SAS® Visual Analytics 8.5: 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, except the generalized additive model and the nonparametric logistic regression. 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 Viya 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**
- **Automatically convert measure variables with two levels to category** applies when a data set is first opened in SAS Visual Analytics. You can manually convert the category back to a measure variable. This setting does not apply if the data source was opened with a default data view applied.
- **Default statistic** specifies your default statistic preference for **Category Response** and **Measure Response** variables to use for all models and model comparison.

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 20 segments the input data into groups that share similar features.
- Decision tree on page 24 creates a hierarchical segmentation of the input data based on a series of rules applied to each observation.
- Generalized additive model on page 33 is an extension of the generalized linear model that allows spline terms in order to predict an interval response.
- Generalized linear model on page 38 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 44 attempts to predict the value of an interval response as a linear function of one or more effect variables.
- Logistic regression on page 48 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.
- Nonparametric logistic regression on page 55 is an extension of the logistic regression model that allows spline terms to predict a binary response.

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, group by variables, partition variables, and stratification variables in generated partitions.

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.

To create an interaction effect, do one of the following:

- In the **Data** pane, select the variables of interest. Right-click one of the selected variables, and then select **New interaction effect**. Select the interaction type that you are interested in.
- At the top of the **Data** pane, select **New data item**, and then select **Interaction effect**. In the New Interaction Effect window, 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.

Spline Effects

SAS Visual Statistics enables you to create one-dimensional or two-dimensional splines from any measure variable. A spline function is a piecewise polynomial function in which the individual polynomials have the same degree and connect smoothly at certain points. Spline functions are used to fit smooth curves to a wide variety of data.

SAS Visual Statistics creates thin-plate regression splines that are based on thin-plate smoothing splines. For more information, see the PROC GAMPL documentation in the *SAS/STAT User's Guide*.

To create a spline effect, do one of the following:

- In the **Data** pane, right-click the variable of interest, and then select **New spline effect**.
- At the top of the **Data** pane, select **New data item**, and then select **Spline effect**. In the New Spline Effect window, specify the variables of interest and the spline type.

Splines are required inputs to the generalized additive model and the nonparametric logistic regression model.

Variable Selection

Variable selection is the process of reducing the number of input variables to include just the most significant variables. The linear regression, logistic regression, and generalized linear models 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 methods:

- **Forward** — Candidate effects are added one at time to the model based on how much each effect improves the model. Variable selection continues until no effects are remaining or no effect significantly improves the model.
- **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.
- **Fast Backward** — Available for logistic regression models, this technique uses a numeric shortcut to compute the next selection iteration quicker than backward selection.
- **Stepwise** — A combination of forward and backward selection. Candidate effects are added one at a time based on their significance. However, at each step, an effect might be removed if it is deemed not significant.

- **Lasso** — Adds and removes candidate effects based on a version of ordinary least squares, where the sum of the absolute regression coefficients is constrained. Multiple effects can enter the model in a single step.
- **Adaptive Lasso** — Available for linear regressions, this is a modification to lasso where selection weights are applied to each of the parameters used to create the lasso constraint.

Partition Data and Validate the Model

A partition variable is used to perform model validation for the decision tree, generalized additive model, generalized linear model, linear regression, logistic regression, and nonparametric logistic regression. The observations whose partition column value corresponds to training are used for preliminary model building. The observations that correspond to validation are fed through the model, and the results can be used to prevent overfitting the training data and to compare models. In addition, for the decision tree object, you can use the validation data for pruning. If a test partition is specified, the observations that correspond to testing are used for a final assessment of the model.

If your data source includes a column named *_PartInd_*, then SAS Visual Analytics automatically assigns it as a partition variable. In the **Data** pane, right-click the *_PartInd_* variable and select **Edit partition** to change the default partition assignments.

To specify a partition variable based on a variable that already exists in your data:

1. In the **Data** pane, select a partition variable.
2. Right-click the partition variable, and select **New partition**.
3. In the New Partition window, specify the **Number of partitions**. If you select **2**, specify which variable value corresponds to the **Training data value** and which variable value corresponds to the **Validation data value**. If your variable has a test partition, select **3**, and then also specify which variable value corresponds to the **Test data value**.
4. Click **OK**.

Note: Only category variables with fewer than six distinct values are eligible for use as a partition variable. Measure variables with two to five distinct levels can also be used as partition variables. A copy of the measure variable is automatically created and converted to a category variable before being assigned as a partition variable.

If you do not have a partition variable in your data, complete the following steps so that SAS Visual Statistics can generate a partition variable:

1. In the **Data** pane, click **New data item**, and select **Partition**.
2. In the New Partition window, specify the **Name**, **Sampling method**, **Stratify by**, **Number of partitions**, **Training partition sampling percentage**, **Testing partition sampling percentage**, and the **Random number seed**. Some options are available based on your selections.

   The **Number of partitions** option indirectly specifies whether a test partition is included. When you select **3** for this option, a training partition, validation partition, and testing partition are created. When you select **2**, only a training partition and validation partition are created. At least 1% of the data must be used for the validation data. Therefore, the sum of the training partition and testing partition must be strictly less than 100.
3. Click **OK**.

Note: When you use a generated partition variable, results are nondeterministic. If you close and reopen a report with a generated partition variable, subsequent results might not match the initial

results. When you select the **Random number seed** option and specify a random seed, you might get nondeterministic results. Such results are due to the difference in data distribution and computational threads or to the walker used to sample the partition column.

You cannot use an interactive decision tree or selection filters when a generated partition variable is assigned.

When you duplicate or change an object to a cluster or factorization machine object, the partition variable is dropped. However, you can create a partition variable for the cluster or factorization machine object.

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 26 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.

Integration with Model Studio

SAS Visual Analytics enables you to transfer certain analytical models from SAS Visual Analytics to Model Studio. To move a model from SAS Visual Analytics to Model Studio, click the **Create Pipeline** button.

This action creates a new project in Model Studio that contains the following elements:

- the active data set

- score code to apply all data processing, filtering, and transformations
- score code to run the model that was exported

The properties of the modeling nodes can be edited, and subsequently models can be retrained in Model Studio. Right-click the modeling node, and then select **Enable properties** to edit the node. Model interpretability properties are always automatically enabled. You can add and delete nodes in this pipeline as in any other Model Studio pipeline. You can use the Model Studio model comparison and pipeline comparison tools to evaluate your transferred models against any new models.

At this time, the supported models are Bayesian network, decision tree, forest, generalized linear model, gradient boosting, linear regression, logistic regression, neural network, and support vector machine. There also exist exceptions within these models:

- You cannot copy a decision tree with a binned measure response.
- You cannot copy a generalized linear model, linear regression, or logistic regression that uses a frequency, weight, offset, or group by variable.
- You cannot copy a logistic regression with a non-binary response variable.
- You cannot copy a neural network with a weight variable.

There are a few caveats when transferring a model from SAS Visual Analytics to Model Studio.

- In order to add a SAS Visual Analytics model to an existing Model Studio project, the target variable name, type, and event level must match. The data node in the existing Model Studio project must have also been previously run.
- Instead of using the variable name that exists in the original data set, SAS Visual Analytics prefers to use the variable label. However, Model Studio prefers to use the variable name as it exists in the original data set. Therefore, if the variable names and variable labels in your input data are different, you might experience some unexpected naming issues when a model is transferred. Model Studio displays both the variable name and the variable label in the **Variables table** layout of the **Data** pane.
- SAS Visual Analytics creates a custom name for a target variable. This new variable is indicated with a label in Model Studio.
- When you are exporting from the model comparison object, only the champion model is exported.
- Partition variables must be numeric variables that contain only the values 0 for training data, 1 for validation data, and 2 for testing data. The testing data can be omitted. If the partition information from SAS Visual Analytics differs from Model Studio, then the partitioning of the Model Studio project is used.
- Category target variables cannot contain any special characters, including a comma, semicolon, open parenthesis, or close parenthesis. Special characters in the target variable of a Model Studio pipeline cause model creation to fail.
- You cannot transfer a model from Model Studio to SAS Visual Analytics. However, you can copy the input data to SAS Visual Analytics for exploration and visualization.
- Certain actions that create a data item in SAS Visual Analytics are performed in the **Visual Data Preparation** node in Model Studio. For example, when you derive a cluster ID item in SAS Visual Analytics, a data item is created in the **Data** pane. If you specify this created data item in a SAS Visual Analytics model that is transferred to Model Studio, it does not appear in the **Data** pane of your project. Instead, it is re-created when the **Visual Data Preparation** node runs.
- If you enable the properties of a transferred **SAS Visual Analytics** modeling node in **Model Studio**, the model is retrained, which might yield new results. Once properties have been enabled, the model cannot be switched back to use the original **SAS Visual Analytics** score code. Model interpretability properties do not require retraining the model. Therefore, they are automatically enabled and do not affect the original **SAS Visual Analytics** score code.

