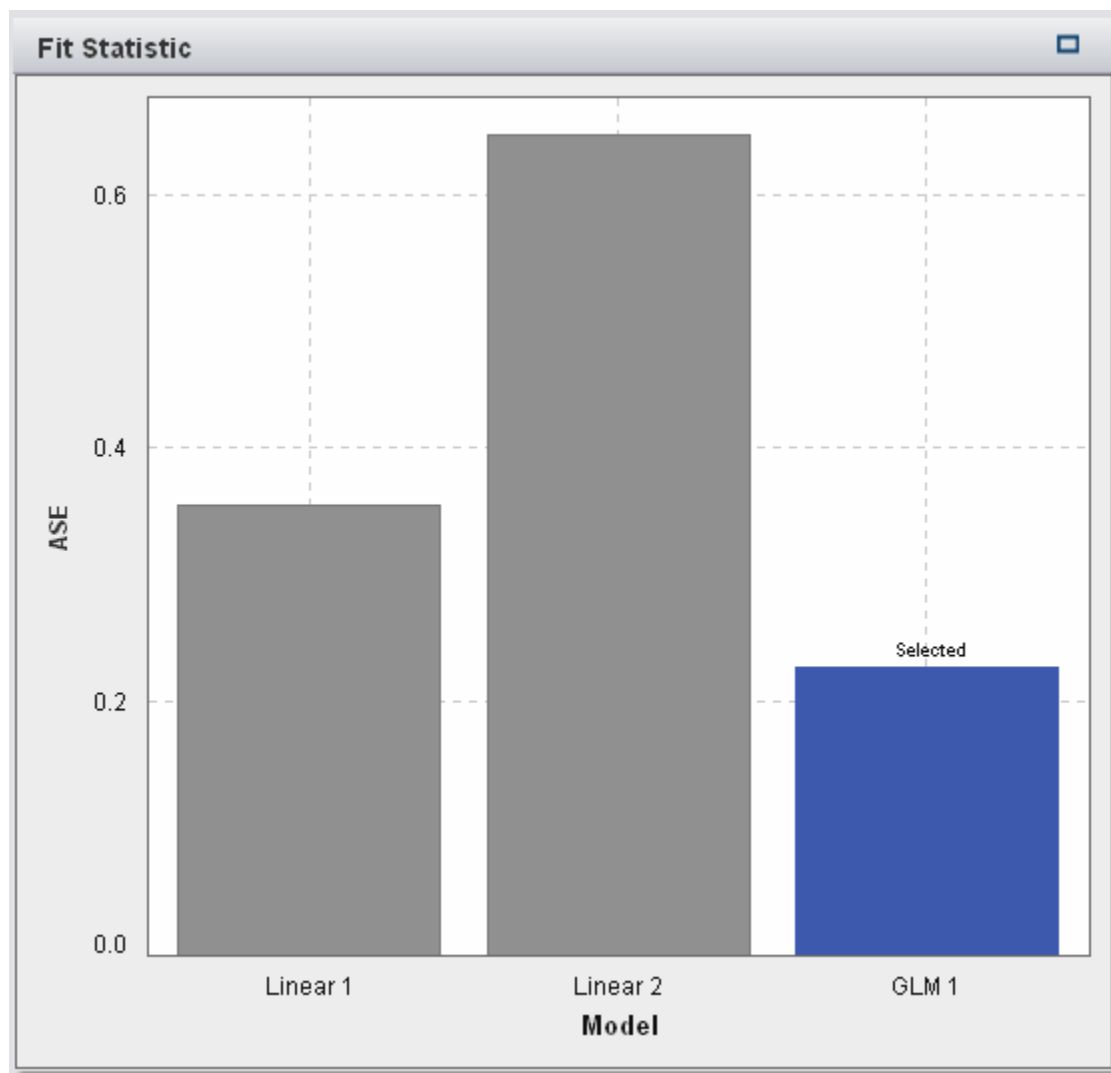
### **Assessment**

The assessment plots available depend on the models being compared. For classification models, the plots displayed are Lift, ROC, and Misclassification. For

numerical models, the plots displayed are observed response value and predicted response value.

## Fit Statistic

The Fit Statistic plot displays the criterion specified in the **Fit statistic** property. In the following image below, the observed average value is plotted for a linear regression and a GLM model. The champion model is indicated in the plot. It is displayed different from the other models.



## Summary Table

When you click **Show Summary Table** at the top of the model pane, the summary panel is displayed at the bottom of the model pane. The summary table contains the following information:

### Statistics

Provides summary statistics for each model in the comparison. The value in the **Selected** column, either **Yes** or **No**, indicates which model the model comparison tool prefers based on the criterion specified in the **Fit statistic** property. However, the statistics listed in the summary table can differ from those listed in the **Fit statistic** property.

### Variable Importance

Indicates which variables had the greatest impact on each of the models in the comparison.





# 10

## SAS Visual Statistics Example

<i>Overview</i> .....	97
<i>Create the Project</i> .....	98
<i>Create a Decision Tree</i> .....	98
<i>Create a Linear Regression</i> .....	100
<i>Create a GLM</i> .....	103
<i>Perform a Model Comparison</i> .....	105

### Overview

This example uses the Framingham Heart Study data set, located in SASHELP.HEART, to compare the performance of a linear regression model and a generalized linear model (GLM). The goal is to predict a person’s age of death based on a collection of health factors. These factors include gender, weight, height, whether the person is a smoker, blood pressure, and more. The focus of this example is how to use SAS Visual Statistics, not how to build the best model.

This example assumes that you have access to the SASHELP.HEART data set. It is beyond the scope of this example to provide instructions about how to access individual data sets at your location. Your system administrator should be able to provide you access to this data set.

