SAS® Visual Analytics 7.5: Working with SAS® Visual Statistics

About SAS Visual Statistics

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. SAS Visual Analytics 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 SAS Visual Analytics. 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 compare models, you can export model score code for all models that you create. With exported model score code, you can easily apply your model to new data.

Specifying Settings for SAS Visual Statistics

There are settings that are specific to SAS Visual Statistics, and there are global settings that are applied to all SAS web applications. For more information about global settings, see *SAS 9.4 Web Applications: General Usage Help* or "Modify SAS Visual Analytics Settings" in *SAS Visual Analytics: Designing Reports* .

Settings for SAS Visual Statistics are saved on a per-user basis. All of your settings persist between sessions.

1. In the application bar, click your name, and then click **Settings**.
2. Under **SAS Visual Analytics**, select **Visual Statistics**.
3. You can change the following settings:
   - **Fit summary p-value precision** determines the minimum number of decimal places used when displaying *p*-values.
   - **Sort categorical response levels in descending order**
   - **Default statistic for Model Comparison** specifies your default statistic preference for **Category Response** and **Measure Response** variables within the **Model Comparison** object.

4 Click **Close** to apply your changes.

> **TIP** When you click **Reset**, the settings revert to their original SAS Visual Statistics configuration.

Modeling Information

Available Models

The following models are available in SAS Visual Statistics:

- Cluster on page 17 segments the input data into groups that share similar features.
- Decision Tree on page 21 creates a hierarchical segmentation of the input data based on a series of rules applied to each observation.
- Generalized Linear Model on page 28 is an extension of a traditional linear model that allows the population mean to depend on a linear predictor through a nonlinear link function.
- 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 38 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.

Overview of 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 measure variable by clicking ⌄. Change **Classification** to **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.

To obtain summary statistics and a histogram for each measure variable, click in the **Data** pane, and then click **View measure details**.

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 Effects

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.

You can create an interaction by selecting the variables of interest in the **Data** pane. Right-click one of the selected variables and select **New interaction effect**. Select the interaction type that you are interested in. You can also create an interaction by selecting **+New data item** at the top of the **Data** pane, and then selecting **Interaction effect**. The New Interaction Effect window will appear and you can specify the variables of interest and the interaction type.

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 objects provide a property to

automatically perform variable selection. 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.

SAS Visual Statistics provides the following variable selection method:

- **Backward** — All candidate effects are included in the initial model. The least significant effects are removed one at a time until the model is significantly weakened by removing an effect.

Missing Values

By default, SAS Visual Statistics handles missing values by dropping all observations that contain a missing value in any assigned role. However, the linear regression, logistic regression, and generalized linear model provide the **Informative missingness** option. In some cases, the fact that an observation contains a missing value provides relevant modeling information. Selecting this option 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.

The decision tree on page 22 model handles missing values differently.

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. 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 group names are concatenated group values using the order of the order in which the variables appear under **Group by**. In the Fit Summary window, groups are ordered by their statistic value.

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

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.

Score Code

Model scoring refers to the process of generating predicted values for a data set that might not 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.

To generate model score code, right-click on the model canvas, and select **Export model**. In the Export Model window, click **Export**.

Score code is saved as a .sas file and can be viewed in any text editor.

Note: It is possible for your exported score code to have lines of code that exceed the maximum line length of 32768. There are two solutions for this issue. The first solution requires that you edit the exported text file to include a line break on each of the long lines and to insert `/ lrecl=1000000` in the %INCLUDE statement. The second solution requires that you open the exported text file in a SAS Program Editor and insert a line break on each of the long lines. In the SAS Program Editor, there is a limit of 6000 characters per line.

Derive Predicted Values

For many of the predictive models, SAS Visual Statistics can create variables that contain prediction information for each observation in the data set. After these variables are created, they can be used in any other object, including other predictive models.

To create the new variables, complete the following steps:

1. Create a valid predictive model.
2. Right-click on the model canvas, and select **Derive predicted**, **Derive a leaf ID variable**, or **Derive cluster ID items**.

   Derive predicted is available in the **Linear Regression**, **Generalized Linear Model**, and **Logistic Regression** objects. **Derive a leaf ID variable** is available in the **Decision Tree** object. **Derive cluster ID items** is available in the **Cluster** object.
3. In the New Prediction Items window, enter a name for the **Predicted values** and either the **Residual values** or the **Probability values**. Residual values are available for linear regressions and generalized linear models. Probability values are available for logistic regressions.

   In the New Leaf ID window, enter a name for the **Leaf ID** variable.

   In the New Cluster ID Items window, enter a name for the **Cluster ID** and the **Distance from centroid value**.
4. Click **OK**. The predicted values appear in their own section of the **Data** pane.

Depending on the model, the information contained in each variable is slightly different.

Predicted values
For linear regressions and generalized linear models, this is a numeric value that is generated by the model. Or, this is the value that would have been generated by the regression model if the observation was scored by the model.

For logistic regressions, this is the decision generated by the model based on the calculated probability and **Prediction cutoff** parameter. All observations are classified as one of the following: the event level of interest, not in the event level of interest, or missing.

Residual values
The computed residual for each observation. This value is available for linear regressions and generalized linear models.

Probability values
The computed probability for each observation to take the event level of interest. Observations with probability values that are greater than or equal to the **Prediction cutoff** parameter are predicted to be in the event level of interest. Observations with probability values that are less than the **Prediction cutoff** parameter are considered to not be in the event level or interest. That is, there is no prediction made regarding each individual measurement level; there is a prediction made only between the measurement level of interest and everything else.

Leaf ID
The leaf ID that contains the observation.

Cluster ID
The cluster that contains the observation.

Distance from centroid value
The distance from the observation to the centroid of the cluster that contains the observation.

Note: When you derive a parameter from one data set and then change data sets, the derived parameters are still available on the **Data** pane. However, you should use these derived parameters only when the active data set is the data set that you used to create them.

Getting Started with SAS Visual Statistics

Overview

This is a brief overview of using SAS Visual Statistics to derive a new variable, create two different models, and compare those models. This example uses the Framingham Heart Study data set, located at http://support.sas.com/documentation/onlinedoc/viya/examples.htm, to compare the performance of a linear regression model and a generalized linear model. 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.

Download the Sample Data

1. In a web browser, navigate to http://support.sas.com/documentation/onlinedoc/viya/examples.htm.
2. Download the file **heart.csv** to your local machine.

Create the Report

This example assumes that you have already signed in to SAS Home.

Here are the steps to create the report:

1. From SAS Home, click **Report Builder**. This opens SAS Visual Analytics where you can open a data source, create a model, or load a project.
2. Click the **Data** button in the welcome window to load your data. A window appears that enables you to select the data source for this project.
3. Click the **Import** button, and then click **Text File**.
4. Navigate to the location where you saved **heart.csv**, select **heart.csv**, and click **Open**.
5. Ensure that **Comma** is selected in the **Specify a file delimiter** field. Enter `2` in the **Data records begin on row** field. Enter `HEART` in the name field.
6. Click the **Import** button. A new report automatically opens with the table loaded.
7. By default, the report is named `Report 1`, which is displayed at the top of the page. Before continuing with the example, rename the project by saving it.

   In the upper right corner of the report, click the Menu button ⋮, and then select **Save**. This opens the Save As window. In the **Folders** pane, navigate to a location where you have Write permission. In the **Name** field, enter `Heart Study`, and click **Save**.

   Typically, you can save your work in `My Folder`.

For more information about reports, see *SAS Visual Analytics: Designing Reports*.

Create a Decision Tree

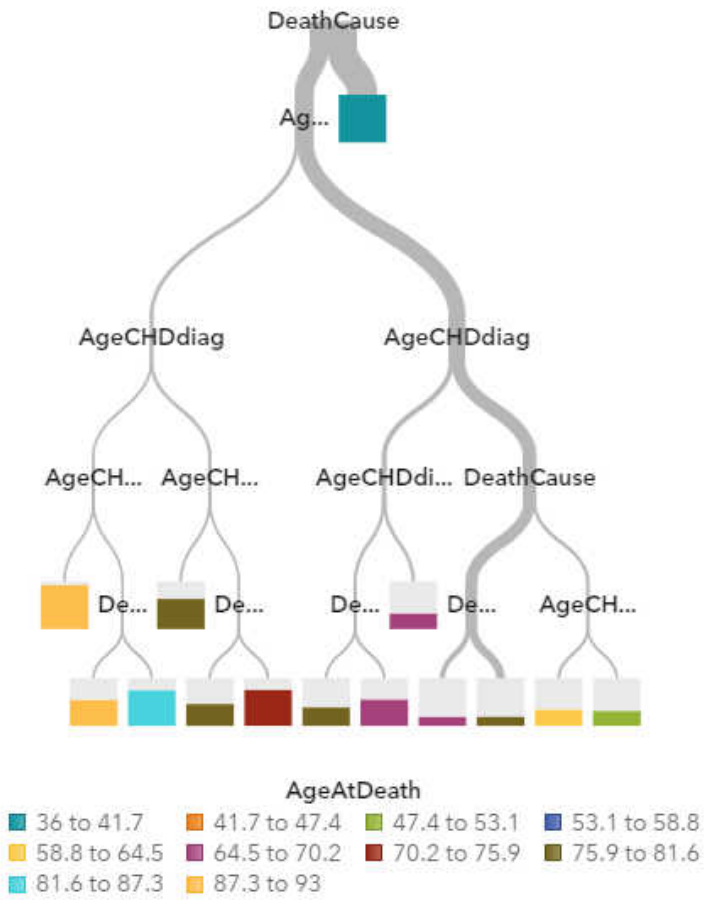
Here are the steps to create a decision tree:

1. From the left pane, click Objects to select an object. Drag and drop the Decision Tree onto the canvas to create a decision tree.