- You cannot save your pipeline as a template to The Exchange.
- If you transfer a gradient boosting, decision tree, or forest model, the value for the **Maximum levels** property that is mapped to Model Studio is one fewer than the value that you specified in SAS Visual Analytics.

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.

Fit Statistics

Several assessment measures are computed 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 measures. The following are all fit statistics that are available. The fit statistics that are available for a particular modeling object depend on the model and whether the response is a category or measure data item.

The C statistic, Gini, Gamma, and Tau are all used with ordinal variables. The predicted mean score of an observation is the sum of the ordered response values minus one, weighted by the corresponding predicted probabilities for that observation. A pair of observations with different observed responses is said to be concordant if the observation with the lower ordered response value has a lower predicted mean score than the observation with the higher ordered response value. If the observation with the lower ordered response value has a higher predicted mean score than the observation with the higher ordered response value, then the pair is discordant. If the pair is neither concordant nor discordant, it is a tie.

Let N be the sum of observation frequencies in the data. Suppose that there are a total of t pairs with different responses: n_c of them are concordant, n_d of them are discordant, and $t - n_c - n_d$. The C statistic, Gini, Gamma, and Tau can be defined as the following:

- C statistic = $(n_c + 0.5(t - n_c - n_d)) / t$

  The C statistic is an estimate of the area under the ROC curve when you have a binary response.
- Gini (Somers' D) = $(n_c - n_d) / t$
- Goodman-Kruskel's Gamma = $(n_c - n_d) / (n_c + n_d)$
- Kendall's Tau = $(n_c - n_d) / (0.5N(N - 1))$

-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.

Adjusted R-Square
The adjusted R-square value attempts to account for the addition of more effect variables. Values are in the range 0–1. 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.

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.

Cumulative % Captured

The cumulative number of events observed up to and including the specified percentile bin divided by the total number of events, sorted in descending order of the predicted event probabilities.

Cumulative % Events

The cumulative number of events observed up to and including the specified percentile bin divided by the total number of observations in those bins, sorted in descending order of the predicted event probabilities.

Cumulative Lift

The lift, calculated using all of the data up to and including the current percentile bin, sorted in descending order of the predicted event probabilities.

F1 Score

The harmonic mean of the sensitivity and the precision. Sensitivity is the true positive rate. Precision is the ratio of observations that the model correctly identified as an event to all observations that the model identified as an event. F1 scores closer to 1 indicate better models.

FDR

The false discovery rate (FDR) is the expected proportion of false positives.

FPR

The false positive rate (FPR).

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.

GACV

Generalized approximate cross validation (GACV) is a technique to estimate the smoothing parameter in penalized likelihood regression.

Gain

The percentage increase of the event rate up to the specified percentile bin of the data (in descending order of the predicted event probabilities) over the baseline event rate.

GCV

Generalized cross validation (GCV) is a cross validation approximation technique that does not require retraining the model on a subset of the input data. Smaller values are preferred.

KS (Youden)

The maximum distance between the ROC curve and the baseline model.

Lift
The ratio of the percent of captured responses within each percentile bin to the average percent of responses for the 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 are in the range 0–1. Values closer to 1 are preferred.

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.

Misclassification Rate
The misclassification rate of the model.

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

Observed Average
The average of the observed response values in the specified percentile bin of the data, sorted in descending order of the predictions.

R-Square
The R-square value is an indicator of how well the model fits the data. R-square values are in the range 0–1. 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.

SSE
The sum of squared errors (SSE) is the sum of the squared differences between the observed values and the predicted values.

UBRE
Unbiased risk estimator (UBRE) is a scaled version of Akaike's information criterion.

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. All variables used by the model in any capacity are included in the score code. This includes interaction terms and group by variables.

To generate model score code, right-click in 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 word processing program.

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.

Generalized additive models and nonparametric logistic regression models export an analytic store table in addition to SAS code. See Export and Score a Model for more information.

Register Models and Access the Model Repository

To register a model, right-click in the model canvas, and select **Register model**. In the Register Model window, enter a name for your model, and then click **OK**.

When a model is registered, it appears in your model repository, which you can access through SAS Environment Manager. To access the model repository:

1. In the upper left corner of the window, click the ☰ icon and select **Manage Environment**.
2. In the left pane, click ▤ to access the Content window.
3. Below **SAS Content**, select **Model Repositories**, and then select **VARepository**. The **VARepository** folder is where any SAS Visual Statistics models and SAS Visual Data Mining and Machine Learning models that you register are stored.

If SAS Model Manager is installed on the system, then the model is registered in SAS Model Manager. SAS Model Manager is used to store and organize models in a common model repository. It allows for model governance and model change control over time. To view the selected models in SAS Model Manager, click ☰ and select **Manage Models**. For more information about SAS Model Manager, see *SAS Model Manager: User's Guide*

If the CASHostAccountRequired group has been created, then Read and Write permissions must be configured for the ModelStore caslib and the ModelPerformanceData caslib. These permissions enable members of the CASHostAccountRequired group to register analytic store models. For more information about how to configure permissions, see File System Directory Permissions.

Note: If you update and re-register a model during the same user session, you must refresh the model in **SAS Model Manager** to see the updates.

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. You can also create variables with information about the leaves in a decision tree and about generated clusters. After these variables are created, they can be used in any other object, including other predictive models.

To create variables that contain the prediction information:

1. Create a decision tree model, linear regression model, generalized linear model, generalized additive model, logistic regression, or nonparametric logistic regression model.
2. Right-click on the model canvas, and select **Derive predicted**.
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, generalized linear models, generalized additive models, and regression decision trees. Probability

values are available for logistic regressions, nonparametric logistic regressions, and classification decision trees.

For classification decision trees, you can specify whether you want to derive predicted values and probability values for all levels of the response or for only the event level.

4 Click **OK**. The derived variables appear in a new section of the **Data** pane.

To create a leaf ID variable:

1 Create a decision tree model.

2 Right-click on the model canvas and select **Derive a leaf ID variable**.

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

4 Click **OK**. The leaf ID variable appears in a new section on the **Data** pane.

To create cluster ID variables:

1 Generate clusters with the Cluster object.

2 Right-click on the canvas and select **Derive cluster ID items**.

3 In the New Cluster ID Items window, enter a name for the **Cluster ID**, **Distance from centroid value**, **Interval distance from centroid value**, **Nominal distance from centroid value**, and **Standardized distance from centroid value** variables. The variable types that were used to create the cluster determine which cluster variables are available.

4 Click **OK**. The cluster variables appear in a new section on the **Data** pane.

Note: The variables that are available in the New Cluster ID Items window vary based on the type (category or measure) of the variables that are assigned to the Cluster object.

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.

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

Predicted values

For linear regressions, generalized linear models, generalized additive models, and regression decision trees, 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, nonparametric logistic regressions, and classification decision trees, this is the decision generated by the model based on the calculated probability and **Prediction cutoff** parameter. For logistic regressions, observations are classified into either the event level of interest, not in the event level of interest, or missing. For classification decision trees, you can specify whether you want to classify observations into all levels of the response or for only the event level.