---

## Create the Project

This example assumes that you have already signed on to SAS Visual Analytics and you are on the home page. From the home page, click the **Create Analytical Model** icon in the **Create Content** group. This opens SAS Visual Statistics and enables you to open a recent project or to create a new project. Click **Select a Data Source**, located under **Start a new model**, to create a new project.


A window appears that enables you to select the data source for this project. Select the data source that corresponds to SASHELP.HEART. Click **Open**.

By default, the project is named **Project 1**, which is displayed in the upper left corner of SAS Visual Statistics. Before continuing with the example, rename the project by saving it. Click **File ► Save** from the main menu. This opens the Save As window. In the **SAS Folders** pane, navigate to a location where you have Write permission. Typically, you can save your work in **My Folder**. In the **Name** field, enter **Heart Study**, and click **Save**.

By default, a linear regression model is immediately available for usage. You can change the default model type in the Preferences window. However, in this example, a decision tree is created in order to derive a leaf ID variable. The leaf ID variable is then used in the linear regression model and GLM.

---

## Create a Decision Tree

From the toolbar, click the  icon to create a decision tree. From the **Data** pane, drag and drop the **Age at Death** variable into the **Response** field in the right pane. In the **Data** pane, select **Diastolic**, **Weight**, **Height**, **Cholesterol**, **Age CHD Diagnosed**, **Sex**, and **Cause of Death**. Drag and drop these items into the model pane. The decision tree automatically updates.




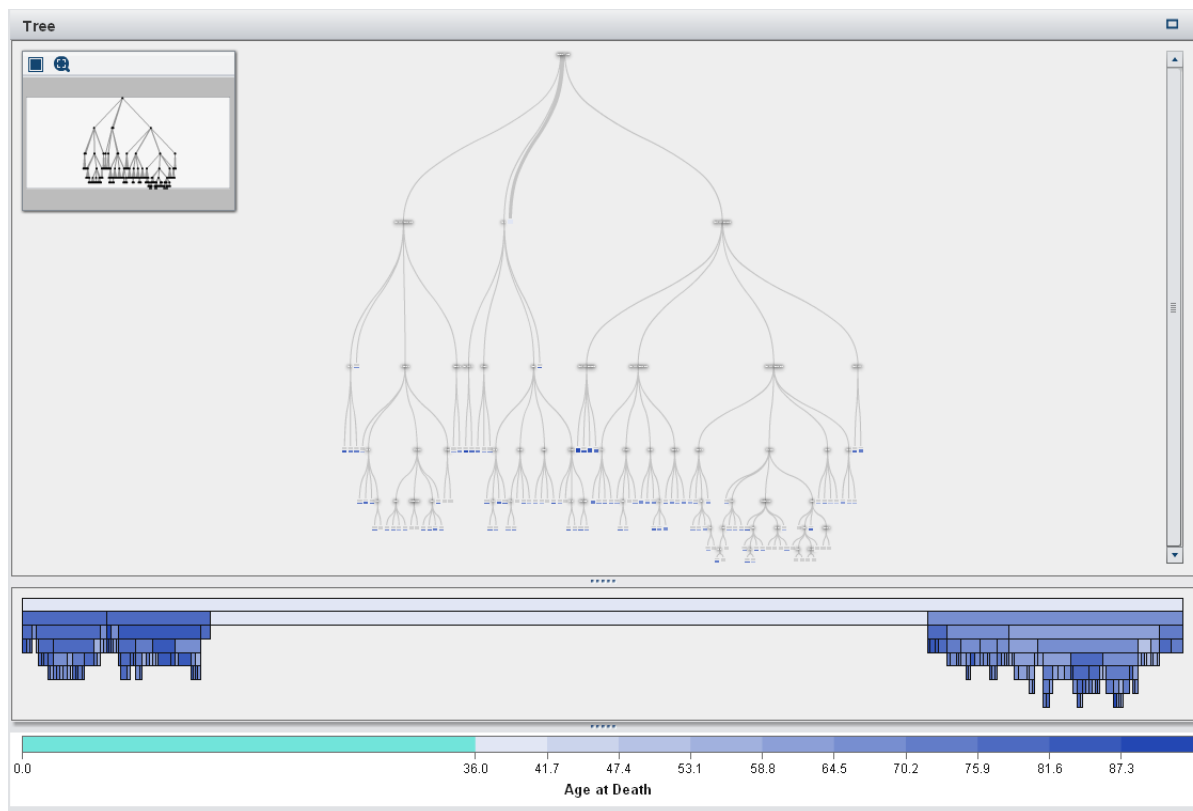
Click **Show Summary Table** in the summary bar. In the summary table, select the **Node Rules** tab. Notice that the only predictors used were **Age CHD Diagnosed** and **Cause of Death**. You can adjust the decision tree properties to include more predictors in the model.

Click the **Properties** tab in the right pane. The most obvious property to change is **Reuse predictors**. When you deselect this property, each predictor variable is used in at most one split. However, assume that reusing predictors creates the best split in each node for this example. This might not always be the case for your data.

Instead, set the value of **Maximum levels** to 10. The decision tree now has a maximum depth of 10 levels, instead of the default 6. On the **Node Rules** tab of the summary table, every predictor is used at least once.

Set the value of **Maximum branches** to 4. This allows each non-leaf node to split into at most four new nodes.

Select the **Show tree overview** property. In the Tree Overview window, click the  icon to fit the entire decision tree into the Tree Overview window. Although each node is difficult to see, your decision tree should resemble the following:



In the Tree window, right-click, and select **Derive a Leaf ID Variable**. A new category variable, named **Leaf ID 1**, appears in the **Data** pane.

Save the project.

