Analytics Proc in Enterprise Guide and SAS Studio

Pia Rønnevik, FANS, Customer Success Manager

pia.roennevik@sas.com

Descriptiv statistics

- SGSCATTER (Creates a paneled graph of scatter plots for multiple combinations of variables)
- MEANS (Data summarization tools to compute descriptive statistics for variables across all observations and within groups of observations)
- UNIVARIATE (Produces a variety of statistics that summarize the data distribution of each analysis variable)
- FREQ (One-way to *n*-way frequency and contingency (crosstabulation) tables)
- CORR (Computes Pearson correlation coefficients, three nonparametric measures of association, polyserial correlation coefficients, and the probabilities associated with these statistics)
- TTEST (Performs *t* tests and computes confidence limits for one sample, paired observations, two independent samples, and the AB/BA crossover design)

Visual Data – graph

https://go.documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/grstatproc/n0yjdd910dh59zn1toodgupaj4v9.htm

- SGPLOT (Bar Chart, *vbar*)
- SGPLOT (Bar-Line Chart, *vbar/vline*)
- SGPLOT (Boxplot, *vbox*)
- SGPLOT (Bubble Plot, *bubble*)
- SGPLOT (Heat Map, *heatmap*)
- SGPLOT (Histogram, *histogram*)
- SGPLOT (Line Chart, *vline*)
- FREQ (Mosaic Plot, *MosaicPlot*)
- TEMPLATE (Pie Chart, *piechart*)
- SGPLOT (Scatter Plot, *scatter*)
- SGPLOT (Series Plot, *series*)

Visual Data – Map

- SGMAP (Choropleth Map, *choromap*)
- SGMAP (Text Map, *text*)
- SGMAP (Bubble Map, *bubble*)
- SGMAP (Scatter Map, *scatter*)
- SGMAP (Series Map, *series*)

Visual Data – control charts

- SHEWHART (Box Chart)
- SHEWHART (C Chart)
- SHEWHART (Individual Measurements Chart)
- SHEWHART (Mean and Range Chart)
- SHEWHART (Mean and Standard Deviation Chart)
- SHEWHART (np Chart)
- SHEWHART (p Chart)
- SHEWHART (u Chart)

https://support.sas.com/documentation/onlinedoc/qc/141/shewhart.pdf

http://www.math.wpi.edu/saspdf/qc/chap31.pdf

Regression Models

- GLMSELECT/REG (Linear Regression with classification/continous variables)
- COUNTREG (Count Regression in which the dependent variable takes nonnegative integer or count values)
- GLM (One- Way Anova with categorical variables)
- GLM (N-Way Anova with factors)
- GLM (Analysis of Covariance with categorical variables/continuous covariate)
- NPAR1WAY (Nonparameteric One-Way ANOVA with classification variable)
- LOGISTIC (Binary Logistic Regression with classification/continuous variables, Link function: Logit, Probit and LogLog)
- GLMSELECT (Predictive Regression Models with classification/continuous variables)
- GENMOD (Generalized Linear Models with classification/ continuous variables, distributions: Normal, Binomial, Gamma, Inverse Gaussian, Multinomial, Negative binomial, Poisson, Tweedie, Zero-inflated negative binomial/Poisson)
- GAM (Generalized additive models)
- MIXED (Mixed Models with classification and continuous variables (random and fixed effects))
- PLS (Partial Least Squares Regression with classification/continuous variables)
- ROBUSTREG (Robust Regression with different methods: M Estimation, LTS Estimation, S Estimation, MM Estimation and M Estimation (tuned)
- ENTROPY (Entropy used to estimation of simultaneous systems of linear regression models)
- MDC (Multinomial Discrete Choice Modeling is used when the dependent variable takes multiple discrete values)

Survival Analysis

- LIFETEST (Nonparametric Survival Analysis)
- PHREG (Cox Proportional Hazards Regression)

Forecasting

- TIMEDATA (Time Series Data Preparation)
- TIMESERIES (Time Series Exploration)
- ARIMA (Modeling and Forecasting Random walk/Moving average/Arima/Arimax)
- ESM (Exponential Smoothing Models)
- UCM (Unobserved Components Models forecasts equally spaced univariate time series, decomposes the response series into components such as trend, seasonals, cycles, and the regression effects due to predictor serie)