   Note: All SAS Visual Statistics objects are located at the bottom of the **Objects** pane.

2. Click Roles in the right pane. For **Response**, click **Add**, and then select **AgeAtDeath**.
3. For **Predictors**, click **Add**, and then select **DeathCause**, **Sex**, **AgeCHDdiag**, and **Cholesterol**. Click **OK**. The decision tree automatically updates. Notice that **AgeAtDeath** is automatically binned into ten groups.

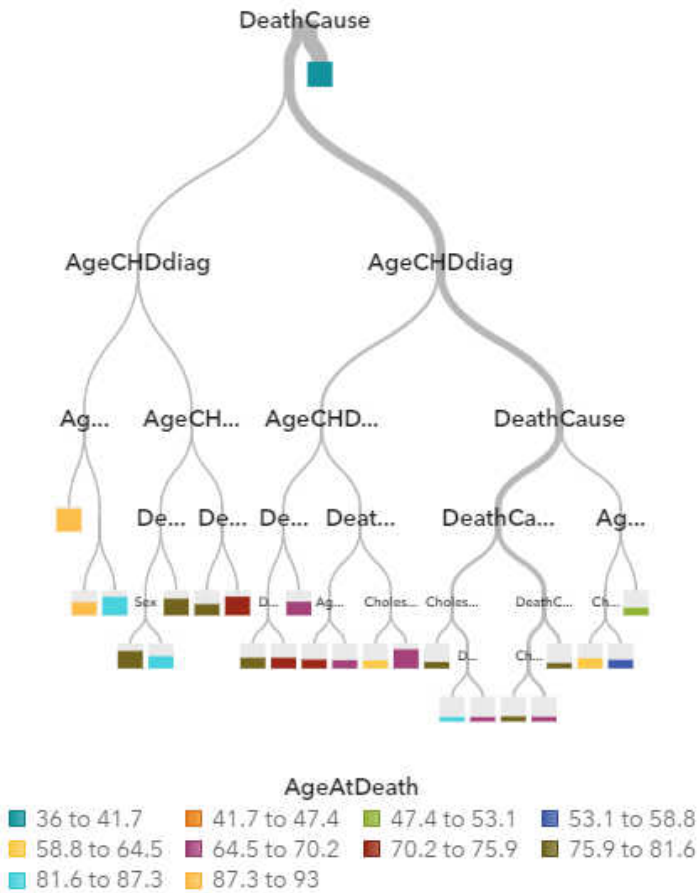
Tree

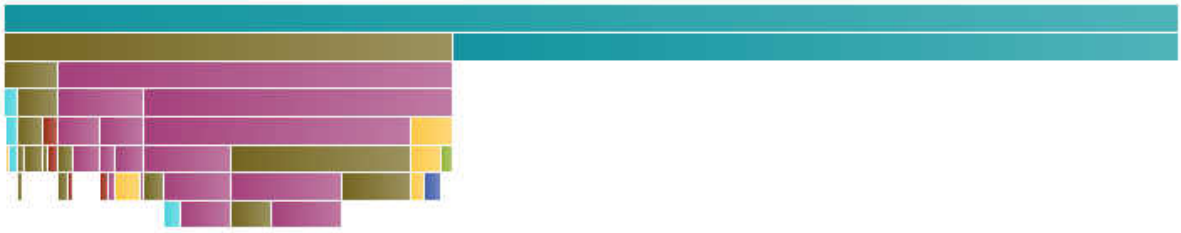


4 Click Options in the right pane. Select 8 for the **Maximum levels** option. This option enables you to increase the size of your tree.

Your decision tree should resemble the following:

Tree





5 Click Maximize ↙↗ to enter maximize mode.

In the details table, select the **Node Rules** tab. Notice that each predictor was used at least once.

6 Click Restore ↘↖ to exit maximize mode.

7 In the model canvas, right-click, and select **Derive a leaf ID variable**. The default name for this variable is **Leaf ID (1)**.

In the New Leaf ID window, click **OK**. The **Leaf ID (1)** item appears in the **Data** pane.

8 Click Save to save the report.

Create a Linear Regression

Here are the steps to create a linear regression:

1. Click Add + to add a new page.
2. From the left pane, click Objects to select an object. Drag and drop the Linear Regression onto the canvas to create a linear regression.
3. In this example, the variable of interest is **AgeAtDeath**.

   Click Roles in the right pane. For **Response**, click **Add**, and then select **AgeAtDeath**.
4. Next, you must 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 then 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 effect for these variables.

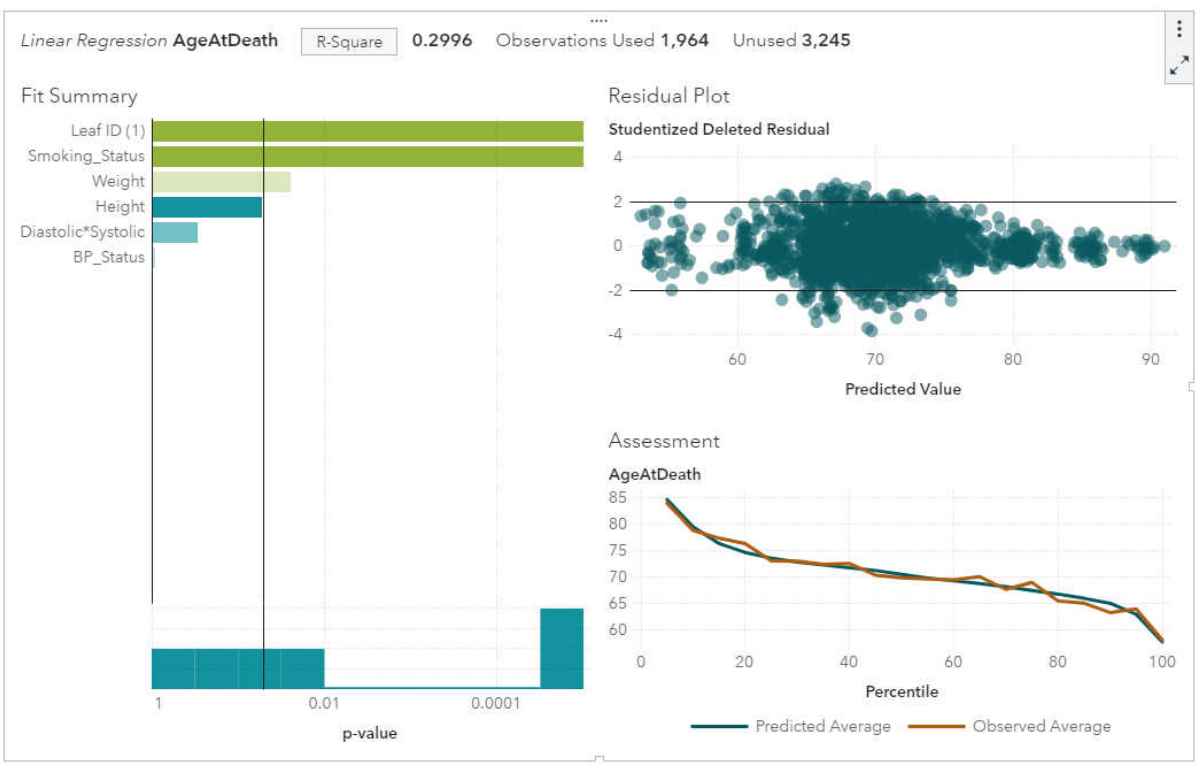
   Follow these steps to create an interaction term:

   a. In the **Data** pane, click **New data item**, and select **Interaction effect**.
   b. In the New Interaction Effect window, move **Diastolic** and **Systolic** from the **Available columns** area into the **Effect elements** area.
   c. Click **OK**.

   The interaction term **Diastolic*Systolic** appears in the **Interaction Effect** group of the **Data** pane.
5. Drag and drop **Diastolic*Systolic** onto the canvas. A model is created based on that single effect because **Enable auto-refresh** is enabled.

> **TIP** Each time a change is made to the model, the linear regression automatically updates. In the upper right corner of the report, click the Menu button ⋮, and select **Disable auto-refresh** to prevent automatic updates. You might want to disable automatic updates if you anticipate making many changes or if you are experiencing server performance issues. When automatic updates are disabled, you must click Refresh to update the model.

6 In the **Data** pane, select **BP_Status**, **Smoking_Status**, **Height**, **Weight**, and **Leaf ID (1)**. Drag and drop these variables onto the canvas. The linear regression updates to include these effects.



7 In the right pane, select Options. In this model, **Informative missingness** is not selected. Disabling **Informative missingness** means that observations with missing values are not included in the analysis.