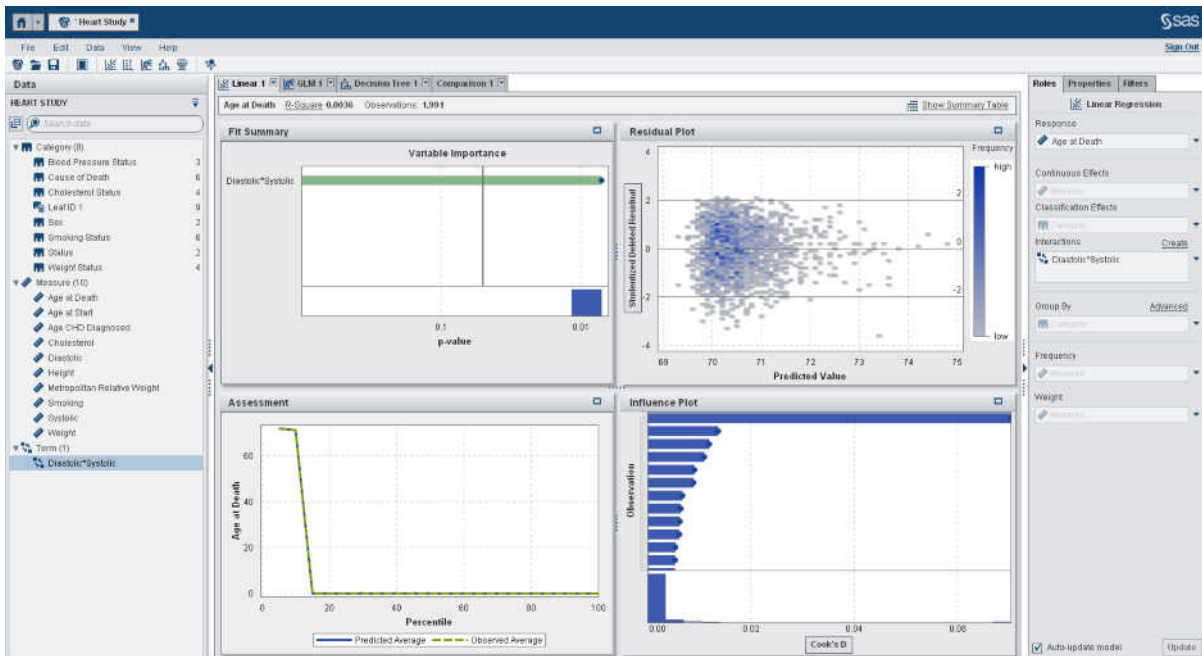
## Create a Linear Regression

In the model pane, select the linear regression model that was created when the project was created.

In this example, the variable of interest is **Age at Death**, which should be the first variable listed in the **Measure** section of the **Data** pane. Because you want this variable to be the response variable, click, drag, and drop **Age at Death** from the **Data** pane onto the model pane. Notice that **Age at Death** now appears in the **Response** field on the **Roles** tab.

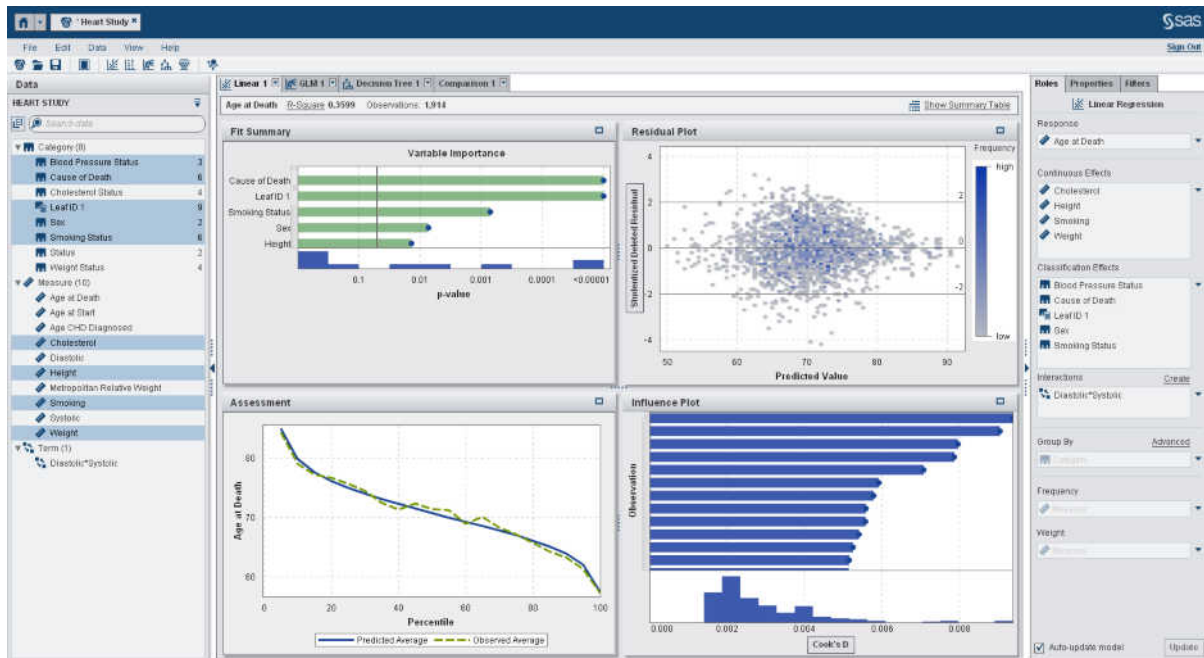
The next step is to choose the effect variables or interaction terms that you want to include in the analysis. One option is to make every available variable an effect variable and let SAS Visual Statistics perform variable selection. However, this is not always feasible from a computational resources perspective. This example creates an interaction term to use as an effect variable and includes a few other variables as effect variables.

Because you suspect that systolic blood pressure and diastolic blood pressure interact with each other, create an interaction term for these variables. In the **Data** pane, select **Diastolic**. Hold down the Ctrl key, and click **Systolic**. Both variables should be selected. Right-click **Systolic**, and select **Create a Single Interaction**. The interaction term **Diastolic\*Systolic** appears in the **Term** group of the **Data** pane.