Residual values

The computed residual for each observation. This value is available for linear regressions, generalized linear models, generalized additive models, and regression decision trees.

Probability values

The computed probability for each observation. 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.

For classification decision trees, you can specify whether you want the probability values for all levels of the response or for only the event level.

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.

Interval distance from centroid value
The distance from the observation to the centroid of the cluster that contains the observation based only on the measure variables.

Nominal distance from centroid value
The distance from the observation to the centroid of the cluster that contains the observation based only on the category variables.

Standardized distance from centroid value
The distance from the observation to the centroid of the cluster that contains the observation after measure variables have been standardized.

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 Drive.

1 From **SAS Drive**, click ☰ and select **Explore and Visualize**. SAS Visual Analytics opens, and you can open a data source, create a new report, or load a report.

2 Click the **Start with Data** button in the home pane to load your data. The Choose Data window appears.

3 On the **Import** tab, click **Local files**, and then click **Local file**. Navigate to the location where you saved **heart.csv**, select **heart.csv**, and click **Open**.

4 In the Choose Data window, click **Import Item**. After the table is successfully imported, click **OK**.

5 Rename the project by saving it. By default, the report is named `Report 1`, which is displayed at the top of the page.

   In the upper right corner of the window, click ⋮, and then select **Save**. 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`.

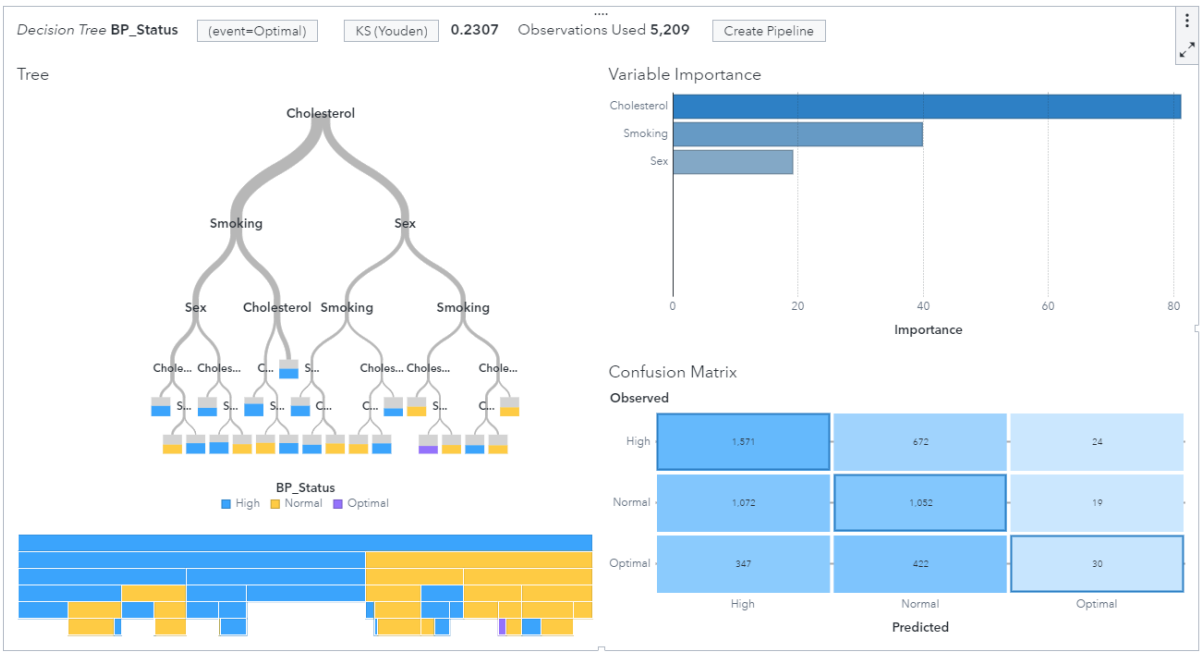
For more information about reports, see *SAS Visual Analytics: Designing Reports*.

Create a Decision Tree

1 In the left pane, click ⟋ to select an object. Drag the icon onto the canvas to create a decision tree.

2 Click the **Assign Data** button in the middle of the report canvas. For **Response**, click **Add**, and select **BP_Status**.

3 For **Predictors**, click **Add**, and select **Sex**, **Cholesterol**, and **Smoking**. Click **OK**, and then click **Close**. The decision tree automatically updates.



4 Observe the confusion matrix. The values on the diagonal of the matrix represent the frequency at which the model correctly predicted the observed value. For example, 1,571 observations with a blood pressure status of high were correctly classified as high by the model. The off-diagonal values represent the frequency at which the model incorrectly predicted the observed value. For

example, 422 observations with a blood pressure status of optimal were incorrectly classified as normal by the model.

5 Click ↙↗ to enter maximize mode.

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

6 Click ↘↖ 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 to save the report.

Create a Linear Regression

1 Click + to add a new page.

2 In the left pane, click to select an object. Drag the icon onto the canvas to create a linear regression.

3 In this example, the variable of interest is **AgeAtDeath**.

Click in the right pane. For **Response**, click **Add**, and select **AgeAtDeath**.

4 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. In this example, you create an interaction effect, and then you create 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. In the **Data** pane, select **Diastolic**. The blue dot next to **Systolic** indicates that the variables **Diastolic** and **Systolic** are also highly correlated with each other.

5 Create an interaction term:

a In the **Data** pane, select **Diastolic** and **Systolic**.

b Right-click **Systolic**, and then select **New interaction effect** ⇨ **Add one interaction effect**.

The interaction term **Diastolic*Systolic** appears in the **Interaction Effect** group of the **Data** pane.

6 Drag **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 window, click ⋮, select **Interface options**, and then 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.

7 In the **Data** pane, select **Smoking_Status**, **Cholesterol**, **Weight**, and **Leaf ID 1**. Drag these variables onto the canvas. The linear regression updates to include these effects.

8 In the right pane, click . 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.

9 Save the report.

Create a Generalized Linear Model

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

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

3 Click 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 , and select **View measure details**. In the Measure Details window, select **AgeAtDeath**.

4 Notice that **AgeAtDeath** is not normally distributed 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**.

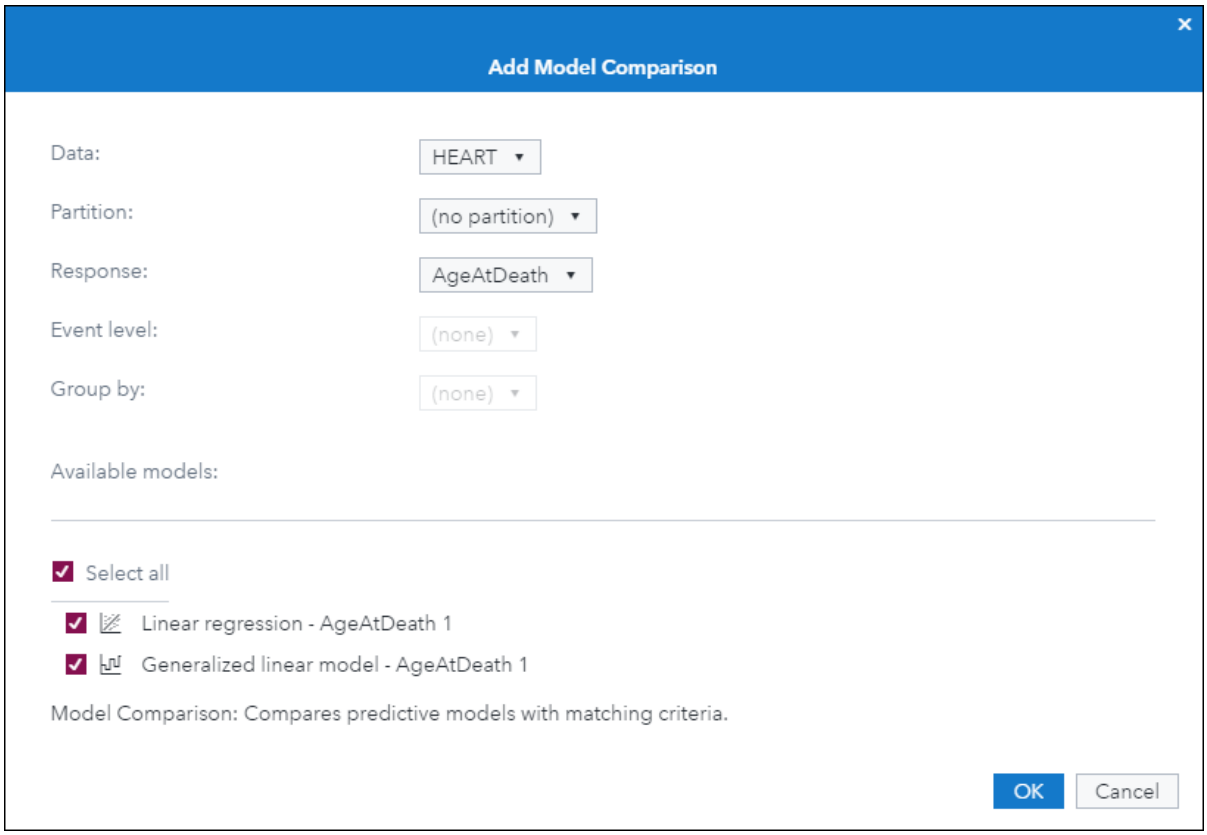
6 (Optional) Repeat this example with different distributions and link functions and compare their performances, familiarizing yourself with SAS Visual Statistics.