Econometrics

- MODEL (Causal Models with exogenous/endogenous/excluded instrumental variables)
- AUTOREG (Cross-sectional Data Linear Models, Regression with autocorrelated and hetroscedastic errors)
- GLIM (Cross-sectional Data Logit/Probit/Censored/Truncated Models)
- PANEL (Panel Data Linear Models with cross-sectional/time ID)
- GLIM (Panel Data Logit/Probit/Censored/Truncated Models)
- COUNTREG (Panel Data Poisson/Negative Binomial Models)
- COUNTREG (Cross-sectional Data Poisson/Poisson Zero-inflated/ Negative Binomial/Negative Binomial Zero-inflated)
- SEVERITY (Severity Models continuous and categorical variables)
- SPATIALREG (Spatial Regression Models with continuous and categorical variables, analyzes spatial econometric models for cross-sectional data whose observations are spatially referenced or georeferenced)
- ARIMA (Univariate Time Series Analysis ARIMA/ARIMAX)
- VARMAX (Multivariate Time Series Analysis, variables arn't only contemporaneously correlated with each other, but also with each other's past values)
- PDLREG (Estimates regression models for time series data in which the effects of some of the regressor variables are distributed across time)
- TSCREG (Time Series Cross Section Regression , panel data sets that consist of time series observations on each of several cross-sectional units)

Multivariate Analysis

- PRINCOMP (Principal Component Analysis)
- FACTOR (Factor Analysis)
- CANCORR (Canonical Correlation)
- DISCRIM (Discriminant Analysis)
- CORRESP (Correspondence Analysis)
- PRINQUAL (Multidimensional Preference Analysis)
- COPULA (COPULA enables you to fit multivariate distributions or copulas from a given sample data set)
- EXPAND (Expand is useful when you need to combine series with different sampling intervals into a single data set)
- SIMILARITY (Computes similarity measures associated with time-stamped data, time series, and other sequentially ordered numeric data)
- SPECTRA (Spectral and cross-spectral analysis of time series, used to look for periodicities or cyclical patterns in data)

Cluster Analysis

- DISTANCE (Compute Similarities and Distances)
- VARCLUS (Cluster Variables)
- STDIZE/FASTCLUS (K-Means Clustering)
- DISTANCE/CLUSTER (Cluster Observations)
- ACECLUS (Estimate Within-Cluster Covariances)

High-Performance Models

- HPCOUNTREG (High-performance Count Regression in which the dependent variable takes on nonnegative integer or count values)
- HPPANEL (High-performance Panel analyze a class of linear econometric models that commonly arise when time series and cross-sectional data are combined.
- HPCDM (High-performance Compound distribution Model are modeling the severity of loss and the frequency of loss separately)
- HPQLIM (High-performance Qualitative and Limited dependent variable model analyzes univariate limited dependent variable models)
- HPSEVERITY (High-performance Severity provides a default set of probability distribution models; Burr, exponential, gamma, generalized Pareto, inverse Gaussian, lognormal, Pareto, Tweedie, and Weibull distributions)

Other Models

- QLIM (Qualitative and Limited dependent variable Model, univariate and multivariate limited dependent variable models, include logit, probit, tobit, selection, and multivariate models)
- SSM (State Space Models used for analyzing continuous response variables that are recorded sequentially according to a numeric indexing variable)

Simulation

- HPCOPULA (High-Performance Copula is a high-performance version of the SAS/ETS COPULA procedure, which simulates data from a specified Copula)
- SIMLIN (Perform simulation or forecasting of the endogenous variables)

SAS Visual Statistics in SAS Viya

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

SAS Visual Data Mining and Machine Learning in Viya

Machine Learning Techniques (Visual)

- Bayesian Network
- Factorization Machine
- Forest
- Gradient Boosting
- Neural Network
- Support Vector Machine

Machine Learning Procedures (SAS Studio Programmatic Interface)