Click, drag, and drop **Diastolic\*Systemic** onto the model pane. A model is created based on that single effect because the **Auto-update model** option is selected in the right pane. Each time a change is made to the model, the linear regression automatically updates. If you anticipate making many changes or if you are experiencing server performance issues, deselect the **Auto-update model** option.. When auto-updates are disabled, you must click **Update** in the right pane to update the model.

Next, add some effects to the model. Hold down the Ctrl key, and select **Blood Pressure Status, Cause of Death, Leaf ID 1, Sex, Smoking Status, Cholesterol, Height, Smoking, and Weight**. Drag and drop these variables onto the model pane. The linear regression updates to include these effects.




In the right pane, select the **Properties** tab. In this model, **Informative missingness** and **Use variable selection** are not selected. Disabling **Informative missingness** means that observations with missing values are not included in the analysis. Disabling **Use variable selection** means that all variables are used in the model, regardless of how significant they are to the model. For this model, keep the default properties settings.

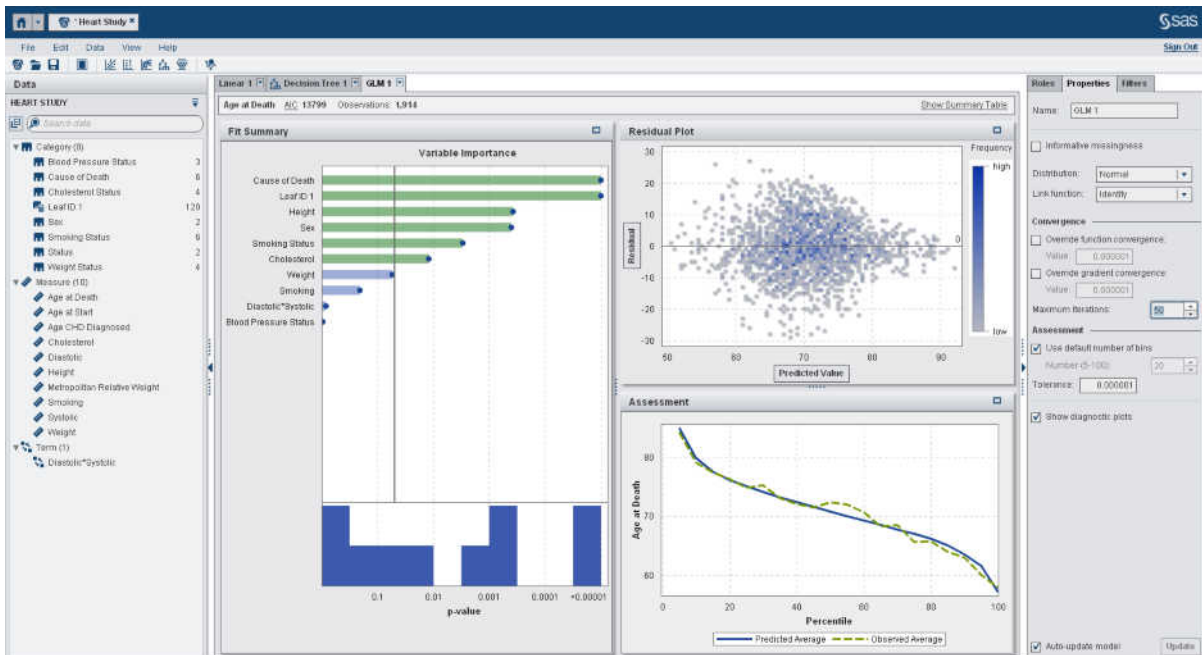
The Fit Summary window indicates that **Cause of Death**, **Leaf ID 1**, and **Height** are the three most important effects in this model.

The Assessment window indicates that the observed average and predicted average are approximately equal for most bins.

Save the project.