7 Save the report.

Perform a Model Comparison

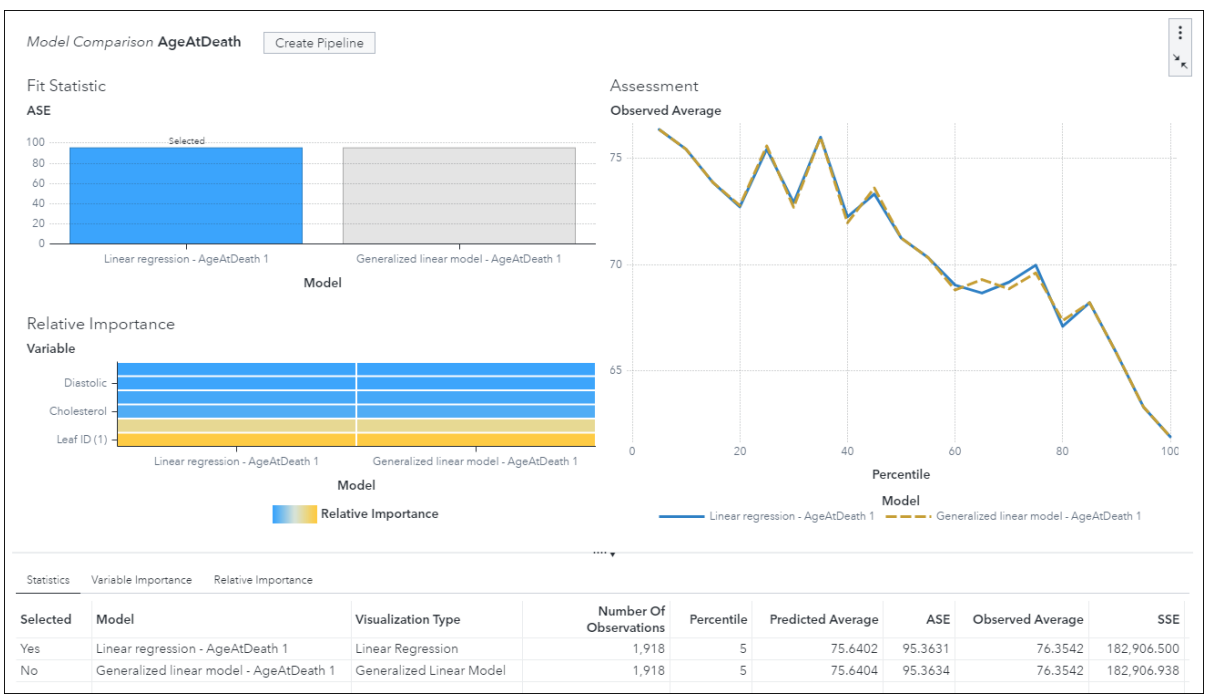
1 Click + to add a new page.

2 In the left pane, click to select an object. Drag the icon onto the canvas to create a model comparison.



Set **Response** to **AgeAtDeath**. The **Event level** and **Group by** options are 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 click **OK**.



Selected	Model	Visualization Type	Number Of Observations	Percentile	Predicted Average	ASE	Observed Average	SSE
Yes	Linear regression - AgeAtDeath 1	Linear Regression	1,918	5	75.6402	95.3631	76.3542	182,906.500
No	Generalized linear model - AgeAtDeath 1	Generalized Linear Model	1,918	5	75.6404	95.3634	76.3542	182,906.938

By default, the fit statistic average square error, **ASE**, is used to compare the models. The **Fit statistic** option enables you to change the fit statistic that 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.

Note: Click ↙↗ to enter maximize mode and to view the summary table.

4 The champion model is **Linear regression - AgeAtDeath 1** 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.

   a Click ↘↖ to exit maximize mode.

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

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

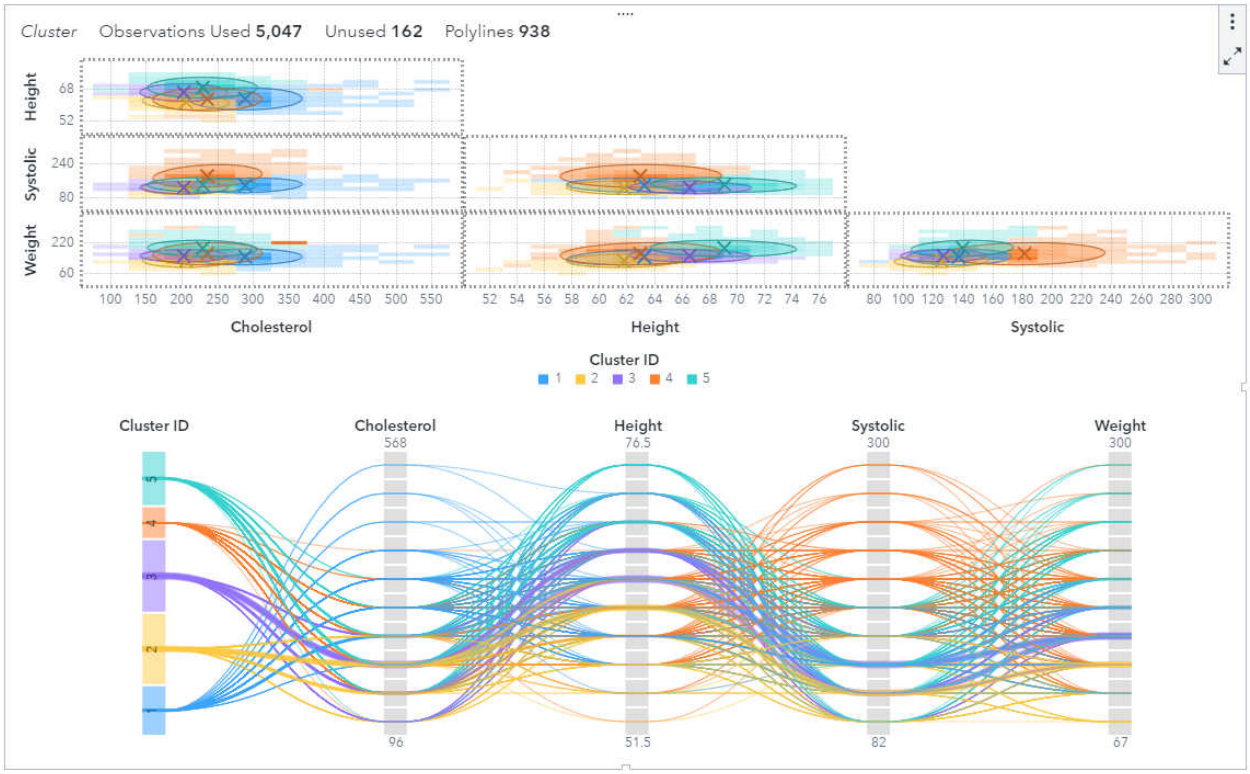
   d Save the report.