The Fit Summary indicates that **Leaf ID (1)** and **Smoking Status** are the most important effects in this model.

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

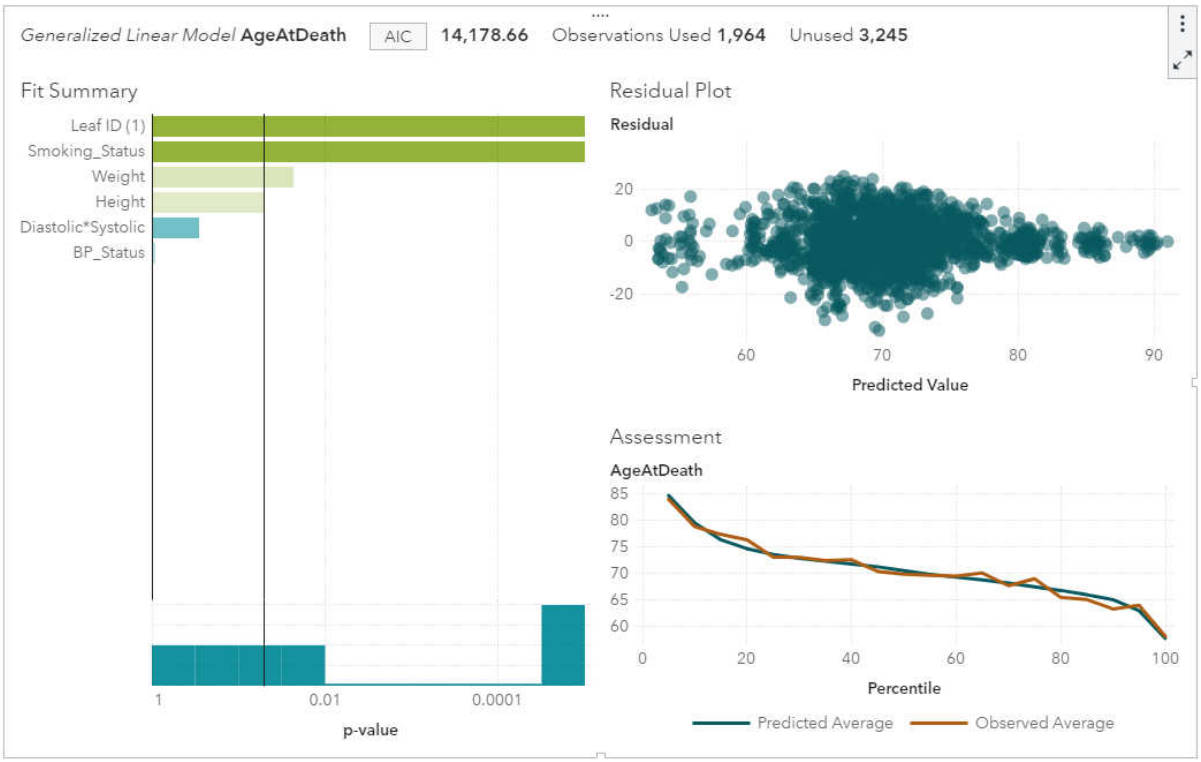
8 Save the report.

Create a Generalized Linear Model

Here are the steps to create a generalized linear model:

1 From the linear regression, right-click on the model canvas, press and hold the Alt key, and then select **Duplicate on new page as** ⇨ **Generalized Linear Model**.

2 The same variables used to train the linear regression model are used for the generalized linear model.



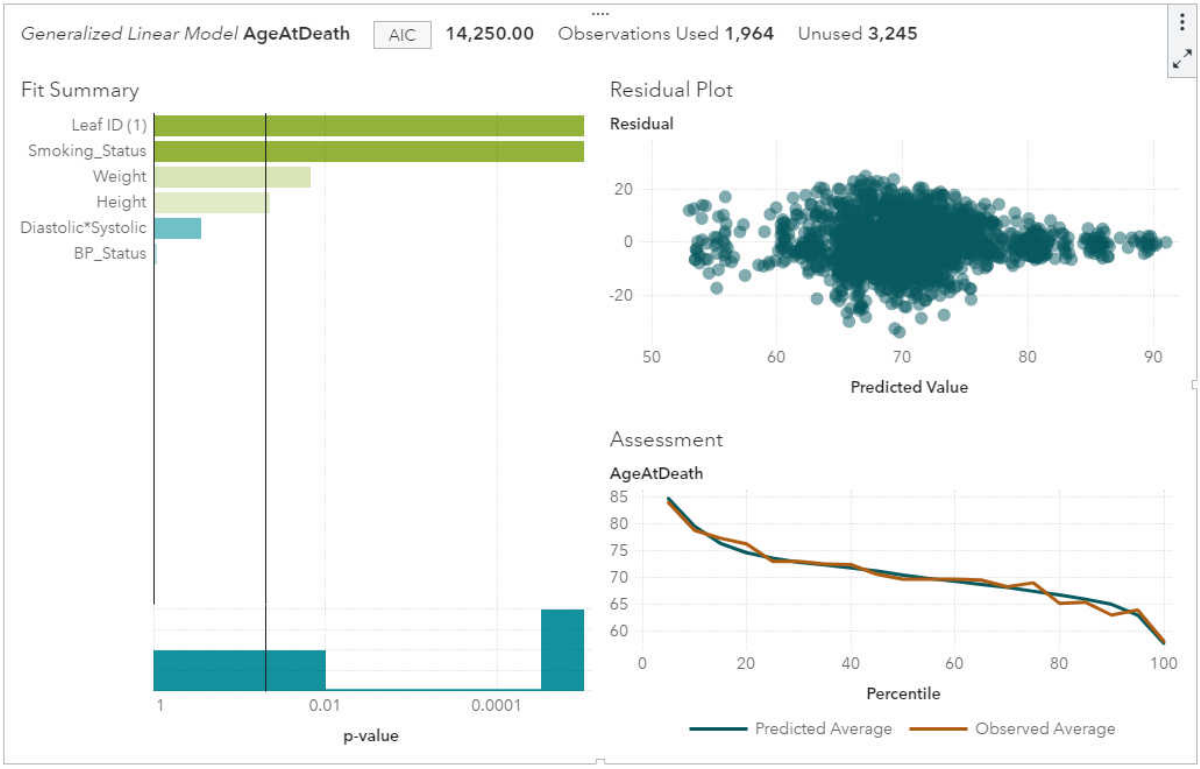
3 Click Options in the right pane. The **Distribution** option enables you to specify the distribution of the response variable and to build a model based on that distribution. The default distribution is **Normal**.

To determine whether the normal distribution applies to the response variable, in the **Data** pane, click Actions, and select **View measure details**. In the Measure Details window, select **AgeAtDeath**.

4 Notice that **AgeAtDeath** is not normally distributed since the distribution has two peaks and is slightly skewed left. Click **Close**.

5 Although the distribution is not exactly Poisson, use the Poisson distribution for this example. For the **Distribution** option, select **Poisson**. Next, select **Identity** for **Link function**.

Note: You are encouraged to repeat this example with different distributions and link functions and compare their performances, familiarizing yourself with SAS Visual Statistics.



6 Save the report.

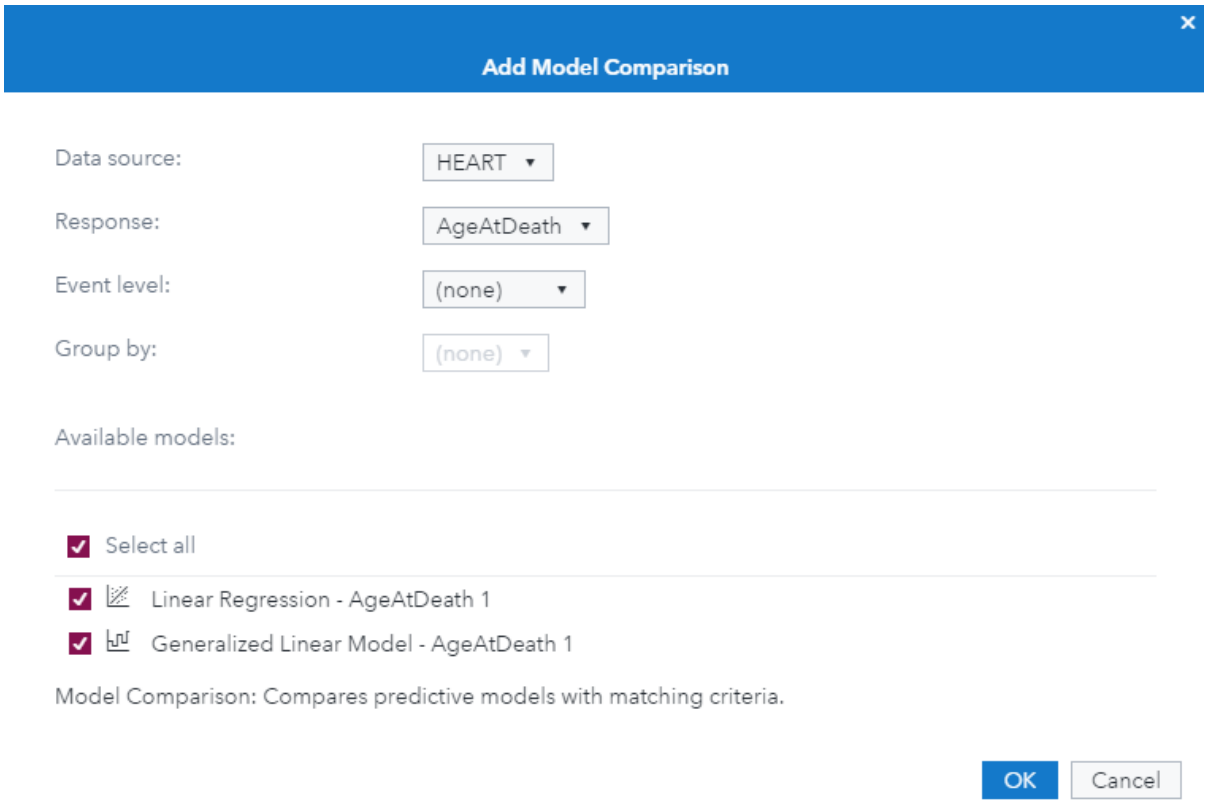
Perform a Model Comparison

Here are the steps to perform a model comparison:

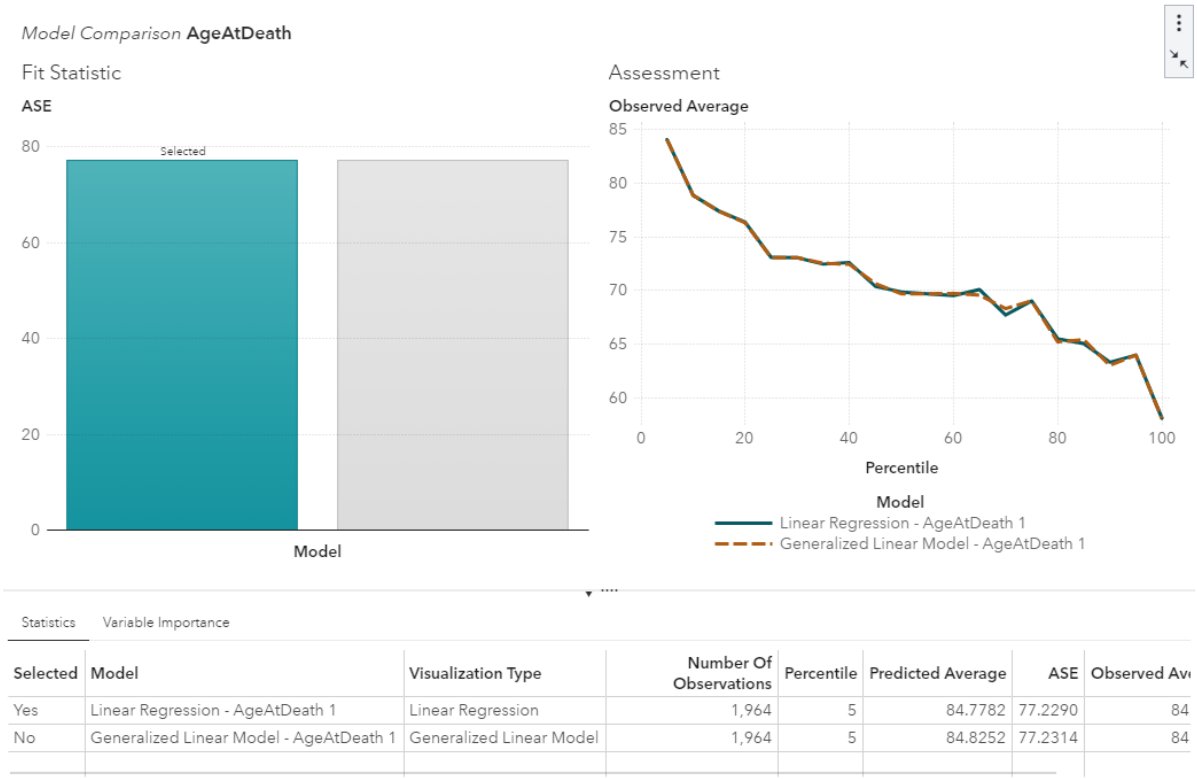
1 Click Add + to add a new page.

2 From the left pane, click Objects to select an object. Drag and drop the Model Comparison onto the canvas to create a model comparison.

The **Response** variable is already set to **AgeAtDeath**. Set the **Event Level** to **(none)**. The **Group by** option is unavailable. With these settings, the available models are **Linear Regression – AgeAtDeath 1** and **Generalized Linear Model – AgeAtDeath 1**.



3 Select **Linear Regression – AgeAtDeath 1** and **Generalized Linear Model – AgeAtDeath 1**, and then click **OK**.



Selected	Model	Visualization Type	Number Of Observations	Percentile	Predicted Average	ASE	Observed Av
Yes	Linear Regression - AgeAtDeath 1	Linear Regression	1,964	5	84.7782	77.2290	84
No	Generalized Linear Model - AgeAtDeath 1	Generalized Linear Model	1,964	5	84.8252	77.2314	84

Note: Click Maximize ⤢ to enter maximize mode and to view the summary table.

4 By default, the fit statistic average square error, **ASE**, is used to compare the models. The other available fit statistics are **SSE** and **Observed Average**. Because smaller values are preferred, the linear regression is chosen as the champion when **ASE** or **SSE** is the criterion. The models are very similar.

   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. In some percentiles, the generalized linear model might be chosen over the linear regression.

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

5 The champion model is **Linear Regression – AgeAtDeath1** and is marked as **Selected** in the Fit Statistic plot and summary table. You can export the model score code for that model to score new data.

   Here are the steps to export the model score code:

   a Click Restore to exit maximize mode.

   b Right-click in the model canvas, and select **Export selected model**.

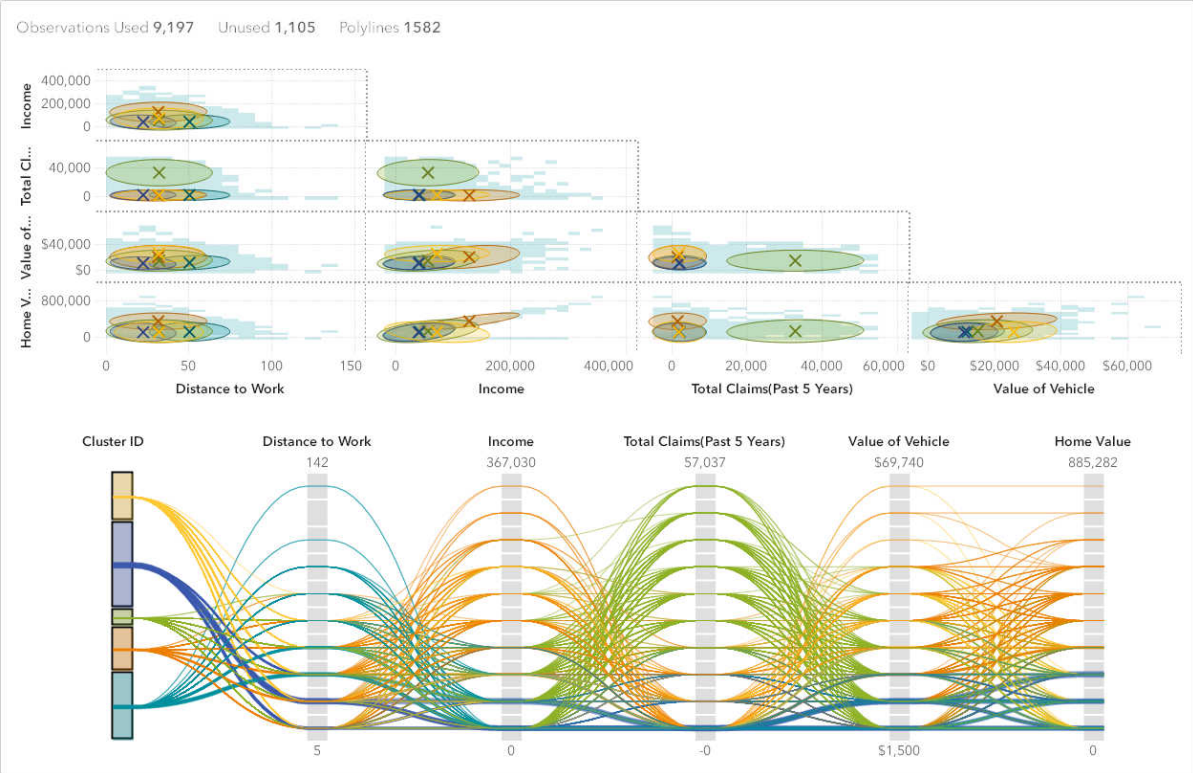
   c In the Export Model window, click **Export**.

Working with Clusters

Overview of Clusters

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 exactly one cluster. From the clustering analysis, you can generate a cluster ID variable to use in other models.

Figure 1 *Example Cluster*



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. To view a larger plot for an effect pair, right-click inside that plot, and click **Isolate**. If **Clustered heat map** is specified for the **Binned plot style**, then stacked bar charts that show additional information about the effect pair are displayed.

