# SAS Visual Statistics in VA – Interactive modeling

Pia Rønnevik, Customer Success Manager, FANS, SAS Institute

pia.roennevik@sas.com

# SAS Visual Statistics i Visual Analytics

**Modeling Techniques (Visual Interface)**

- Linear Regression
- Logistic Regression
- Nonparametric Logistic
- GLM Regression
- GAM Regression
- Clustering
- Decision Tree

**Analytical Procedures (SAS Studio Programmatic Interface)**

- GENSELECT (Generalized Linear Model)
- KCLUS (k-means and k-modes Clustering)
- NMF (Nonnegative Matrix Factorization)
- SANDWICH (Sandwich Variance Estimator)
- PCA (Principal Component Analysis)
- LOGSELECT (Logistic Regression)
- NLMOD (Nonlinear Regression)
- REGSELECT (Linear Regression)
- TREESPLIT (Decision Trees)
- PLSMOD (Partial Least Square)
- QTRSELECT (Quantile Regression)
- SPC (Statistical Process Control)
- LMIXED (Linear Mixed Models)
- MBC (Model-Based Clustering)
- SIMSYSTEM (Simulate Univariate Data)
- GAMMOD (Generalized Additive Model)
- GAMSELECT (Model Selection for GAM)
- PHSELECT (Proportional Hazard Model)
- ICA (Independent Component Analysis)
- MODELMATRIX (Matrix of Covariates

# Multiple Linear Regression Model

**Multiple Linear Regression models** enable you to predict the value of a response variable (dependent variable) as a linear function of one or more effects (independent or predictor variables):

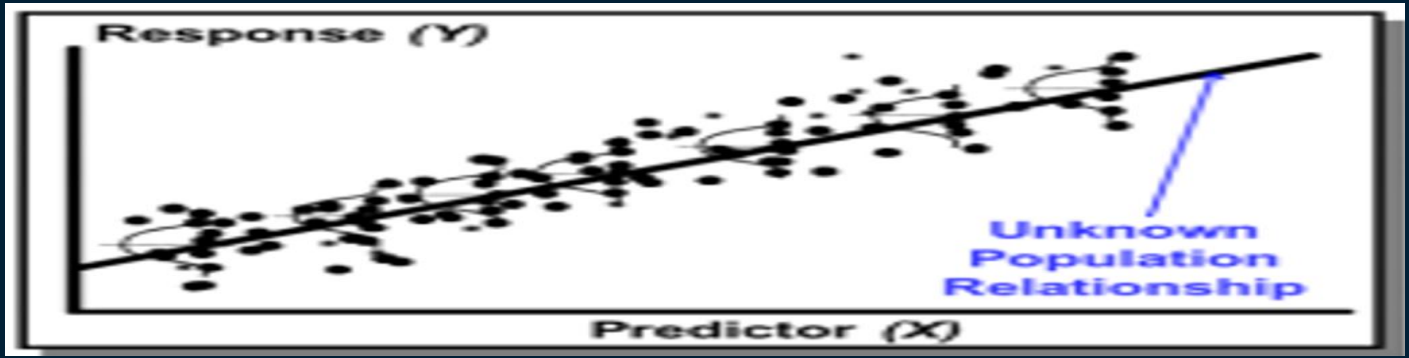
• Linear regression models assume that the relationship between the response variable and the input variables is linear:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon$$

• Linear regression uses the least squares method to determine parameter estimates.

• Model effects or explanatory variables can be one of the following effects:
  - Continuous
  - Categorical
  - Interaction terms

# The Assumptions of Multiple Linear Regression Model




- The mean of the response (Y) is acurately modeled by a function of the predictors (X) that is linear in the parameters.
- The random erro term, Ɛ, is assumed to have a normal distribution with a mean of zero and a constant variance, σ.
- The errors are independent

# Multicollinearity

Multicollinearity is about linear dependence between two or more independent variables, and explains that two variables are correlated.

**How to measure:**

- Calculating correlations between the independent variables or calculating VIF coefficients.
- Correlation close to 1 or -1 indicate a high degree of linear relationship.
- Some believe that a VIF of 5.5 is the upper limit for how strong a degree of collinearity can be tolerated. Others claim that collinearity becomes a problem when VIF exceeds 10.