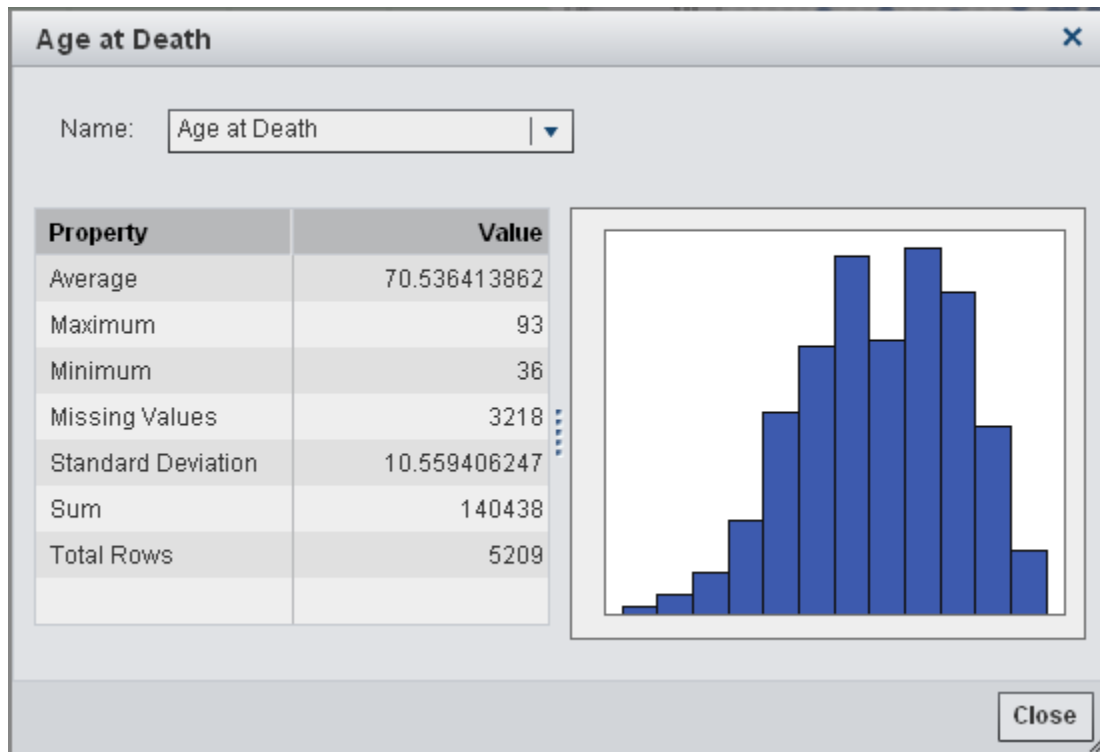
## Create a GLM

From the toolbar, click the  icon to create a new GLM. From the **Data** pane, drag and drop the **Age at Death** variable into the **Response** field in the right pane. In the **Data** pane, hold down the Ctrl key, and select **Blood Pressure Status**, **Cause of Death**, **Leaf ID 1**, **Sex**, **Smoking Status**, **Cholesterol**, **Height**, **Smoking**, **Weight**, and **Diastolic\*Systemic**. Drag and drop these variables onto the model pane.



Click the **Properties** tab in the right pane. The **Distribution** property enables you to specify the distribution of the response variable and to build a model based on that

distribution. The default distribution is **Normal**. To determine whether the normal distribution applies to the response variable, right-click **Age at Death** in the **Data** pane, and select **Properties**.



**Age at Death** is not normally distributed. Although the distribution is not a negative binomial, use the negative binomial distribution for this example. For **Distribution**, select **Negative Binomial**. Next, select **Identity** for **Link function**.

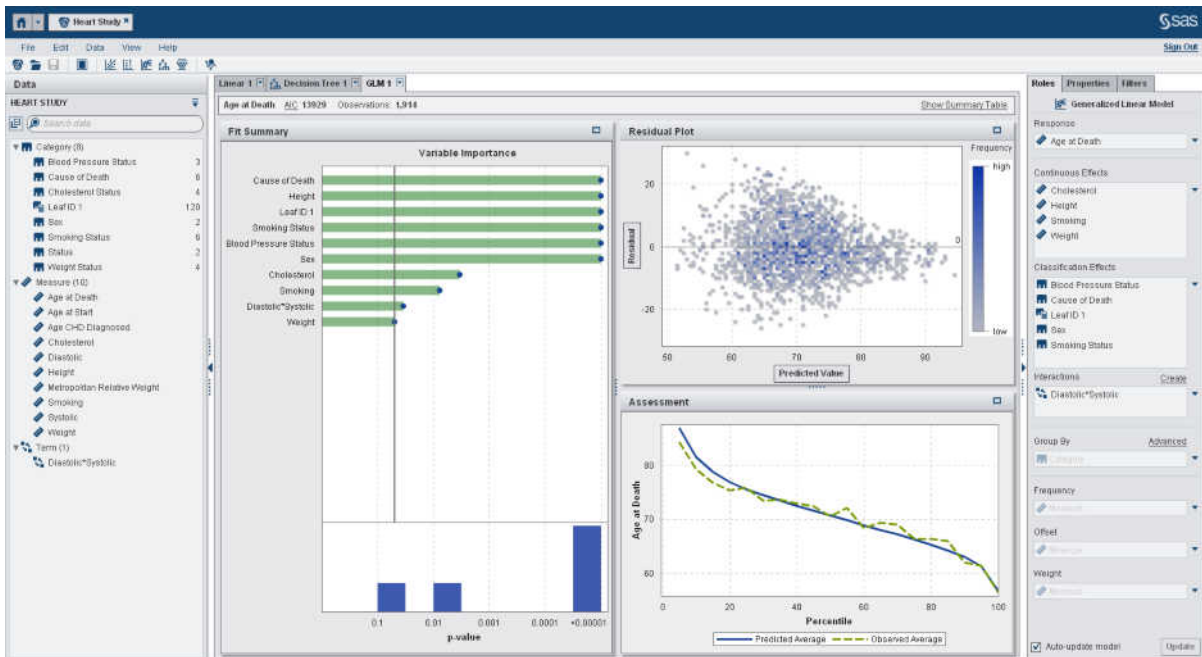
**Note:** You are encouraged to repeat this example with different distributions and link functions and compare their performance and to familiarize yourself with SAS Visual Statistics.

A Convergence Status window appears. This window indicates that the maximum number of iterations was reached, but the model did not converge. To fix this problem, you must increase the value of the **Maximum iterations** property. Click **Close** in the Convergence Status window.

Set the value of **Maximum iterations** to 100, which is the maximum value. The model still does not converge. In this case, you can adjust the function convergence criterion to