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 variable pairs. For each pair of variables and for each cluster ID, the standard deviation of the two variables and their correlation is calculated. The ellipse that is plotted is based on the standard deviations, and the tilt of the ellipse is based on the correlation value. These projections are useful for spotting cluster similarities and differences within the plotted variable pairs. To view a larger plot for a variable 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 variable pair are displayed.

Each cluster is assigned a unique color. Although each cluster is unique in *n*-space, the two-dimensional projections 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 adjust 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.

Create a Cluster

1 In the left pane, click to select an object. Drag the icon onto the canvas.

2 Click in the right pane. Specify at least two variables for **Variables**. You cannot specify an interaction term or spline 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.
- **Range**: Linearly transforms the variable values to the range [0, 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 objects on the canvas automatically. **Stack** displays the objects as if they are in a slide deck where only one object is displayed at a time. When **Stack** is specified, a control bar instead of a scroll bar lets you move between objects.

Cluster Diagram

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

Fill ellipses
specifies whether cluster ellipses are filled.

Visible roles
determines how many variables are shown in the cluster matrix. You must specify an integer value in the range 2–6.

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 matrix.

Parallel Coordinates Plot

Number of bins
specifies the number of bins to use in the Parallel Coordinates plot. You must specify an integer value in the range 2–16.

Maximum polylines
specifies the maximum number of polylines generated by the parallel coordinates algorithm. You must specify an integer value in the range 2–100,000.

Visible roles
determines how many variables are shown in the Parallel Coordinates plot. You must specify an integer value in the range 1–10.

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** is 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.
- **Interval distance from centroid value** specifies the distance from the observation to the centroid of the cluster that contains the observation based only on the measure variables.
- **Nominal distance from centroid value** specifies the distance from the observation to the centroid of the cluster that contains the observation based only on the category variables.
- **Standardized distance from centroid value** specifies the distance from the observation to the centroid of the cluster that contains the observation after measure variables have been standardized.

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.
- **Average Distance**: The average distance to the centroid of all observations within a cluster.
- **Within Distance**: The sum of distances between each observation and the centroid.

Model Information
Gives an overview of the model.

Within Cluster Statistics
Provides the mean and standard deviation for each variable within each cluster.

Iteration History
Provides the convergence history of the cluster-creation process.

Frequency
Provides the number of observations in each cluster, sorted by measurement level values.

Standardization
Provides the standardization results for each variable.

Interval Information
Provides the original mean and standard deviation for each variable.

Parallel Coordinates Plot
Provides the cluster ID and the values of each variable for each polyline in the Parallel Coordinates plot.

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*. The final nodes are called *leaves*. The segmentation from the root node to a leaf forms a *branch*. 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 or a spline term.

The decision tree can create both classification trees and regression trees. A classification tree is used to model categorical data and a regression tree is used to model measure data. For a measure response variable, choosing whether to bin the response variable determines whether a classification tree or regression tree is created. Bin the response variable to create a classification tree or keep it unmodified to create a regression tree.

Create a Decision Tree

To create a decision tree, complete the following steps:

1. In the left pane, click to select an object. Drag 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**.
4. (Optional) Specify **Partition ID**.

About Interactive Mode

The decision tree enables you to manually train and prune nodes by entering interactive mode. In interactive mode, properties are locked. To enter interactive mode, right-click a node, 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 clearing the partition column that is assigned to the decision tree
- 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 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 appears that indicates 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.

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 category response variable contains more than two levels, SAS Visual Analytics treats all observations in the level of interest as an event and all other observations as nonevents.

Autotune
enables you to specify the hyperparameters that control the autotuning algorithm. The hyperparameters determine how long the algorithm can run, how many times the algorithm can run, and how many model evaluations are allowed.

The autotuning algorithm selects the **Maximum levels**, **Leaf size**, and **Predictor bins** values that produce the best model.

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

- **None**: Observations with missing values are excluded from the model.
- **Use in search**: If the number of observations with missing values is greater than or equal to **Minimum value**, then for category variables, missing values are considered a unique measurement level and are included in the model. For measure variables, missing values are used in the calculation of the worth of a splitting rule. Consequently, a splitting rule is produced that assigns the missing values to the branch that maximizes the worth of the split.
- **As machine smallest**: Missing interval values are set to the smallest possible machine value such that the observations are always in the split with the lower variable values. Missing category values are treated as a unique measurement level.
- **Popular**: Observations with missing values are assigned to the child node with the most observations.
- **Similar**: Observations with missing values are assigned to the node deemed most similar by a chi-square test for category responses or an *F* test for measure responses.

Minimum value
specifies the minimum number of observations allowed to have missing values before missing values are treated as a distinct category level. This option is used only when **Missing assignment** is set to **Use in search**.

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 5.

Bin response variable
specifies whether a measure response variable is binned. When a response variable is binned, a classification tree is created. Otherwise, a regression tree is created.

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 50.

Bin method
Specifies the method that is used to bin the measure predictors. Select **Bucket** to divide the measure predictors into evenly spaced intervals based on the difference between maximum and minimum values. Select **Quantile** to divide the measure predictors into approximately equal sized groups. The default value is **Quantile**.

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. Measure responses without binning use a variable as the selection criterion. When enabled, categorical and binned measure responses use the gain ratio criterion. Otherwise, the information gain criterion is used.

Prune with validation data
specifies whether the decision tree is pruned using a cost-complexity algorithm applied to the validation data. When enabled, other pruning options are ignored.

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 in the range 5–100.

Prediction cutoff
specifies the value at which a computed probability is considered an event.

Statistic percentile
specifies the depth for the percentile bins that are used to calculate the observed average, lift, cumulative lift, cumulative percentage captured, cumulative percentage events, and gain.

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 objects are displayed on the canvas. **Fit** aligns all of the objects 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 objects.

Statistic to show
specifies which assessment statistic to display in the model. If you are using a partition variable, the following fit statistics (with the exception of SSE) are available for each partition. The object toolbar contains submenus for each partition type that is available (training, validation, and test).

- **ASE**
- **Observed Average**
- **SSE**
- **C Statistic**
- **Cumulative % Captured**
- **Cumulative % Events**

- **Cumulative Lift**
- **F1 Score**
- **FDR**
- **FPR**
- **Gain**
- **Gamma**
- **Gini**
- **KS (Youden)**
- **Lift**
- **Misclassification Rate**
- **Misclassification Rate (Event)**
- **Tau**

See Fit Statistics on page 8 for more information about the fit statistics that are available.

> **Note:** **ASE**, **Observed Average**, and **SSE** are available only for regression trees and classification trees that have a measure response with more than ten bins. All other fit statistics are available only for classification trees with a category response or a measure response with ten bins or fewer.

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.

Variable Importance / Leaf Statistics Plot

Plot type
specifies whether the Variable Importance plot or Leaf Statistics plot is displayed.

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

Assessment Plots

Display test partition
specifies whether to display the assessment plot for the test partition. This option is available only when you specify a partition ID with a test partition.

Plot to show
specifies which assessment plot is displayed. Select **Confusion matrix**, **Lift**, **ROC**, or **Misclassification**. This option is hidden when you have a regression tree or have a classification tree that has 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 regression tree or have a classification tree that has a measure response with more than ten bins.

Legend visibility
specifies whether the legend is displayed in the Confusion Matrix, Lift plot, ROC plot, or Misclassification 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:

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

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

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.
- To derive predicted values, right-click in the Tree window, and select **Derive predicted**. For a category response, three variables that contain the predicted values, probability values, and prediction cutoffs for each observation are created. For a binned measure response, two variables that contain the predicted values and probability values for each observation are created. For a measure response that is not binned, two variables that contain the predicted values and the residual values for each observation are created. You can use these variables 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. In addition, it splits the node based on the variable with the best information gain when **Rapid growth** is disabled.