**How to solve:**

- Remove some of the highly correlated independent variables.
- Linearly combine the independent variables, such as adding them together.
- Center the variables.
- Redefine the predictor variables.
- Use LASSO, Ridge regression or principal component regression

# The Multiple Linear Regression Model Hypothesis Test

**Null Hypothesis: $H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0$**

The regression model doesn't fit the data better than the mean model.

**Alternative Hypothesis: $H_1$: Not all $\beta_i$s equal zero.**

The regression model does fit the data better than the mean model.

# Generalized Linear Models (GLM)

**Generalized linear models** extend the theory and methods of linear models to data that is not normally distributed.

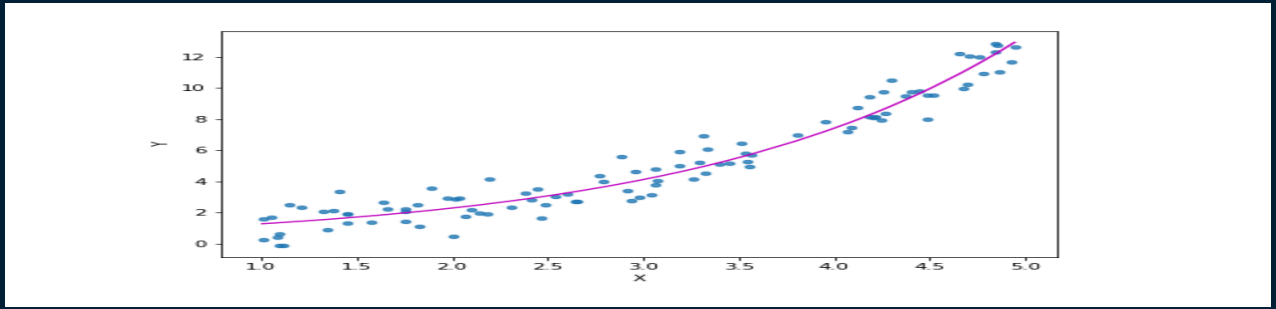
- The distribution of the observations can come from the exponential family of distributions:

$$g(E(y_i)) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} = X\beta$$

- Parameter estimation uses maximum likelihood estimation (MLE) rather than ordinary least squares (OLS)

- The models have three components:
  - Random component: identifies the response variable and its probability distribution.
  - Systematic component: specifies the predictor variables used in a linear predictor.
  - Link function: specifies the link between the random and the systematic components. It indicates how the expected value of the response relates to the linear combination of explanatory variables

# The Assumptions of GLM




- The dependent variable Y is assumed to come from a distribution from an exponential family.
- A GLM does assume a linear relationship between the transformed expected response in terms of the link function and the explanatory variables.
- Explanatory variables can be nonlinear transformations of some original variables.
- The homogeneity of variance does NOT need to be satisfied.
- Errors need to be independent but NOT normally distributed.

# Two popular examples of GLM

**Binary Logistic Regression**

Binary logistic regression models how the odds of "success" for a binary response variable Y depend on a set of explanatory variables:

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i$$

- Random component - The distribution of the response variable is assumed to be binomial with a single trial and success probability E(Y)=∏.
- Systematic component - x is the explanatory variable (can be continuous or discrete) and is linear in the parameters. As with the above example, this can be extended to multiple variables of non-linear transformations.
- Link function - the log-odds or logit link,

$$\eta = g(\pi) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$$

is used.

Poisson Regression

models how the mean of a discrete (count) response variable Y depends on a set of explanatory variables:

$$\log \lambda_i = \beta_0 + \beta x_i$$

- Random component - The distribution of Y is Poisson with mean λ.
- Systematic component - x is the explanatory variable (can be continuous or discrete) and is linear in the parameters. As with the above example, this can be extended to multiple variables of non-linear transformations.
- Link function - the log link is used.

# The distribution of the response variable for Generalized Linear Models

| Distribution | Range Requirements | Available Link Functions |
|---|---|---|
| Beta | Values must be between 0 and 1, exclusive. | Logit, Probit, Log-log, C-log-log |
| Binary | Exactly two distinct values | Logit, Probit, Log-log, C-log-log |
| Exponential | Nonnegative real values | Log, Identity |
| Gamma | Nonnegative real values | Log, Identity, Reciprocal |
| Geometric | Positive integers | Log, Identity |
| Inverse Gaussian | Positive real values | Power(-2), Log, Identity |
| Negative Binomial | Nonnegative integers | Log, Identity |
| Normal | Real values | Identity, Log |
| Poisson | Nonnegative integers | Log, Identity |
| Tweedie | Nonnegative real values | Identity, Log |

# Variable selection methods

Variable selection methods is a method to be used to reduce the number of input variables to include only the most significant variables:

– **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 can significantly improve 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.