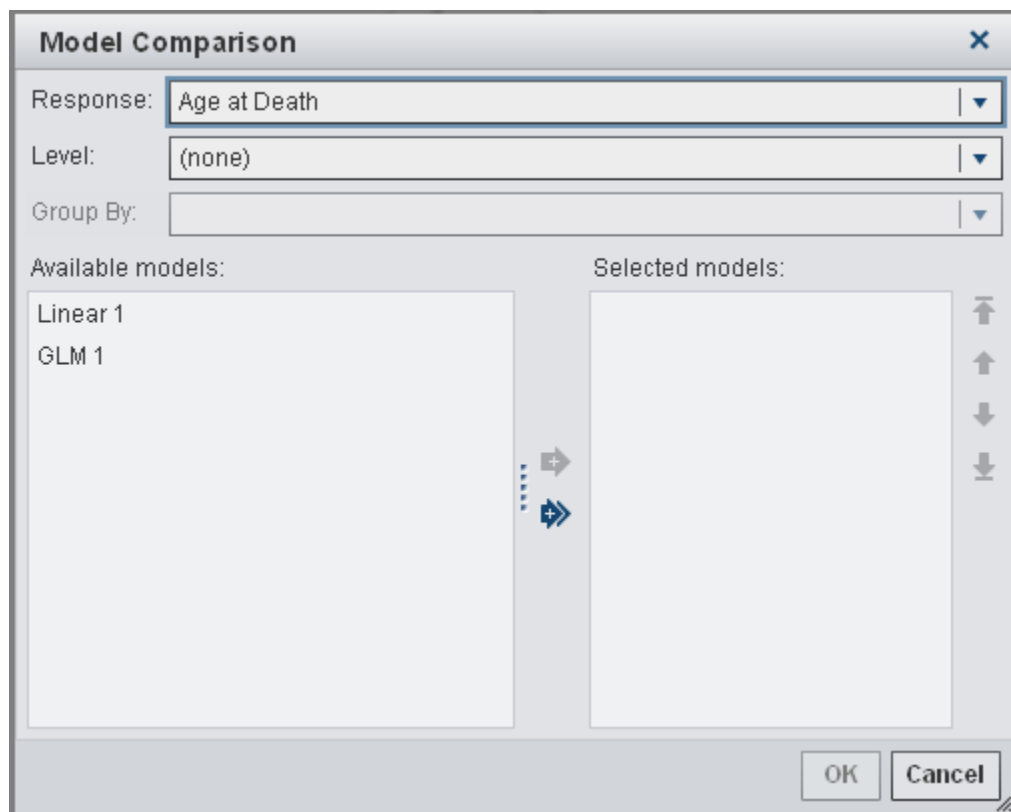
ensure convergence. Select **Override function convergence**. Set the **Value** to 0.00001. Your model should now converge.



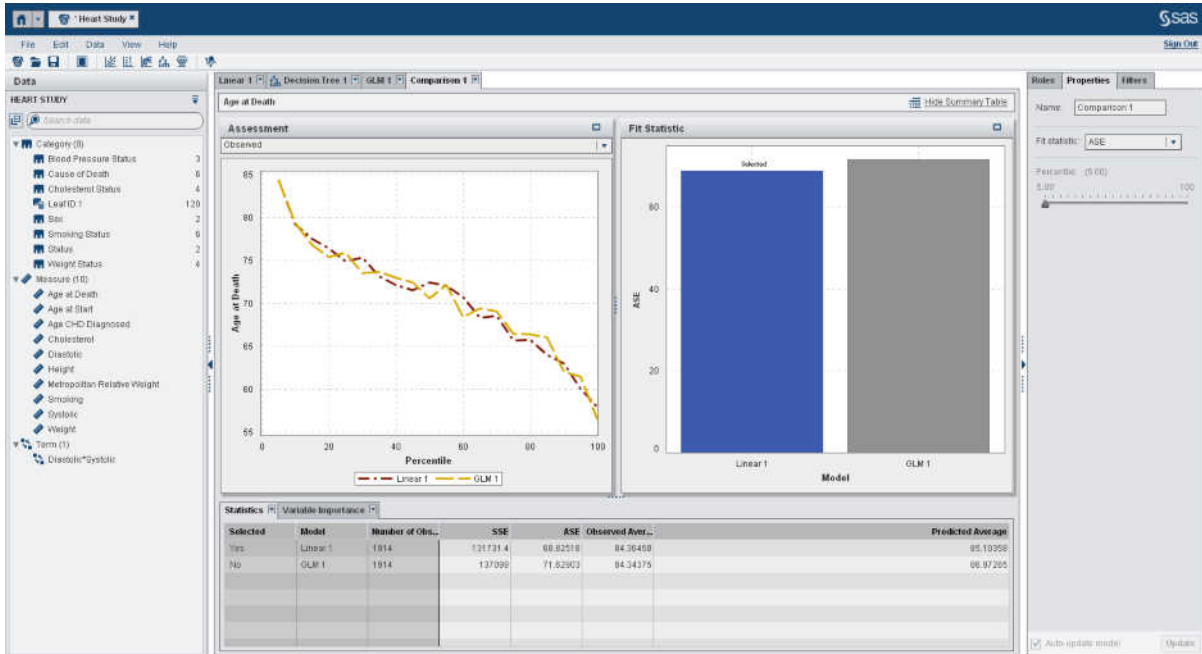
Save the project.

## Perform a Model Comparison

From the toolbar, click the  icon to create a new model comparison. The Model Comparison window appears.



The **Response** variable is already set to **Age at Death**, and **Level** is set to **(none)**. With these settings, the available models are **Linear 1** and **GLM 1**. Click the ➡ to select both models for comparison. Click **OK**.



By default, the fit statistic average squared error **ASE** is used to compare the models. The other available fit statistics are **SSE** and **Observed Average**. Because smaller values are preferred, **Linear 1** is chosen as the champion when **ASE** or **SSE** is the criterion.

When the fit statistic is **Observed Average**, the **Percentile** slider is available. This slider specifies the percentile where the observed average and predicted average are compared.

If you view the Assessment plot, both the **Observed** and **Predicted** plots show that the models are relatively similar.

Now that you have a champion model, you can export the model score code for that model to score new data. Click **File** ► **Export** ► **Model Score Code** from the main menu. In the Export Model Score Code window, select **Linear 1** because it has a better ASE and SSE. Click **OK**. In the Save As window, navigate to a file system location where you have Write privileges. Save the model score code.

Save the example.