Each cluster is assigned a unique color. Although each cluster is unique in *n*-space, the two-dimensional projections will overlap. It is important to note that every observation can belong to exactly one cluster. However, because the Cluster Matrix displays a projection in just two dimensions, multiple clusters can overlap an observation.

When a heat map is not used, individual observations are color-coded to indicate cluster membership.

The Parallel Coordinates plot enables you to make several inferences. You can interact with the plot to explore the data based on cluster membership, a specified range for one or more variables, or both. By restricting the display to specific clusters and data ranges, you can focus on the data that interests you. You can also reverse the sorting of one or more variables on the axis from the pop-up menu. This can be useful when comparing variables that are inversely correlated.

How to Create a Cluster Object

To create a cluster object, complete the following steps:

1 Drag and drop the icon onto the canvas.

2 Click in the right pane. Specify at least two measure variables for **Variables**. You cannot specify an interaction term.

Cluster Options

The following options are available for the cluster object:

Count
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

- **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.

Standardization

- **None** — No standardization is applied.
- **Standard deviation** — Transforms the measure variables so that they have a mean of zero and a standard deviation of 1.

Cluster Display Options

The following display options are available for the cluster object:

General

Plot layout
specifies how the subplots within objects are displayed on the canvas. **Fit** aligns all of the subplots on the canvas automatically. **Stack** displays the subplots as if they are in a slide deck where only one plot is displayed at a time. When **Stack** is specified, a control bar instead of a scroll bar lets you move between the subplots.

Cluster Diagram

Show ellipse outlines
enables you to display the cluster projection ellipses in the cluster matrix.

Fill ellipses
specifies whether cluster ellipses are filled.

Visible roles
determines how many effects are shown in the cluster matrix. Valid values are integers between 2 and 6, inclusive.

Binned plot style
specifies how the heat map of the observations is displayed. Possible values are **Clustered heat map**, **Heat map**, and **Bubble plot**.

Legend visibility
specifies whether the legend is displayed in the cluster diagram.

Parallel Coordinates Plot

Number of bins
specifies the number of bins to use in the Parallel Coordinates plot. Valid values are integers between 2 and 16, inclusive.

Maximum polylines
specifies the maximum number of polylines generated by the parallel coordinates algorithm.

Visible roles
determines how many effects are shown in the Parallel Coordinates plot. Valid values are integers between 2 and 10, inclusive.

Derive Cluster ID Items

After clustering is finished, you can create several new data items based on the results of the cluster object. To derive these items, right-click anywhere in the cluster object, and select **Derive cluster ID items**. Based on the data items in the cluster, the following new data items can be created:

- **Cluster ID** — the cluster that contains the observation.
- **Distance from centroid value** — specifies the distance from the observation to the centroid of the cluster that contains the observation.

Details Table

When you click ↙↗ in the object toolbar of the canvas, the details table is displayed at the bottom of the canvas. The details table contains the following information:

Centroids
Provides the definition for each cluster.

Cluster Summary
Provides summary statistics for each cluster. Available statistics include:

- **Cluster ID**
- **Observations**
- **RMS of STD** — The root mean square distance between observations in the cluster. This is a measure of within-cluster homogeneity. Smaller values are preferred.
- **Within cluster SS** — The sum of squares computed on the observations within a cluster. This is a measure of within-cluster homogeneity.
- **Min centroid-to-observation** — The distance between the centroid and the observation that is closest to the centroid.
- **Max centroid-to-observation** — The distance between the centroid and the observation that is farthest from the centroid.
- **Nearest Cluster** — The cluster that contains the nearest cluster centroid.
- **Centroid Distance** — The distance between the cluster centroid and the nearest cluster centroid.

Model Information
Gives an overview of the model.

Working with Decision Trees

Overview of Decision Trees

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

The decision tree creates classification trees that model categorical data. Measure response variables are automatically binned to create a classification tree.

How to Create a Decision Tree

To create a decision tree, complete the following steps:

1. Drag and drop the icon onto the canvas.
2. Click in the right pane. Specify a single variable as the **Response** variable.
3. Specify at least one variable for **Predictors**.

About Interactive Mode

The decision tree enables you to manually train and prune nodes by entering interactive mode. To enter interactive mode, right-click in the Tree window, and select **Enter interactive mode**. You can also right-click a node and select the train, split, or prune action to enter interactive mode. To leave interactive mode, right-click in the Tree window, and select **Exit interactive mode**.

While in interactive node, adding a new predictor has no effect on the decision tree. However, you can manually make new splits with the added predictors. Prompt filters and action filters also have no effect while in interactive mode.

There are several actions that cause you to automatically exit interactive node. Here are some examples:

- editing or deleting a calculated column that is assigned to the decision tree

- editing or deleting a custom category that is assigned to the decision tree
- editing or deleting a data source filter that is on the data that is assigned to the decision tree
- editing a parameter that is used by a data item that is assigned to the decision tree
- removing a predictor that is assigned to the decision tree

If you attempt any of these actions, a warning message will pop up indicating that your interactive changes will be removed from the model if you continue.

Note: When you leave interactive mode, you lose all of your changes. The interactive decision tree feature is disabled when you have a measure response with more than ten bins.

Decision Tree Options

The following options are available for the decision tree:

General

Event level
enables you to choose the event level of interest. When your response variable contains more than two levels, SAS Visual Statistics treats all observations in the level of interest as an event and all other observations as nonevents.

Missing assignment
specifies how observations with missing values are included in the model.

- **None** — observations with missing values are excluded from the model.
- **As machine smallest** — missing interval values are set to the smallest possible machine value such that the observations will always be in the split with the lower variable values. Missing category values are treated as a unique measurement level.

Growth strategy
specifies one of three default sets of options or **Custom**. The subsequent **General** options are hidden if you select one of the predefined growth strategies.

Maximum branches
specifies the maximum number of branches allowed when splitting a node. The default value is 2.

Maximum levels
specifies the maximum depth of the decision tree. The default value is 6.

Leaf size
specifies the minimum number of observations allowed in a leaf node. The default value is 10.

Response bins
specifies the number of bins used to categorize a measure response variable. The default value is 10.

Predictor bins
specifies the number of bins used to categorize a predictor that is a measure variable. The default value is 20.

Rapid growth
enables the k-means fast search algorithm and ignores bin ordering. Disable this option to use the greedy search method and respect bin ordering. When enabled, categorical and binned

measure responses use the gain ratio criterion. Otherwise, the information gain criterion is used.

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

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

Assessment

Number of bins
specifies the number of bins to use for calculating the assessment statistics. 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.

Decision Tree Model Display Options

The following display options are available for the decision tree:

General

Plot layout
specifies how the subplots within objects are displayed on the canvas. **Fit** aligns all of the subplots on the canvas automatically. **Stack** displays the subplots as if they are in a slide deck where only one plot is displayed at a time. When **Stack** is specified, a control bar instead of a scroll bar lets you move between the subplots.

Statistic to show
specifies which assessment statistic to display in the model. Possible statistics are **KS (Youden)** and **Misclassification Rate (Event)**. This option is hidden when you have a measure response with more than ten bins. In that case, the average squared error (ASE) is used as the assessment statistic.

Decision Tree / Icicle Plot

Statistic to show
specifies which size statistic to display in the decision tree plot. Possible values are **Count** and **Percent**.

Legend visibility
specifies whether the legend is displayed in the Decision Tree plot.

Leaf Statistics Plot

Y axis
specifies whether the Y axis of the Leaf Statistics plot should show counts or percentages.

Assessment Plots

Plot to show
specifies which assessment plot is displayed. Possible plots are **Lift**, **ROC**, or **Misclassification**. This option is hidden when you have a measure response with more than ten bins. In that case, an assessment plot that displays the average predicted and average observed response values against the binned data is shown.

Y axis
specifies whether a standard Lift plot or a Cumulative lift plot is displayed. This option is hidden when you have a measure response with more than ten bins.

Legend visibility
specifies whether the legend is displayed in the assessment plot.

Growth and Pruning Details

The following additional options are available for the decision tree:

Information Gain and Gain Ratio Calculations

When the **Rapid growth** option is enabled, node splits are determined in part by the 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:

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

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:

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

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:

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

In this equation, T_j represents the observations that contain the jth 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:

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

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

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

Pruning

The **Pruning** option of the decision tree determines how aggressively your decision tree is pruned. The growth algorithm creates a decision tree based on the properties that you specify. The pruning algorithm considers each node to be a root node of its own subtree, starting from the bottom. If the misclassification rate of the subtree is significantly better than the misclassification rate of the root node, then the subtree is kept. If the misclassification rate of the subtree is similar to the misclassification rate of the root node, then the subtree is pruned. In general, smaller decision trees are preferred.

For a smaller value of the **Pruning** option, the difference in the misclassification rates can be relatively small. If the **Pruning** value is larger, then the difference in the misclassification rates must be relatively large.

Variables that are not used in any split can still affect the decision tree, typically due to one of two reasons. It is possible for a variable to be used in a split, but the subtree that contained that split might have been pruned. Alternatively, the variable might include missing values, but the **Missing assignment** option is set to **None**.