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. For regression trees, the Leaf Statistics plot displays a bar chart of the average response values for the leaf.

Variable Importance

The Variable Importance plot displays the importance of each variable.

Assessment Plot

The confusion matrix displays the classification results for categorical response models. After a model is created, each observation has an observed value and a predicted value. The total number of each observed-predicted pair is calculated. The confusion matrix displays how many observations fall into each pair. A perfect model always predicts the observed value, and all values lie on the diagonal of the matrix. Any off-diagonal values represent a misclassification. Cells are shaded based on the proportion of the value in each cell to the number of observed values for that level. Darker shaded cells show the concentration of the predictions for the observed level. For a binary response, the information in the confusion matrix is identical to the Misclassification 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.

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**, **Predicted Value**, **Average**, and **Std. Dev**.

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.

Variable Importance
Provides the variable importance information for each variable used in the model.

Cost-Complexity Pruning
Displays all of the trees that were created by the model when **Prune with validation data** is enabled. The **Best Tree** column indicates which decision tree was selected.

Confusion Matrix
Provides a summary of the correct and incorrect classifications for the model that is used to generate the confusion matrix.

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
Provides 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.

Note: The **Assessment** detail table is available only for regression trees and classification trees that have a measure response with more than ten bins. The **Lift**, **ROC**, and **Misclassification** detail tables are available only for classification trees with a category response or a measure response with ten bins or fewer.

Working with Generalized Additive Models

Overview of Generalized Additive Models

A generalized additive model is an extension of the generalized linear model. A generalized additive model relaxes the linearity assumption in a generalized linear model by allowing spline terms to characterize nonlinear dependency structures. Each spline term is constructed by the thin-plate regression spline technique. A roughness penalty is applied to each spline term by a smoothing parameter that controls the balance between goodness of fit and the roughness of the spline curve.

A generalized additive 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
Binary	Two distinct values
Gamma	Nonnegative real values
Inverse Gaussian	Positive real values
Negative Binomial	Nonnegative integers
Normal	Real values
Poisson	Nonnegative integers
Tweedie	Real values

Create a Generalized Additive Model

1 In the left pane, click to select an object. Drag 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 spline effect for **Spline effects**.

4 Specify one or more **Continuous effects**, **Classification effects**, or **Interaction effects**. These effects are optional.

5 (Optional) Specify **Partition ID**, **Frequency**, **Weight**, or **Offset** variables.

Generalized Additive Model Options

The following options are available for the generalized additive model:

General

Distribution
specifies the distribution used to model the response variable.

Power specification
determines how the Tweedie distribution is configured. Specify **Initial value** to provide a starting value that is iteratively improved. Specify **Fixed value** to provide a constant value. Specify **Automatic** to let SAS Visual Statistics determine the best power parameter.

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
Binary	C-log-log, Logit, Log-log, Probit
Gamma	Identity, Log, Reciprocal
Inverse Gaussian	Identity, Log, Power(-2)
Negative Binomial	Identity, Log
Normal	Identity, Log
Poisson	Identity, Log
Tweedie	Identity, Log

Model evaluation criterion
specifies the criterion for selecting the smoothing parameters of spline effects. Available criteria are **GCV** (generalized cross validation), **GACV** (generalized approximate cross validation), and **UBRE** (unbiased risk estimator).

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 performance iterations
specifies the maximum number of performance iterations 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.

Spline

Confidence band
specifies the confidence level used to generate the confidence band of the one-dimensional spline plot.

Univariate max degrees of freedom
specifies the maximum degrees of freedom for one-dimensional spline effects. You must specify an integer value in the range 2–200.

Bivariate max degrees of freedom
specifies the maximum degrees of freedom for two-dimensional spline effects. You must specify an integer value in the range 3–200.

Assessment

Number of bins
specifies the number of bins to use in the assessment. You must specify an integer value in the range 5–100.

Statistic percentile
specifies the depth for the percentile bins that are used to calculate the observed average.

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 Additive Model Display Options

The following display options are available for the generalized additive 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. If you are using a partition variable, **ASE** and **Observed Average** are available for each partition. The object toolbar contains submenus for each partition type that is available (training, validation, and test). **GACV**, **GCV**, and **UBRE** are available only if they are specified as the **Model evaluation criteria**.

- **AIC**

- **AICC**
- **ASE**
- **BIC**
- **GACV**
- **GCV**
- **Observed Average**
- **SSE**
- **UBRE**

See Fit Statistics on page 8 for more information about the fit statistics that are available.

Fit Summary

Plot type
specifies which effects are shown in the Fit Summary plot. Because the *p*-values for spline effects, classification effects, and continuous effects are not fully analogous in a generalized additive model, they are plotted separately.

Spline Plot

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

Assessment Plot

Display test partition
specifies whether to display the assessment plot for the test partition. This option is available only when you specify a partition ID with a test partition.

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

Generalized Additive Model Results

Fit Summary Plot

The Fit Summary plot displays the importance of each spline or 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. Because the *p*-values for spline effects, classification effects, and continuous effects are not fully analogous in a generalized additive model, they are plotted separately.

Iteration Plot

The Iteration plot displays the change in objective function value at each step of the model creation process. The vertical lines in the plot represent the first inner iteration of each performance iteration. The objective function value might increase at each vertical line, but it should always decrease within a performance iteration.

Spline Plot

The Spline plot displays one of the splines used in the generalized additive model. For a one-dimensional spline, a line plot with a confidence band is displayed. For a two-dimensional spline, a contour plot is displayed.

Assessment Plot

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

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 Plot
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.

Estimates for Smoothing Components
Provides the estimated values for the spline parameters.

Tests for Smoothing Components
Provides the effective degrees of freedom results for the spline parameters.

Parameter Estimates
Gives the estimated values for the model parameters.

Spline(Effect)
Each spline effect in the model is provided on a tab. One-dimensional splines display the spline creation results and confidence band values. Two-dimensional splines display the spline creation results.

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 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
Tweedie	Nonnegative real values

Create a Generalized Linear Model

1. In the left pane, click to select an object. Drag 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 (Optional) Specify **Partition ID**, **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.

Variable selection method
specifies whether variable selection is performed. For more information, see Variable Selection on page 4. Methods available are **None**, **Forward**, **Backward**, **Fast Backward**, **Stepwise**, and **Lasso**.

Maximum effects
specifies the maximum number of effects included in the model when performing variable selection.

Selection criterion
specifies the statistic used to determine whether an effect is included in the model when performing variable selection. Available criteria are **AIC**, **AICC**, **SBC**, **Significance level**, and **Validation** (ASE of the validation partition).

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 4 . When **Significance level** is specified for the **Selection criterion**, this property is available for **Forward**, **Backward**, **Stepwise**, and **Fast Backward** variable selection methods.

Distribution
specifies the distribution used to model the response variable.

Power specification
determines how the Tweedie distribution is configured. Specify **Initial value** to provide a starting value that is iteratively improved. Specify **Fixed value** to provide a constant value. Specify **Automatic** to let SAS Visual Statistics determine the best power parameter.

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	Identity, Log
Gamma	Identity, Log, Reciprocal
Geometric	Identity, Log

Distribution	Available Link Functions
Inverse Gaussian	Identity, Log, Power(-2)
Negative Binomial	Identity, Log
Normal	Identity, Log
Poisson	Identity, Log
Tweedie	Identity, Log

Standardize continuous effects
specifies whether continuous effects are standardized.

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 in the range 5–100.

Statistic percentile
specifies the depth for the percentile bins that are used to calculate the observed average.

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. If you are using a partition variable, **ASE** and **Observed Average** are available for each partition. The object toolbar contains submenus for each partition type that is available (training, validation, and test).

- **-2 Log Likelihood**
- **AIC**
- **AICC**
- **ASE**
- **BIC**
- **Observed Average**
- **SSE**

See Fit Statistics on page 8 for more information about the fit statistics that are available.