# Part 3

## Administrative Tasks

### *Chapter 11*

### *Installation and Configuration* ..... 111



# 11

## Installation and Configuration

<i>Installation</i> .....	111
<i>Configuration</i> .....	111
SAS Visual Statistics Capabilities .....	111
Thresholds for High-Cardinality Data .....	114

---

### Installation

SAS Visual Statistics is installed via the SAS Deployment Wizard. There are no prompts or pages specific to SAS Visual Statistics during the installation. During the installation of SAS Visual Statistics, refer to the *SAS Visual Analytics: Installation and Configuration Guide* and the related resources mentioned in that guide.

---

### Configuration

#### SAS Visual Statistics Capabilities

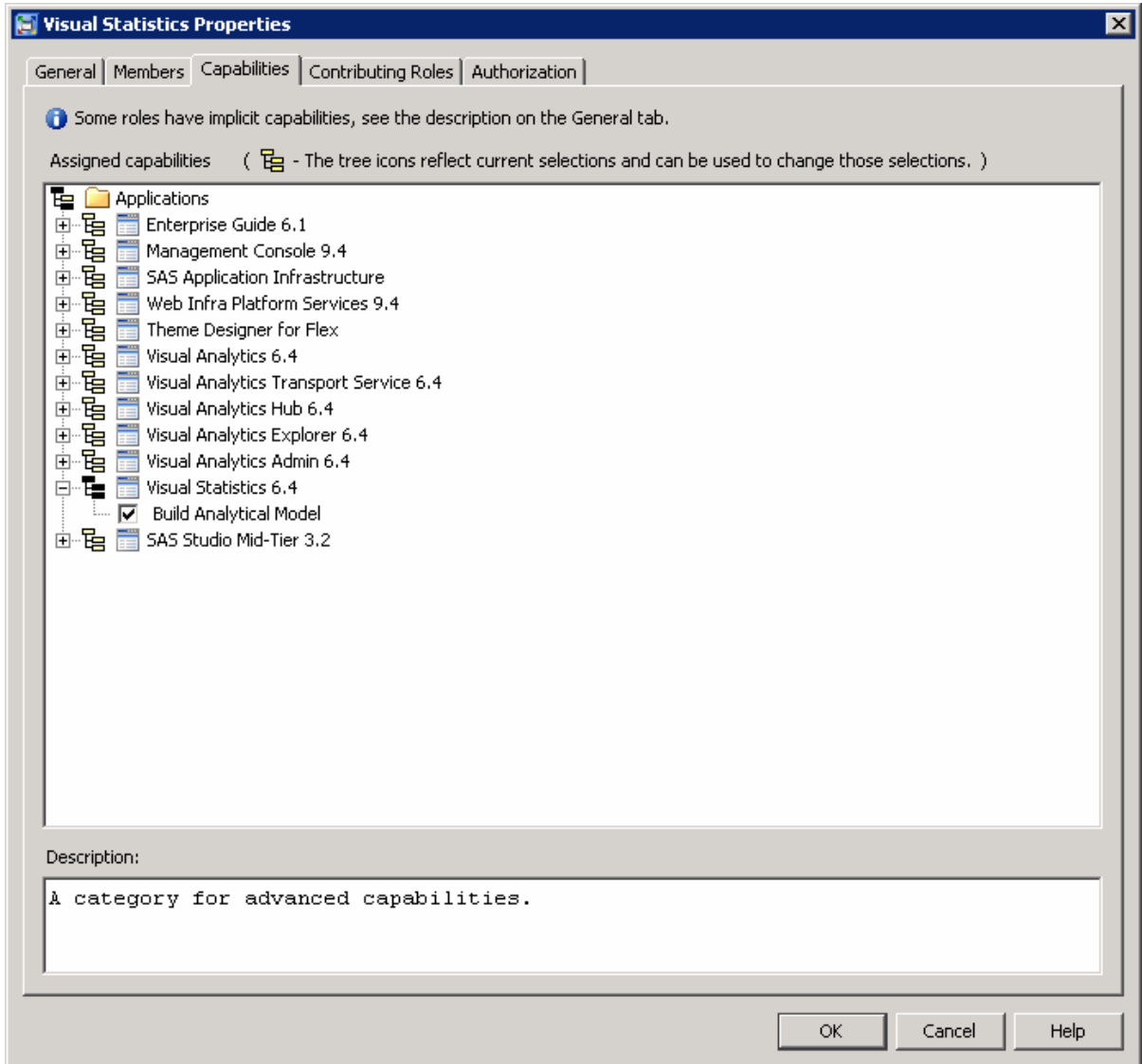
It is strongly recommended that you create a separate role in SAS Management Console that provides access to SAS Visual Statistics specifically for users who need to build models in SAS Visual Statistics. The primary reason is because model building in SAS Visual Statistics can have significant adverse effects on the performance of your

computing environment. Limiting access to the model building capabilities of SAS Visual Statistics mitigates this risk.

To create a new role that provides access to model building in SAS Visual Statistics, complete the following steps:

- 1** In SAS Management Console, access the **User Manager** plug-in. You must have authorization to use the **User Manager** plug-in and permission to create a new role.
- 2** Right-click inside the **User Manager** plug-in, and select **New ► Role**. The New Role Properties window appears.
- 3** On the **General** tab, enter `visual Statistics` in the **Name** field. You can add a **Display Name**, **Description**, or both.
- 4** On the **Capabilities** tab, select the **Build Analytical Model** capability.





## 5 Click **OK**.

You now have a specific role for a user who wants to access the model building capabilities of SAS Visual Statistics. To assign the **Visual Analytics: Analysis** capabilities to this role, use the **Contributing Roles** tab in the Visual Statistics Properties window.

## Thresholds for High-Cardinality Data

High-cardinality data has one or more columns that contain a very large number of unique values. For example, user names, e-mail addresses, and bank account numbers can be high-cardinality data items. SAS Visual Statistics provides metadata properties that constrain high-cardinality data to ensure proper performance, meaningful results, or both. Each constraint is independent of other constraints and should be adjusted based on the needs of your data and system performance. You must restart the SAS Web Application Server to apply your changes.

