SAS® VISUAL STATISTICS

WORKSHOP
SAS FORUM 2015

MICHAEL SPERLING
About SAS® Visual Statistics

Whether you are a business user, a statistician or a data scientist, faster insights are at your fingertips. Get all the benefits of data visualization tools with absolutely no coding required. SAS Visual Statistics offers advanced analytics, including decision trees, logistic regression, supervised and unsupervised clustering and generalized least squares. Furthermore, one can compare model performance, export-scoring code and the results from one analysis can be used in new visualizations.

About this workshop:

- Introduces some of the capabilities in SAS Visual Statistics
- Show how results from statistical modelling can be used in descriptive statistics afterwards
- The sections work independently (with a few exceptions) so you may jump ahead if you like.
1. Open data source and validate

1. Click Google Chrome in the task bar at the bottom of the screen.

2. Select SAS Visual Analytics Hub from the menu.

3. Log on as the user sdkstudent\sasdemo with the password Orion123. If you see this userid and dots for password already predefined, then click Sign In.

4. Click Exploration

5. To open the Data Explorer double click the application shortcut for the Data Explorer on the SAS® Home page.

6. Open the exploration SAS Forum 2015 Visual Statistics located at InsightToys\Reports and Analyses\Hands-on\ 

An empty data exploration appears with the data source Insighttoy5_transformed selected.
2. Cluster analysis

1. Start with a blank workspace, select “Gruppe” or “Cluster” from the toolbar.

2. Drag variables to the canvas:
   a. Vendor Satisfaction
   b. Order Total
   c. Order Distribution Cost

3. Maximize Parallel Coordinates

4. Un-click Auto-update

5. Under properties set:
   a. Number of Clusters: 3
   b. Visible Roles: 3
   c. Number of bins: 4
   d. Maximum polylines: 3.000
   e. Visible roles: 3
5.

6. Change name to “Klyngeanalyse”

7. Visual evaluation:
   a. Click on the individual Clusters in the bar to the left in order to see how the individual clusters are distributed on the three variables
   b. Click just below “Order Distribution Cost” and pull downwards to see where the observations (automatically filtered) are distributed on the clusters

8. Click on and choose “Derive a Cluster ID Variable”
9. Name the new variable “Segment” and press “OK”

10. Notice that a new variable has appeared on the left side

Why do I get 4? – Because a cluster for missing values has been created.

You have now created an un-supervised clustering analysis and the result, Segment, is ready to use in another visualization.
3. Logistic Regression

1. Add a new visualization and select “Logistic Regression” from the toolbar.

2. Assign “High_Low_Market_Penetration” as response variable

3. Drag the variable Segments\(^1\) as the Classification Effect

4. Add additional Classification Effects:
   a. Manufacturing Facility
   b. Product Make
   c. Vendor Loyalty Program
   d. Vendor Type

5. Add Continuous Effects
   e. Order Sales Cost
   f. Sales Rep Actual
   g. Sales Rep Orders
   h. Sales Rep Vendors
   i. Vendor Distance

6. Note your R-square score
   j. If you wish to you can change it to -2 Log Likelihood, AIC, AICC or BIC

7. In the upper left corner you see a chart for the Variable Importance. Mouse-over to see p-value

8. In the lower left corner you can choose between Lift, ROC and misclassification as your assessment chart.

9. In the lower right corner, one can evaluate influential observations. Click on \(\) to change the measure.

\(^1\) IF you have completed section 2, if not continue to task 3 (adding additional variables)
TIP: If you click on one of the horizontal bars, you can see the residual for that observation in the Residual Plot in the upper right corner.

10. Residual plot
   a. Change the residual to *Standardized Pearson Residual*
   b. Select a number of observations in the upper left of the Residual Plot (those with the largest residual)
   c. Note that the bars in the Influence Plot in the lower right corner is marked also
   d. Right-click and click on *Exclude Selected*
   e. Right-click in the residual diagram to choose *Use Histogram*
11. Click on **Data** and **New Interaction Effect** in the Data Panel

12. Add Segment and Vendor distance as a (two-way) interaction

13. Pull the interaction effect to the visualization

14. Click on and pick **Export Score Code**
The score code can be exported

15. In the Properties Pane on the right you may name the visualization Logistic Model/(Logistisk Model)

16. Pick \[\text{VA Explorer and Predicted Values}\]

17. In the panel to the left, you now see two new variables. Add the prediction variable to the new visualization.
These new derived variables can be used in new visualizations.
4. Decision tree

1. Add a new visualization and select “Decision Tree” or Beslutningstræ from the toolbar.

2. Assign “High_Low_Market_Penetration” as response variable

3. Add Segment as a predictor

4. In the properties tab

   a. Change the name to “Beslutningstræ”
   b. Change the Maximum branches to 4
c. Change Maximum levels to 4

5. Add (some of the following) predictors
   
d. Manufacturing Facility
e. Product Make
f. Vendor Loyalty Program
g. Vendor Type
h. Order Sales Cost
i. Sales Rep Actual
j. Sales Rep Orders
k. Sales Rep Vendors
l. Vendor Distance

Notice that some predictors are shaded. Those are excluded from the model (based on their limited relationship with the response variable).
6. If you have time right-click on the elements in visualization to *Split, Train or Prune* tree.
   
   Note: once you embark on this manually driven exploration the options in the Properties Pane is no longer active.

7. In the properties Show diagnostic plots

   ![Show diagnostic plots](image)

   m. As in the Logistic Regression visualization, you may choose between Lift, ROC and misclassification.

8. Name the visualization
5. Compare the models

1. Add a new Visualization
2. Select “Model Comparison” from the toolbar.

3. Click on the Model Comparison icon

4. Select the Logistic model and Beslutningstræ to the selected models area and click “OK”.
5. Based on different measures SAS Visual Statistics now compares the two model.
   a. If you have time try to change the different measures
### 6. Change the Response variable

1. By now, you have experienced, that one can easily change the explanatory variables. However, it is just as easy to change the response variable (provided that the variables has the same format).

2. In the roles tab right-click on the variable -> click on “Replace High_Low_Market_Penetration” and pick Large Orders (which is a binary representation of the Orders Total variable).

3. Do the same for the logistic regression and repeat section 5.
7. Generalized Models

The Order size may be modelled by a generalized linear model.

1. Add a new Visualization
2. Select “Generalized Linear Model” from the toolbar.
3. Add Order Total as the response variable
4. Add a number of explanatory variables
5. Go to the Properties tab in the panel to the right
6. Name the visualization
7. Pick Poisson distribution in the drop-down-menu next to “Distribution”
1. In the assessment and residual plot windows you can evaluate the model and in the fit summary you can point out which variables are significant and which are not.

2. When you are satisfied with your model, you can derive predicted values residuals from the menu. One can use these new variables in other visualizations, e.g. use the residuals as filter to filter out outliers.
8. New Custom Category

The variables High_low_Market_Penetration and Large Orders has been created by the “New Custom Category” functionality. Below is described how this was done. If you have time, you may right-click on the variables and change this. Afterwards redo the analysis above to see how this change affects your results.

1. Pull the variable “Market Penetration” to the Visualization canvas

2. Click the Data panel and then “New Custom Category”

3. Enter a name: for instance “High_Low_Market_Penetration”

4. At a label = “Low”

5. Click below the label and add “0” and “0.00007” in the boxes

6. Click on “New Label”

7. At a label = “High”
8. Click below the label and add “0.00007” and “1” in the boxes

9. Click “OK”

10. Creating Large Orders: repeat step 2 to 9 to end up with a menu like the one shown below