Parameter Estimate Plot

Legend visibility
specifies whether the legend is displayed in the Parameter Estimate plot.

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 in the Residual plot.

Variable Selection Plot

Plot to show
specifies whether to display the Variable Selection plot.

Assessment Plot

Display test partition
specifies whether to display the assessment plot for the test partition. This option is available only when you specify a partition ID with a test partition.

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

Generalized Linear Model Results

Fit Summary

The Fit Summary plot displays the 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 plot displays a Goodness of Fit plot and a Variable Importance plot.

Parameter Estimate

The Parameter Estimate plot displays the median change in response value for each unit change of the effect. For classification effects, the estimate is computed individually for each level of the effect.

This plot is displayed instead of the Fit Summary plot when you specify the **Lasso** variable selection.

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

Variable Selection Plot

Displays the change in value of the variable selection statistic as effects are added to or removed from the model. This plot is available only when a variable selection method other than **None** is specified.

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*.

Selection Info
Provides a summary of the variable selection methodology.

Selection Summary
Provides a summary of the variable selection results at each step in the selection process.

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.

Create a Linear Regression

1. In the left pane, click to select an object. Drag 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. (Optional) Specify **Partition ID**, **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 4. Methods available are **None**, **Forward**, **Backward**, **Stepwise**, **Lasso**, and **Adaptive lasso**.

Maximum effects
specifies the maximum number of effects included in the model when performing variable selection.

Selection criterion
specifies the statistic used to determine whether an effect is included in the model when performing variable selection. Criterion available are **AIC**, **AICC**, **SBC**, **Significance level**, and **Validation** (ASE of the validation partition).

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 4 . When **Significance level** is specified for the **Selection criterion**, this property is available for **Forward**, **Backward**, and **Stepwise** variable selection methods.

Assessment

Number of bins
specifies the number of bins to use in the assessment. You must specify an integer value in the range 5–100.

Statistic percentile
specifies the depth for the percentile bins that are used to calculate the observed average.

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. If you are using a partition variable, **ASE** and **Observed Average** are available for each partition. The object toolbar contains submenus for each partition type that is available (training, validation, and test).

- **Adjusted R-Square**
- **AIC**
- **AICC**
- **ASE**
- **F Value of Model**
- **Mean Square Error**
- **Observed Average**
- **R-Square**
- **Root MSE**
- **SBC**
- **SSE**

See Fit Statistics on page 8 for more information about the fit statistics that are 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 in the Residual plot.

Influence Plot / Variable Selection Plot

Plot to show
specifies whether the Influence plot, Variable Selection plot, or neither 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

Display test partition
specifies whether to display the assessment plot for the test partition. This option is available only when you specify a partition ID with a test partition.

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

Linear Regression Results

Fit Summary Window

The Fit Summary plot displays the 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 plot displays a Goodness of Fit plot and a Variable Importance plot.

Parameter Estimate

The Parameter Estimate plot displays the median change in response value for each unit change of the effect. For classification effects, the estimate is computed individually for each level of the effect.

This plot is displayed instead of the Fit Summary plot when you specify either the **Lasso** or **Adaptive lasso** variable selection methods.

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

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 absolute value of the **DFFITS** and **Likelihood Displacement** measurements are shown in the influence plot. Negative values of these measurements are displayed with a different color.

The following options are available in the Influence plot:

- To change the statistic 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.

Variable Selection Plot

Displays the change in value of the variable selection statistic as effects are added to or removed from the model. This plot is available only when a variable selection method other than **None** is specified.

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*.

Selection Info
Provides a summary of the variable selection methodology.

Selection Summary
Provides a summary of the variable selection results at each step in the selection process.

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.

Create a Logistic Regression

1. In the left pane, click to select an object. Drag 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. (Optional) Specify **Partition ID**, **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 4. Methods available are **None**, **Forward**, **Backward**, **Fast Backward**, **Stepwise**, and **Lasso**.

Maximum effects
specifies the maximum number of effects included in the model when performing variable selection.

Selection criterion
specifies the statistic used to determine whether an effect is included in the model when performing variable selection. Criterion available are **AIC**, **AICC**, **SBC**, **Significance level**, and **Validation** (ASE of the validation partition).

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 4 . When **Significance level** is specified for the **Selection criterion**, this property is available for the **Forward**, **Backward**, **Fast Backward**, and **Stepwise** variable selection methods.

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.

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

- **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 in the range 5–100. By default, measure variables are grouped into 20 bins.

Prediction cutoff
specifies the value at which a computed probability is considered an event.

Statistic percentile
specifies the depth for the percentile bins that are used to calculate the observed average, lift, cumulative lift, cumulative percentage captured, cumulative percentage events, and gain.

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. If you are using a partition variable, **C Statistic**, **Cumulative % Captured**, **Cumulative % Events**, **Cumulative Lift**, **F1 Score**, **FDR**, **FPR**, **Gain**, **Gamma**, **Gini**, **KS (Youden)**, **Lift**, **Misclassification Rate (Event)**, and **Tau** are available for each partition. The object toolbar contains submenus for each partition type that is available (training, validation, and test).

- **-2 Log Likelihood**
- **AIC**
- **AICC**
- **BIC**
- **C Statistic**
- **Cumulative % Captured**
- **Cumulative % Events**
- **Cumulative Lift**
- **F1 Score**
- **FDR**
- **FPR**
- **Gain**
- **Gamma**
- **Gini**
- **KS (Youden)**
- **Lift**
- **Max-rescaled R-square**
- **Misclassification Rate (Event)**
- **R-Square**
- **Tau**

See Fit Statistics on page 8 for more information about the fit statistics that are 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 is displayed in the Residual plot.

Influence Plot / Variable Selection Plot

Plot to show
specifies whether the Influence plot, Variable Selection Plot, or neither is displayed.

Use histogram
specifies whether the Influence plot is a histogram.

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

Assessment Plots

Display test partition
specifies whether to display the assessment plot for the test partition. This option is available only when you specify a partition ID with a test partition.

Plot to show
specifies which assessment plot is displayed. Select **Confusion matrix**, **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 is displayed in the confusion matrix, Lift plot, ROC plot, or Misclassification plot.

Logistic Regression Results

Fit Summary

The Fit Summary plot displays the 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 plot displays a Goodness of Fit plot and a Variable Importance plot.

Parameter Estimate

The Parameter Estimate plot displays the median change in response value for each unit change of the effect. For classification effects, the estimate is computed individually for each level of the effect.

This plot is displayed instead of the Fit Summary plot when you specify the **Lasso** variable selection method.

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 for an event is

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

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

The confusion matrix displays the classification results for categorical response models. After a model is created, each observation has an observed value and a predicted value. The total number of each observed-predicted pair is calculated. The confusion matrix displays how many observations fall into each pair. A perfect model always predicts the observed value, and all values lie on the diagonal of the matrix. Any off-diagonal values represent a misclassification. Cells are shaded based on the proportion of the value in each cell to the number of observed values for that level. Darker shaded cells show the concentration of the predictions for the observed level. For logistic regressions, the information in the confusion matrix is identical to the Misclassification 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 absolute value of the **Likelihood Displacement** measurement is shown in the Influence plot. Negative values of this measurement are displayed with a different color.

The following options are available in the Influence plot:

- To change the statistic that is plotted, change the value of the **X axis** option in the **Influence Plot / Variable Selection 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.

Variable Selection Plot

Displays the change in value of the variable selection statistic as effects are added to or removed from the model. This plot is available only when a variable selection method other than **None** is specified.

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*.

Selection Info
Provides a summary of the variable selection methodology.

Selection Summary
Provides a summary of the variable selection results at each step in the selection process.

Confusion Matrix
Provides a summary of the correct and incorrect classifications for the model that is used to generate the confusion matrix.