Note: If a predictor does not contribute to the predictive accuracy of the decision tree or the contribution is too small, then it is not included in the final, displayed decision tree.

Decision Tree Results

Tree Window

The Tree window contains the decision tree, tree overview, and icicle plot.

TIP 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 the pointer.

The color of the node in the icicle plot indicates the predicted level for that node. When you select a node in either the decision tree or the icicle plot, the corresponding node is selected in the other location. When you select a leaf node, that node is selected in the Leaf Statistics window.

If you zoom in on the decision tree, a bar chart is displayed in each node that shows the distribution of event levels within that node.

The following tasks are available in the Tree window:

- To derive a leaf ID variable, right-click in the Tree window, and select **Derive a leaf ID variable**. This action creates a category variable that contains the leaf ID for each observation. You can use this variable in other objects throughout SAS Visual Analytics.
- Right-click on a leaf node to perform the following:

Split
opens the Split node 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 gain ratio when **Rapid growth** is enabled. It splits the node based on the variable with the best information gain when **Rapid growth** is disabled.

Grow or **Train**
opens the Train node 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 node 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.

Note: Some variables are not available for a split if the contribution is too small or the split would violate the **Leaf size** property. It is possible for there to be no variables available for splitting.

- Right-click on a non-leaf node, and select **Prune** to prune the decision tree at that node. This removes all nodes beneath the selected node and turns that node into a leaf node.

Leaf Statistics

For classification trees, the Leaf Statistics plot displays the distribution of event levels within each leaf node. You can view this information as either within-node percentages or as absolute counts.

Assessment Plot

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.

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 it. 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.

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, the decision tree considers all levels that are not events as equal. A significant number of misclassifications could indicate that your model does not fit the data.

When you have a measure response with more than ten bins, the Assessment plot displays 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 decision tree computes several assessment measures to help you evaluate how well the model fits the data. These assessment measures are available in the object toolbar of the canvas. Click the currently displayed assessment measure to see all available assessment measures.

KS (Youden)
The maximum distance between the ROC curve and the baseline model.

Misclassification Rate (Event)
The misclassification rate of the target event level.

When you have a measure response with more than ten bins, the average squared error (ASE) is used as the assessment statistic.

Details Table

When you click ⤢ in the object toolbar of the canvas, the details table is displayed at the bottom of the canvas. The details table contains the following information:

Node Statistics
Provides summary statistics for each node in the decision tree. Available statistics can include **Node ID**, **Depth**, **Parent ID**, **N Children**, **Type**, **Observations**, **% Observations**, **Percent of Parent**, **N Missing**, **Gain Ratio**, and **Predicted Value**.

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 in any of its parent nodes, then it is listed in the table. Otherwise, the entry is blank.

Lift
Lists the binned assessment results that are used to generate the Lift plot.

ROC
Lists the results that are used to generate the ROC plot.

Misclassification
Provides a summary of the correct and incorrect classifications for the model.

Assessment
Lists the binned assessment results that are used to generate the Assessment plot. This tab is available when you have a measure response with more than ten bins.

Assessment Statistics
Provides the value of any assessment statistic computed for the model.

Working with Generalized Linear Models

Overview of Generalized Linear Models

A generalized linear model 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 generalized linear model 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 distribution that you specify should match the distribution of your response. This 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

How to Create a Generalized Linear Model

To create a generalized linear model, complete the following steps:

1 Drag and drop the icon onto the canvas.

2 Click in the right pane. Specify a single measure variable as the **Response** variable.

3 Specify at least one measure variable for **Continuous effects** field, category variable for **Classification effects**, or an interaction for **Interaction effects**.

4 Optionally, you can specify **Group by**, **Frequency**, **Weight**, or **Offset** variables.

Generalized Linear Model Options

The following options are available for the generalized linear model:

General

Informative missingness

specifies whether the informative missingness algorithm is used. For more information, see Missing Values on page 6.

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, Reciprocal
Geometric	Log, Identity
Inverse Gaussian	Log, Identity, Power(-2)
Negative Binomial	Log, Identity
Normal	Log, Identity
Poisson	Log, Identity

Convergence

Function convergence

specifies the relative convergence criterion for the objective function. When you specify a larger value, the model converges sooner. This reduces the amount of time spent training the model, but it can create a suboptimal model.

Gradient convergence
specifies the relative convergence criterion for the maximum gradient component. When you specify a larger value, the model converges 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 specified criterion is reached. The reason for convergence is provided on the **Convergence** tab of the details table.

Assessment

Number of bins
specifies the number of bins to use in the assessment. 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.

Generalized Linear Model Display Options

The following display options are available for the generalized linear model:

General

Plot layout
specifies how the subplots within objects are displayed on the canvas. **Fit** aligns all of the subplots on the canvas automatically. **Stack** displays the subplots as if they are in a slide deck where only one plot is displayed at a time. When **Stack** is specified, a control bar instead of a scroll bar lets you move between the subplots.

Statistic to show
specifies which assessment statistic to display in the model. Possible statistics are **-2 Log Likelihood**, **AIC**, **AICC**, **ASE**, and **BIC**.

Residual Plot

Use histogram
specifies whether the Residual Plot is a histogram.

X axis
specifies whether the **Linear Predictor** or the **Predicted Value** is plotted in the Residual Plot. If **Use histogram** is selected, then **X axis** specifies which statistic is plotted. Possible statistics are **Residual** and **Standardized Pearson Residual**.

Y axis
specifies which statistic is plotted in the Residual Plot. Possible statistics are **Residual** and **Standardized Pearson Residual**.

Legend visibility
specifies whether the legend is displayed on the Residual Plot.

Assessment Plot

Legend visibility

specifies whether the legend is displayed on the Assessment plot.

Generalized Linear Model Results

Fit Summary Window

The Fit Summary 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.

When your analysis includes a group by variable, the Fit Summary displays a Goodness of Fit plot.

Residual Plot

A Residual Plot shows the relationship between the predicted value of an observation and the residual of an observation. The residual of an observation is the difference between the predicted response value and the actual response value.

When using large data sets, the Residual Plot is displayed as a heat map instead of as a scatter plot. In a heat map, the actual observations are binned, and the color of each point indicates the relative number of observations in that bin. Alternatively, you can plot the residuals in a histogram.

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.

The following options are available in a Residual Plot:

- To change the residual that is plotted, change the value of the **Y axis** or the **X axis** in the **Residual Plot** options.
- To filter out one or more observations, select those observations in the Residual Plot, right-click the Residual Plot, and select **New filter from selection** ⇨ **Exclude selection**.
- To filter in one or more observations, select those observations in the Residual Plot, right-click the Residual Plot, and select **New filter from selection** ⇨ **Include only selection**.
- To examine one or more observations, select those observations in the Residual Plot, right-click the Residual Plot, and select **Show selected**. This opens a data table that contains only the selected observations.

Assessment Plot

For a generalized linear model, the Assessment plot displays 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 generalized linear model computes several assessment measures to help you evaluate how well the model fits the data. These assessment measures are available in the object toolbar of the canvas. Click the currently displayed assessment measure to see all available assessment measures.

-2 Log Likelihood
The likelihood function estimates the probability of an observed sample given the specified 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 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.

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

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.

Details Table

When you click ↙↗ in the object toolbar of the canvas, the details table is displayed at the bottom of the canvas. The details table contains the following information:

Model Information
Gives an overview of the model.

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
Provides 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.

Parameter Estimates
Gives the estimated values for the model parameters.

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*.

Assessment
Lists the binned assessment results that are used to generate the Assessment plot.

Assessment Statistics
Provides the value of any assessment statistic computed for the model.

Working with Linear Regression Models

Overview of Linear Regression Models

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 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.

How to Create a Linear Regression

To create a linear regression, complete the following steps:

1. Drag and drop the icon onto the canvas.
2. Click in the right pane. Specify a single measure variable as the **Response** variable.
3. Specify at least one measure variable for **Continuous effects**, a category variable for **Classification effects**, or an interaction for **Interaction effects**.
4. Optionally, you can specify **Group by**, **Frequency**, or **Weight** variables.

Linear Regression Options

The following options are available for the linear regression:

General

Informative missingness
specifies whether the informative missingness algorithm is used. For more information, see Missing Values on page 6.

Variable selection method
specifies whether variable selection is performed. For more information, see Variable Selection on page 5. Methods available are **None** and **Backward**.

Significance Level
specifies the significance value for an effect to be in the model. For more information about variable selection, see Variable Selection on page 5. This property is available for the **Backward** variable selection methods.

Assessment

Number of bins
specifies the number of bins to use in the assessment. 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.

Linear Regression Model Display Options

The following display options are available for the linear regression:

General

Plot layout
specifies how the subplots within objects are displayed on the canvas. **Fit** aligns all of the subplots on the canvas automatically. **Stack** displays the subplots as if they are in a slide deck where only one plot is displayed at a time. When **Stack** is specified, a control bar instead of a scroll bar lets you move between the subplots.

Statistic to show
specifies which assessment statistic to display in the model. See Fit Statistics on page 36 for more information about the statistics available.

Residual Plot

Use histogram
specifies whether the Residual Plot is a histogram.