The **classCardinalityLimit** and **groupbyCardinalityLimit** properties exist to control the workload during model creation. There are two separate, but related, consequences to these properties. First, you cannot specify an effect variable or a group by variable if it contains more distinct levels than the value specified in the respective property. Second, you cannot add an effect variable or a group by variable to a model if it causes the total number of distinct levels to surpass the value specified in the respective property.

In contrast, the **filterCardinalityLimit** and **responseCardinalityLimit** properties affect only the number of distinct levels that are displayed. This limitation ensures that only a meaningful number of distinct levels are displayed when selecting filter criteria or when choosing a level of interest for the logistic regression model. You can specify a filter variable or logistic regression response variable that contains more distinct levels than this value, but only the top *n* levels are displayed.

To access these metadata properties, open SAS Management Console. On the **Plug-Ins** tab, navigate to **Application Management ► Configuration Manager ► SAS Application Infrastructure ► Visual Analytics ► Visual Statistics**. Right-click **Visual Statistics**, and select **Properties**. Select the **Advanced** tab. The following four metadata properties are shown:

vstat.classCardinalityLimit	1024
vstat.filterCardinalityLimit	1024
vstat.groupbyCardinalityLimit	1024
vstat.responseCardinalityLimit	1024

**classCardinalityLimit**

determines the maximum number of distinct levels allowed for all classification effects and interaction terms in the model. This limitation is on the cumulative total of classification effects and interaction terms included in the model. This limitation is calculated whenever a model is built or updated.

**filterCardinalityLimit**

determines the maximum number of distinct levels displayed for the filter variables. This value is the maximum number of distinct levels displayed on the **Filters** tab in the right pane. A filter variable can contain more distinct levels than this value, but only the first  $n$  levels are displayed.

**groupbyCardinalityLimit**

determines the maximum number of distinct levels allowed for the group by variables. This limitation is on the cumulative total for group by variables. This limitation is calculated whenever a model is built or updated.

**responseCardinalityLimit**

determines the maximum number of distinct levels displayed for the response variable in a logistic regression model. This value is the maximum number of distinct levels displayed when you specify a level of interest. When you specify a response variable with more distinct levels than this value, only the first  $n$  levels are displayed.

You should determine your cardinality limits based on your input data, computer systems, and familiarity with the statistical models.



## Recommended Reading

Here is the recommended reading list for this title:

- *SAS Visual Analytics: User's Guide*
- *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*

For a complete list of SAS publications, go to [sas.com/store/books](https://sas.com/store/books). If you have questions about which titles you need, please contact a SAS Representative:

SAS Books

SAS Campus Drive

Cary, NC 27513-2414

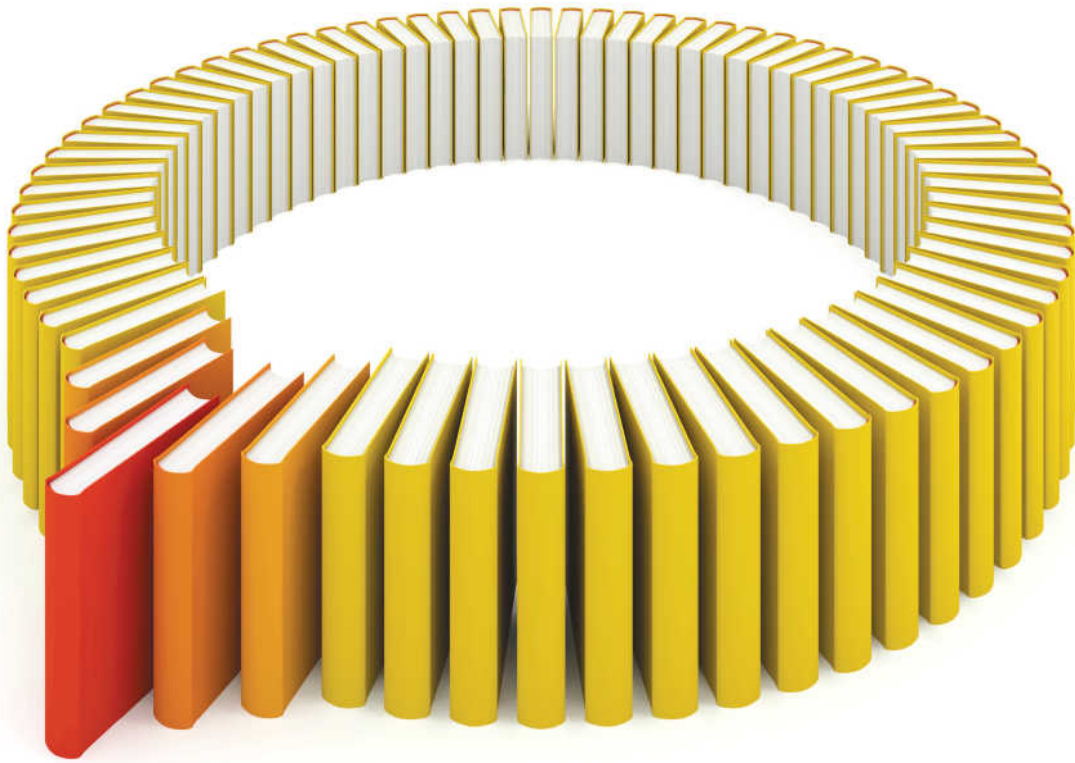
Phone: 1-800-727-0025

Fax: 1-919-677-4444

Email: [sasbook@sas.com](mailto:sasbook@sas.com)

Web address: [sas.com/store/books](https://sas.com/store/books)





# Gain Greater Insight into Your SAS® Software with SAS Books.

Discover all that you need on your journey to knowledge and empowerment.