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 Nonparametric Logistic Regression Models

Overview of Nonparametric Logistic Regression Models

A nonparametric logistic regression attempts to predict the value of a binary response variable. A nonparametric logistic regression is an extension of the logistic regression that allows spline terms to characterize nonlinear dependency structures. Each spline term is constructed by the thin-plate regression spline technique. A roughness penalty is applied to each spline term by a smoothing parameter that controls the balance between goodness of fit and the roughness of the spline curve.

The nonparametric logistic regression requires that you specify a link function. The link function is used to relate the response variable to the effect variables.

Create a Nonparametric Logistic Regression

1 In the left pane, click to select an object. Drag 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 spline effect for **Spline effects**.

4 (Optional) Specify one or more measure variables for **Continuous effects**, a category variable for **Classification effects**, or an interaction for **Interaction effects**.

5 (Optional) Specify **Partition ID**, **Frequency**, **Weight**, or **Offset** variables.

Nonparametric Logistic Regression Options

The following options are available for the nonparametric 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.

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:

- **C-log-log** specifies the complementary log-log function.

$$g(M) = \ln(-\ln(1-p))$$

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

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

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

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

$$g(M) = \log(-\log(p))$$

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

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

Model evaluation criterion
specifies the criterion for selecting the smoothing parameters of spline effects. Available criteria are **GCV** (generalized cross validation), **GACV** (generalized approximate cross validation), and **UBRE** (unbiased risk estimator).

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 performance 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.

Spline

Confidence band
specifies the confidence level used to generate the confidence band of the one-dimensional spline plot.

Univariate max degrees of freedom
specifies the maximum degrees of freedom for one-dimensional spline effects. You must specify an integer value in the range 2–200.

Bivariate max degrees of freedom
specifies the maximum degrees of freedom for two-dimensional spline effects. You must specify an integer value in the range 3–200.

Assessment

Number of bins
specifies the number of bins to use in the assessment. You must specify an integer value in the range 5–100.

Prediction cutoff
specifies the value at which a computed probability is considered an event.

Statistic percentile
specifies the depth for the percentile bins that are used to calculate the observed average, lift, cumulative lift, cumulative percentage captured, cumulative percentage events, and gain.

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.

Nonparametric Logistic Regression Model Display Options

The following display options are available for the nonparametric 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. If you are using a partition variable, **C Statistic**, **Cumulative % Captured**, **Cumulative % Events**, **Cumulative Lift**, **F1 Score**, **FDR**, **FPR**, **Gain**, **Gamma**, **Gini**, **KS (Youden)**, **Lift**, **Misclassification Rate (Event)**, and **Tau** are available for each partition. The object toolbar contains submenus for each partition type that is available (training, validation, and test).

- **AIC**

- **AICC**
- **BIC**
- **C Statistic**
- **Cumulative % Captured**
- **Cumulative % Events**
- **Cumulative Lift**
- **F1 Score**
- **FDR**
- **FPR**
- **GACV**
- **Gain**
- **Gamma**
- **GCV**
- **Gini**
- **KS (Youden)**
- **Lift**
- **Misclassification Rate (Event)**
- **Tau**
- **UBRE**

See Fit Statistics on page 8 for more information about the fit statistics that are available.

Fit Summary
specifies which effects are shown in the Fit Summary plot. Because the *p*-values for spline effects, classification effects, and continuous effects are not fully analogous in a nonparametric logistic regression model, they are plotted separately.

Spline Plot

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

Assessment Plots

Display test partition
specifies whether to display the assessment plot for the test partition. This option is available only when you specify a partition ID with a test partition.

Plot to show
specifies which assessment plot is displayed. Select **Confusion matrix**, **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 is displayed in the confusion matrix, Lift plot, ROC plot, or Misclassification plot.

Nonparametric Logistic Regression Results

Fit Summary

The Fit Summary plot displays the 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. Because the *p*-values for spline effects, classification effects, and continuous effects are not fully analogous in a nonparametric logistic regression model, they are plotted separately.

Iteration Plot

The Iteration plot displays the change in objective function value at each step of the model creation process. The vertical lines in the plot represent the first inner iteration of each performance iteration. The objective function value might increase at each vertical line, but it should always decrease within a performance iteration.

Spline Plot

The Spline plot displays one of the splines used in the nonparametric logistic regression. For a one-dimensional spline, a line plot with a confidence band is displayed. For a two-dimensional spline, a contour plot is displayed.

Assessment Plot

The confusion matrix displays the classification results for categorical response models. After a model is created, each observation has an observed value and a predicted value. The total number of each observed-predicted pair is calculated. The confusion matrix displays how many observations fall into each pair. A perfect model always predicts the observed value, and all values lie on the diagonal of the matrix. Any off-diagonal values represent a misclassification. Cells are shaded based on the proportion of the value in each cell to the number of observed values for that level. Darker shaded cells show the concentration of the predictions for the observed level. For nonparametric logistic regressions, the information in the confusion matrix is identical to the Misclassification 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 nonparametric 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.

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 Plot
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.

Estimates for Smoothing Components
Provides the estimated values for the spline parameters.

Tests for Smoothing Components
Provides the effective degrees of freedom results for the spline parameters.

Parameter Estimates
Gives the estimated values for the model parameters.

Spline(Effect)
Each spline effect in the model is provided on a tab. One-dimensional splines display the spline creation results and confidence band values. Two-dimensional splines display the spline creation results.

Confusion Matrix
Provides a summary of the correct and incorrect classifications for the model that is used to generate the confusion matrix.

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

1. In the left pane, click to select an object. Drag the icon onto the canvas.
2. In the Add Model Comparison window, specify the **Data source**, **Partition**, **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, partition information, 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 that are available depend on the models being compared.

Relative Importance Plot

Legend visibility
specifies whether the legend is displayed in the Relative Importance plot.

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.

Partition
specifies which partition the assessment plot should display.

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

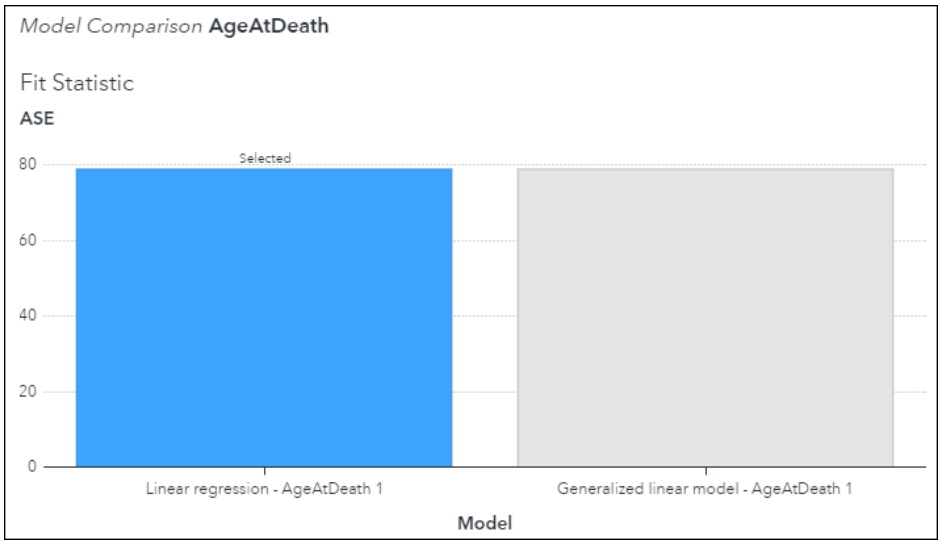
Model Comparison Results

Assessment Plot

The assessment plots available depend on the models being compared. For classification models, the plots displayed are the confusion matrix, Lift, ROC, and Misclassification. For numerical models, the plots displayed are Observed Response Value and Predicted Response Value. When a partition column is used, you can display each partition individually or side by side.

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.



Variable Importance

The Relative Importance plot compares, across all models, the importance of any effect included in any model. Use this plot to determine which variables have the most impact on their models and the size of that impact relative to the other models. The variable importance calculations are not comparable across all models. When the variable importance values are not comparable, this plot is hidden.

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.

Relative Importance
Provides the importance calculation results for each variable.