– **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 considered not significant.

– **Lasso**: This method adds and removes candidate effects based on a version of ordinary least squares, where the sum of the absolute regression coefficients is constrained.

– **Adaptive Lasso**: Available for linear regressions, this is a modification to Lasso where selection weights are applied to each of the parameters that are used to create the Lasso constraint.

# Analyzing the results

- **The Fit Summary window**: shows how significant the effect variables are to the response variable.

- **The Residual Plot window:** shows the relationship between the predicted and the residual data. The residual is the difference between the predicted response value and the actual response value. Residual plots can detect nonconstant variance in the input data, which evident when the relative spread of the residual values changes as the predicted values change. The plot can also help identify outliers

- **The Assessment window:** shows the values for the observed response and the model's predicted response. Large differences in the average predicted and average observed values can indicate a bias.

- **The Influence Plot window:** shows observations that might influence the overall analysis.

- **The Variable Selection Plot window:** shows the change in value of the variable selection statistic as effects are added or removed.

# Generalized Additive Model (GAM)

**Generalized additive models** are an extension of the generalized linear model, where the response variable depends linearly on unknown smooth functions of some predictor variables.

$$g(\mathrm{E}(Y)) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_m(x_m)$$

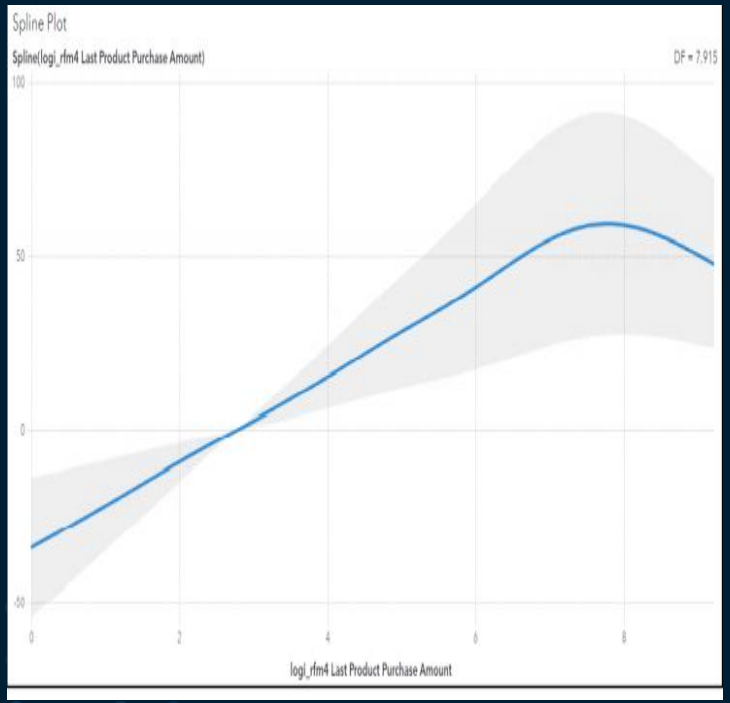
- A distribution is specified for Y (normal, binomial…..) together with a link function.

- Model effects or explanatory variables can be any of the following effects:
  - spline
  - continuous
  - categorical
  - interaction terms

- Advantages of generalized additive models include pattern discovery and potential better predictive capability

- A disadvantage includes the loss of interpretability of a spline effect as a predictor

# What is splines?




The splines are useful for modeling complex nonlinear relationship for which simple polynomials are not sufficient:

- A spline is a smooth function consisting of piecewise polynomials joined at points called knots.
- Splines can take a number of forms depending on the degree of the polynomial and the number of knots that are used.

# Logistic Regression

Logistic regression enables you to investigate the relationship between a discrete response variable and one or more independent variables:

- Mathematically: $$ln\left(\frac{P(y=1)}{1-P(y=1)}\right) = log\left(\frac{P(y=1)}{P(y=0)}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

- Type of Logistic Regression:
  - Binary: the response variable has two outcomes {0, 1}.
  - Multinomial: Three or more outcome without ordering.
  - Ordinal: Three or more outcome with ordering.