Y axis
specifies which statistic is plotted in the Residual Plot. Possible values are **PRESS**, **Residual**, **Studentized Deleted Residual**, and **Studentized Residual**.

X axis
specifies which statistic is plotted in the Residual Plot. This option is available when **Use histogram** is selected. Possible values are **PRESS**, **Residual**, **Studentized Deleted Residual**, and **Studentized Residual**.

Legend visibility
specifies whether the legend is displayed on the Residual Plot.

Influence Plot

Plot to show
specifies whether the Influence Plot is displayed.

Use histogram
specifies whether the Influence Plot is a histogram.

X axis
specifies which statistic is plotted in the Influence Plot. Possible values are **Cook's D**, **Covariance Ratio**, **DFFITS**, **Leverage**, and **Likelihood Displacement**.

Assessment Plot

Legend visibility
specifies whether the legend is displayed in the Assessment Plot.

Linear Regression Results

Fit Summary Window

The Fit Summary 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.

When your analysis includes a group by variable, the Fit Summary displays a bar chart that ranks the groups by goodness of fit.

Residual Plot

A Residual Plot shows the relationship between the predicted value of an observation and the residual of an observation. The residual of an observation is the difference between the predicted response value and the actual response value.

When using large data sets, the Residual Plot is displayed as a heat map instead of as a scatter plot. In a heat map, the actual observations are binned, and the color of each point indicates the relative number of observations in that bin. Alternatively, you can plot the residuals in a histogram.

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.

The following options are available in a Residual Plot:

- To change the residual that is plotted, change the value of the **Y axis** or the **X axis** option in the **Residual Plot** options.
- To filter out one or more observations, select those observations in the Residual Plot, right-click the Residual Plot, and select **New filter from selection** ⇨ **Exclude selection**.
- To filter in one or more observations, select those observations in the Residual Plot, right-click the Residual Plot, and select **New filter from selection** ⇨ **Include only selection**.
- To examine one or more observations, select those observations in the Residual Plot, right-click the Residual Plot, and select **Show selected**. This opens a data table that contains only the selected observations.

Assessment Plot

For a linear regression, the Assessment plot displays 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.

Influence Plot

The Influence Plot displays several measurements that are computed for each observation. A histogram can also be displayed. When the input data contains a large number of observations, the observations are binned. Use these measurements to help identify outliers and other data points that greatly affect the predicted regression model.

To change the computed measurement that is plotted, right-click the measurement name on the X axis, and select a new measurement.

The following options are available in the Influence Plot:

- To change the residual that is plotted, change the value of the **X axis** option in the **Influence Plot** options.
- To filter out one or more observations, select those observations in the Influence Plot, right-click the Influence Plot, and select **New filter from selection** ⇨ **Exclude selection**.
- To filter in one or more observations, select those observations in the Influence Plot, right-click the Influence Plot, and select **New filter from selection** ⇨ **Include only selection**.
- To examine one or more observations, select those observations in the Influence Plot, right-click the Influence Plot, and select **Show selected**. This opens a data table that contains only the selected observations.

Fit Statistics

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

Adjusted R-Square

The adjusted R-square 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. 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.

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

F Value of 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 square 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.

R-Square
The R-square value is an indicator of how well the model fits the data. R-square 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.

Details Table

When you click ↙↗ in the object toolbar of the canvas, the details table is displayed at the bottom of the canvas. The details 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.

Overall ANOVA
Provides the analysis of variance results for the model, error, and corrected total.

Fit Statistics
Lists all of the fit statistics.

Parameter Estimates
Gives the estimated values for the model parameters.

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*.

Assessment
Lists the binned assessment results that are used to generate the Assessment plot.

Assessment Statistics
Provides the value of any assessment statistic computed for the model.

Working with Logistic Regression Models

Overview of Logistic Regression Models

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.

How to Create a Logistic Regression

To create a logistic regression, complete the following steps:

1. Drag and drop the icon onto the canvas.
2. Click in the right pane. Specify a single category variable as the **Response** variable.
3. Specify at least one measure variable for **Continuous effects**, a category variable for **Classification effects**, or an interaction for **Interaction effects**.
4. Optionally, you can specify **Group by**, **Frequency**, **Weight**, or **Offset** variables.

Logistic Regression Options

The following options are available for the logistic regression:

General

Event level
enables you to choose the event level of interest. When your category response variable contains more than two levels, SAS Visual Statistics treats all observations in the level of interest as an event and all other observations as nonevents.

Informative missingness
specifies whether the informative missingness algorithm is used. For more information, see Missing Values on page 6.

Variable selection method
specifies whether variable selection is performed. For more information, see Variable Selection on page 5. Methods available are **None** and **Backward**.

Significance Level
specifies the significance value for an effect to be in the model. For more information about variable selection, see Variable Selection on page 5 . This property is available for the **Backward** variable selection method.

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 listed below:

- **Logit** (default) specifies the inverse of the cumulative logistic distribution function.

$$g\left(M\right) = \log\left(\frac{M}{1-M}\right)$$

- **Probit** specifies the inverse of the cumulative standard normal distribution function.

$$g\left(M\right) = \frac{1}{\Phi(M)}$$

Convergence

Function convergence
specifies the convergence criterion for the objective function. When you specify a larger value, the model converges sooner. This reduces the amount of time spent training the model, but it can create a suboptimal model.

Gradient convergence
specifies the convergence criterion for the maximum gradient component. When you specify a larger value, the model converges 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 specified criterion is reached. The reason for convergence is provided on the **Convergence** tab of the details table.

Assessment

Number of bins
specifies the number of bins to use. You must specify an integer value between 5 and 100. By default, measure variables are grouped into 20 bins.

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.

Logistic Regression Model Display Options

The following display options are available for the logistic regression:

General

Plot layout
specifies how the subplots within objects are displayed on the canvas. **Fit** aligns all of the subplots on the canvas automatically. **Stack** displays the subplots as if they are in a slide deck where only one plot is displayed at a time. When **Stack** is specified, a control bar instead of a scroll bar lets you move between the subplots.

Statistic to show
specifies which assessment statistic to display in the model. See Fit Statistics on page 43 for more information about the statistics available.

Residual Plot

Use histogram
specifies whether the Residual Plot is a histogram.

X axis
specifies whether the **Linear Predictor** or the **Predicted Probability** is plotted in the Residual Plot. If **Use histogram** is selected, then **X axis** specifies which statistic is plotted. Possible statistics are **Deviance Residual**, **Pearson Residual**, **Residual**, and **Standardized Pearson Residual**.

Y axis
specifies which statistic is plotted in the Residual Plot. Possible statistics are **Deviance Residual**, **Pearson Residual**, **Residual**, and **Standardized Pearson Residual**.

Legend visibility
specifies whether the legend of the Residual Plot is displayed.

Influence Plot

Plot to show
specifies whether to display the Influence Plot.

Use histogram
specifies whether the Influence Plot is a histogram. This option is available when **Influence Plot** is selected.

X axis
specifies which statistic is plotted in the Influence Plot. Possible statistics are **CBAR**, **Deviance Change**, **Likelihood Displacement**, and **Pearson Change**.

Assessment Plots

Plot to show
specifies which assessment plot is displayed. Possible values are **Lift**, **ROC**, or **Misclassification**.

Y axis
specifies whether a standard Lift plot or a Cumulative Lift plot is displayed.

legend visibility
specifies whether the legend of the Lift plot or ROC plot is displayed.

Logistic Regression Results

Fit Summary Window

The Fit Summary 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.

When your analysis includes a group by variable, the Fit Summary displays a Goodness of Fit plot.

Residual Plot

A Residual Plot shows the relationship between the predicted value of an observation and the residual of an observation.

For an event with a predicted probability *p*, the residual is

$$\frac{e^p + 1}{e^p}$$

For a non-event with a predicted probability *p*, the residual is

$$-(e^p + 1)$$

When using large data sets, the Residual Plot is displayed as a heat map instead of as a scatter plot. In a heat map, the actual observations are binned, and the color of each point indicates the relative number of observations in that bin. Alternatively, you can plot the residuals in a histogram.

A logistic regression always displays a two-tailed distribution due to the difference in the event and non-event residual calculations. The upper tail displays the event residuals and the residual values should approach zero as the predicted probability increases. The bottom tail displays the non-event residuals and the residual values should approach zero as the predicted probability decreases. This chart might indicate the direction (event or non-event) in which certain bad data points are biased.

The following options are available in a Residual Plot:

- To change the residual that is plotted, change the value of the **Y axis** or **X axis** option in the **Residual Plot** options.
- To filter out one or more observations, select those observations in the Residual Plot, right-click the Residual Plot, and select **New filter from selection** ⇨ **Exclude selection**.
- To filter in one or more observations, select those observations, in the Residual Plot, right-click the Residual Plot, and select **New filter from selection** ⇨ **Include only selection**.
- To examine one or more observations, select those observations in the Residual Plot, right-click the Residual Plot, and select **Show selected**. This opens a data table that contains only the selected observations.

Assessment Plot

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.

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 it. 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.

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, the logistic regression considers all levels that are not events as equal. A significant number of misclassifications could indicate that your model does not fit the data.

Influence Plot

The Influence Plot displays several measurements that are computed for each observation. A histogram can also be displayed. When the input data contains a large number of observations, the observations are binned. Use these measurements to help identify outliers and other data points that greatly affect the predicted regression model.

To change the computed measurement that is plotted, right-click the measurement name on the X axis, and select a new measurement.

The following options are available in the Influence Plot:

- To change the residual that is plotted, change the value of the **X axis** option in the **Influence Plot** options.
- To filter out one or more observations, select those observations in the Influence Plot, right-click the Influence Plot, and select **New filter from selection** ⇨ **Exclude selection**.
- To filter in one or more observations, select those observations in the Influence Plot, right-click the Influence Plot, and select **New filter from selection** ⇨ **Include only selection**.

- To examine one or more observations, select those observations in the Influence Plot, right-click the Influence Plot, and select **Show selected**. This opens a data table that contains only the selected observations.

Fit Statistics

The logistic regression computes several assessment measures to help you evaluate how well the model fits the data. These assessment measures are available in the object toolbar of the canvas. Click the currently displayed assessment measure to see all 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. This value is -2 times the log likelihood. Smaller values are preferred.

AIC
Akaike's information criterion. Smaller values indicate better models. 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.

KS (Youden)
The maximum distance between the ROC curve and the baseline model.

Max-rescaled R-square
The observed R-square value divided by the maximum attainable R-square 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.

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

Misclassification Rate (Event)
The misclassification rate of the target event level.

Details Table

When you click ↙↗ in the object toolbar of the canvas, the details table is displayed at the bottom of the canvas. The details table contains the following information:

Model Information
Gives an overview of the model.

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.

Response Profile
Displays the event and nonevent counts.

Iteration History
Provides the function and gradient iteration 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.

Parameter Estimates
Gives the estimated values for the model parameters.

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*.

Lift
Lists the binned assessment results that are used to generate the Lift plot.

ROC
Lists the results that are used to generate the ROC plot.

Misclassification
Provides a summary of the correct and incorrect classifications for the model.

Assessment Statistics
Provides the value of any assessment statistic computed for the model.

Working with Model Comparison

Overview of Model Comparison

The model comparison 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 two models are trained before you can perform a comparison.

Before performing a model comparison, ensure that all models are initialized and updated. If auto-refresh 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.

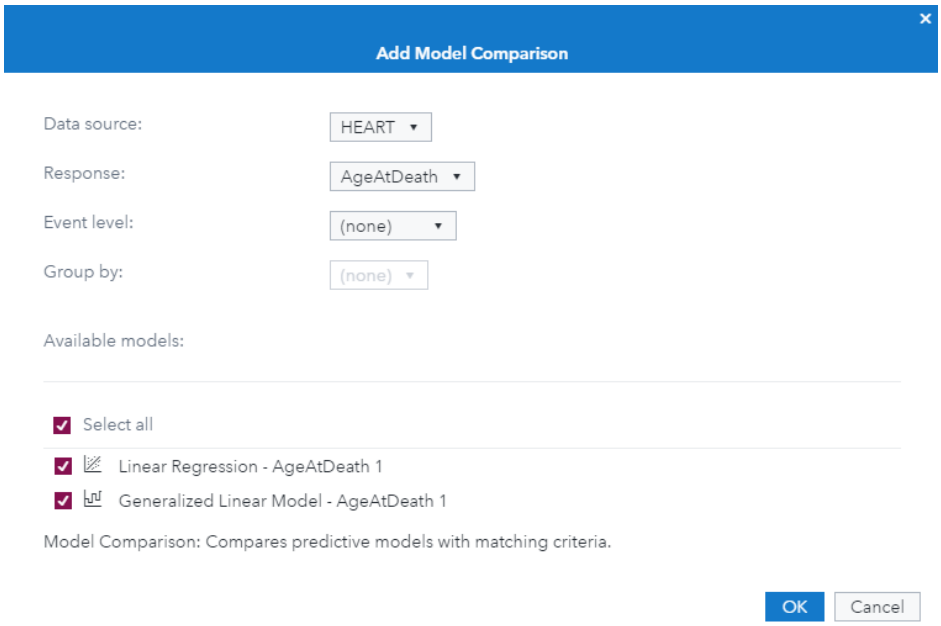
When you change a model after a comparison has been created, changes are not carried over to the model comparison.

Using the Model Comparison

To create a model comparison, complete the following steps:

1 Drag and drop the icon onto the canvas.

2 In the Add Model Comparison window, specify the **Data source**, **Response**, **Event level**, and **Group by**.



3 At the bottom of the window, select all of the models that you want to compare. You must specify 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 Options

The following options are available for model comparison:

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 Display Options

The following display options are available for model comparison:

Comparison Display

Plot layout

specifies how the subplots within objects are displayed on the canvas. **Fit** aligns all of the subplots on the canvas automatically. **Stack** displays the subplots as if they are in a slide deck where only one plot is displayed at a time. When **Stack** is specified, a control bar instead of a scroll bar lets you move between the subplots.

Fit Statistics

Fit statistic

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

Assessment Plots

Show all plots

specifies whether all assessment plots are displayed simultaneously. This option is available only if the models have a category response.

Plot to show

specifies which assessment plot is displayed.

Y axis

specifies which statistic is plotted in the assessment plot.

Legend visibility

specifies whether the legend of the specified assessment plot is displayed.

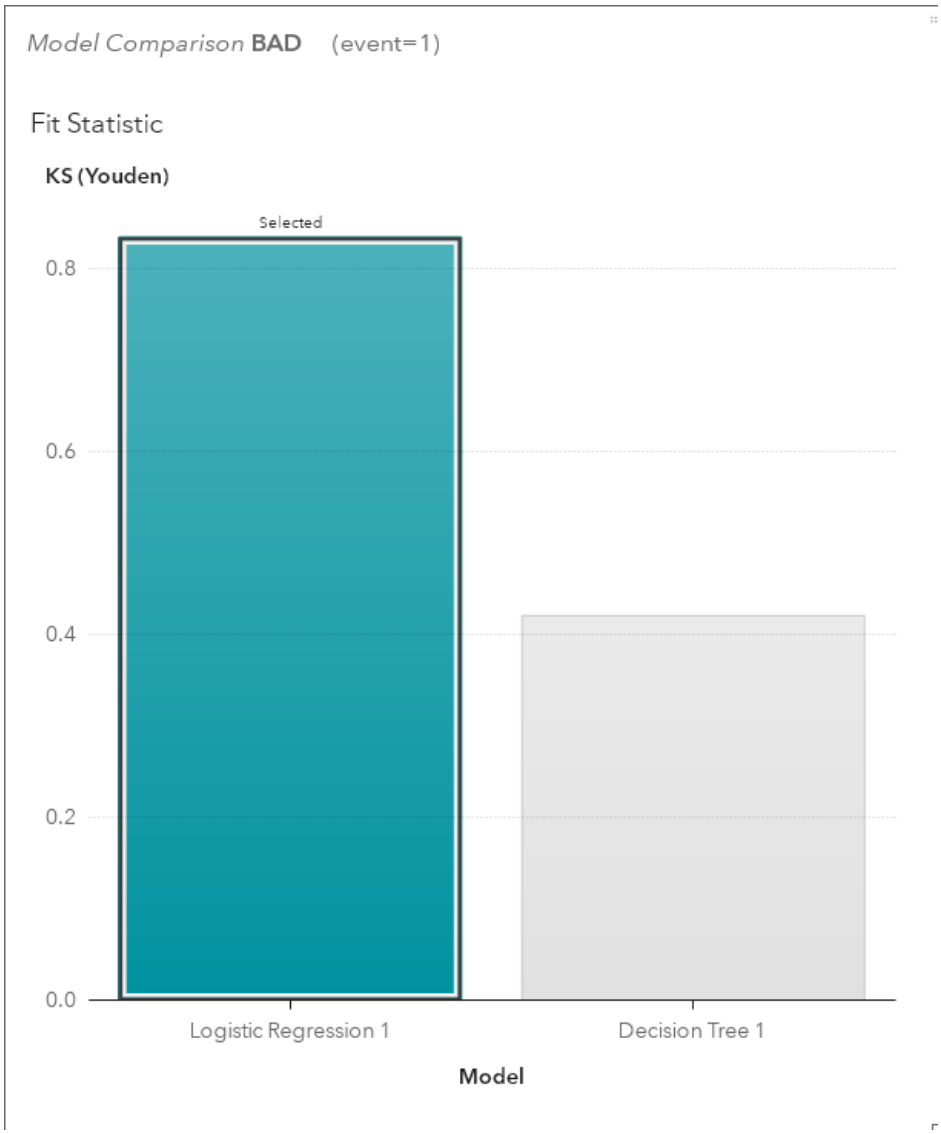
Model Comparison Results

Assessment Plot

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 Statistics

The Fit Statistic plot displays the criterion specified in the **Fit statistic** option. In the following image, the KS (Youden) value is plotted for a logistic regression and decision tree. The champion model is selected in the plot.



Details Table

When you click ↙↗ from the object toolbar on the canvas, the details table is displayed at the bottom of the canvas. The details 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 prefers based on the criterion specified in the **Fit statistic** property. However, the statistics listed in the details table can differ from those listed in the **Fit statistic** option.

Variable Importance

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