- Multiple effects variables can be any of the following:
  - continuous
  - categorical
  - interaction terms

# The Assumptions of Logistic Regression

- Binary logistic regression requires the dependent variable to be binary, and Ordinal logistic regression requires the dependent variable to be ordinal.
- The observations are independent. That is, the observations should not come from repeated measurements/matched data.
- There should not be multicollinearity among the independent variables, meaning variables shouldn't be too highly correlated with each other.
- There are no extreme outliers.
- The independent variables should be linearly related to the log odds.
- Logistic regression typically requires a large sample size. A general guideline is that you need at minimum of 10 cases with the least frequent outcome for each independent variable in your model.

# Analyzing the results with a Logistic Regression

Three panes appear after running Logistic Regression, and can help you analyze the results of the logistic regression model.

- **Fit Summary** : displays the relationship between the predicted and the residual data.

- **Residual Plot**: displays the difference between the predicted and the actual data.

- **Assessment**:
  - Confusion matrix summarizes classifications: number of correct and incorrect predictions are summarized (true positives, false negatives, false positives, and true negatives).
  - Lift measures model effectiveness: is the ratio of the percent of captured responses within each percentile bin to the average percent of responses for the model.
  - ROC (receiver operating characteristic): measures classification accuracy.
  - Misclassification plot: how many observations were correctly and incorrectly classified for each value of the response variable.

# Nonparametric Logistic Regression Model

Nonparametric logistic regression models are an extension of the logistic regression model. They relax the linearity assumption and allow spline terms in order to capture nonlinear dependency structures. Like logistic regression models, nonparametric models enable you to specify a link function.

- Model effects or explanatory variables can be the following effects:
  - spline
  - continuous
  - categorical
  - interaction terms

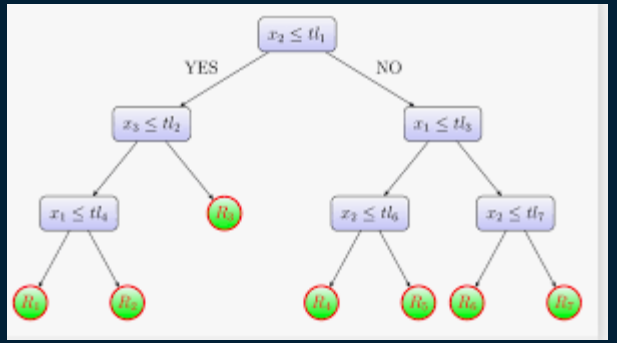
- Advantages of nonparametric models include pattern discovery, and a potentially better predictive capability.

- Disadvantages include the loss of interpretability of a spline effect as a predictor.

# Decision Trees

Decision Trees are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.

- There is only one response variable, but it can be either a category or a measure.
- There can be multiple predictors.
- Both category and measure predictors are accommodated. (interaction terms aren't allowed.)
- You can derive a leaf ID. This ID can be used in other models that are featured in the SAS Visual Statistics functionality.



# Analyzing the results from Decision Tree

Four panes appear after running Decision Tree, and can help you analyze the results of the decision tree model:

- **Tree with Treemap**: shows an interactive and navigational decision tree with node statistics and node rules.

- **Icicle Plot**: shows a hierarchical breakdown of the tree data.

- **Variable Importance Plot**: provides the variable importance information for effects in the tree.

- **Assessment**:
  - Confusion matrix summarizes classifications.
  - Lift measures model effectiveness.
  - ROC (receiver operating characteristic) measures classification accuracy.
  - Misclassification measures predictive accuracy.

# Fit Statistics

There are several assessment measures to hel you evaluate how well the model fits your data, depending on the model and whetever the response is a category or measure variable:

- **-2 Log Likelihood:** estimates the probability of an observed sample given all possible parameter values. Smaller values are preferred.
- **Adjusted R-Square**: 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): 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. Smaller values indicate better models.
- **ASE**(The average square error ): is the sum of squared errors (SSE) divided by the number of observations. Smaller values are preferred.
- **BIC**(The Bayesian information criterion):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. Lower BIC implies either fewer explanatory variables and better fit.
- **Misclassification Rate:** The misclassification rate of the model. Lower values are better.
- **R-Square:** 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.

https://go.documentation.sas.com/doc/en/vacdc/v_012/vaobjvs/n1j5llevh3vltfn1a7b0jqu94e4u.htm#n0gn5

29dn10djvcxz8q3ov

# Thanks for you attention!