- FACTMAC (Factorization Machine Model)
- FOREST (Forest Model)
- GRADBOOST (Gradient Boosting Model)
- NNET (Neural Network)
- SVMACHINE (Support Vector Machine)
- SVDD (Support Vector Data Description)
- BNET (Bayesian Network)
- BOOLRULE (Boolean Rules)
- FASTKNN (k-nearest neighbor)
- GVARCLUS (Variable Clustering and Graphical Modeling)
- MBANALYSIS (Association Rule Mining)
- RPCA (Robust Principal Component Analysis)

Examples

Regression Models

Linear Regression with classification and continous variables

```
proc glmselect data=SASHELP.CARS  outdesign(addinputvars)=Work.reg_design;
        class Origin / param=glm;
        model MPG_Highway=EngineSize  Cylinders Horsepower  Origin/
showpvalues
                selection=none;
run;

proc reg data=Work.reg_design  alpha=0.05  plots(only)=(diagnostics residuals
                observedbypredicted);
        where  Origin is not missing;
        ods select DiagnosticsPanel ResidualPlot ObservedByPredicted;
        model  MPG_Highway=&_GLSMOD  /;
        run;
quit;
```

One-Way Anova with categorical variable

```
proc glm data=SASHELP.CARS;
        class Origin;
        model MPG_Highway=Origin;
        means Origin / hovtest=levene welch plots=none;
        lsmeans Origin / adjust=tukey pdiff alpha=.05;
        run;
quit;
```

Nonparameteric One-Way ANOVA with classification variable

```
proc npar1way data=SASHELP.CARS wilcoxon plots(only)=(wilcoxonboxplot);
        class Origin;
        var MPG_Highway;
run;
```

N-Way Anova with factors

```
proc glm data=SASHELP.CARS;
        class Make Model Origin Type;
        model MPG_Highway=Make Model Origin Type Model*Origin
                        Model*Type Origin*Type  Model*Origin*Type
                        / ss1 ss3;
        lsmeans Make Model Origin Type / adjust=tukey pdiff=all alpha=0.05 cl;
quit;
```

Analysis of Covariance with Categorical variables and continuous covariate

```
proc stdize data=SASHELP.CARS method=mean out=work._ancova_stdize;
        var Cylinders;
run;

proc glm data=work._ancova_stdize;
        class Origin;
        model MPG_Highway=Origin Cylinders Cylinders * Origin;
        lsmeans Origin / adjust=tukey pdiff alpha=.05;
quit;
```

Regression Models

Binary Logistic Regression with classification **variables**

```
proc logistic data=CREDIT_DISCOVERY_FOR_DS_DATA;
        class CREDIT_LIM CREDIT_SCORE / param=glm;
        model writeoff(event='YES')=CREDIT_LIM CREDIT_SCORE / link=logit
technique=fisher;
run;
```

Predictive Regression Models with classification **variables**

```
proc glmselect data=SASHELP.CARS plots=(criterionpanel);
        class Origin Make / param=glm;
        model MPG_Highway=Origin Make EngineSize Cylinders Horsepower /
                selection=stepwise
(select=sbc) hierarchy=single;
run;
```

Generalized Linear Models with classification variables

```
proc genmod data=SASHELP.CARS plots=(predicted resraw(index) stdreschi(index) );
        class Make Origin / param=glm;
        model MPG_Highway=Make Origin EngineSize Cylinders Horsepower /
dist=normal;
run;
```

Mixed Models with classification and continuous variables (random and fixed effects)

```
proc mixed data=SASHELP.CARS method=reml
plots=(residualPanel) alpha=0.05;
        class Make Origin;
        model MPG_Highway= /;
        random Intercept / type=VC subject=Make;
run;
```

Partial Least Squares Regression **with classification variables**

```
proc pls  data=SASHELP.CARS method=pls plots;
        class Make Origin;
        model MPG_Highway=Make Origin EngineSize Cylinders
        Horsepower EngineSize*Cylinders EngineSize*Horsepower
        Cylinders*Horsepower;
run;
```

Forecasting

```
proc sort data=PUBLIC.DATA_FORECAST out=Work.preProcessedData;
        by price discount cost Txn_Month;  run;
```

ARIMAX

```
proc arima data=Work.preProcessedData plots
(only)=(series(corr crosscorr) residual(corr normal) forecast(forecastonly));
        identify var=sale crosscorr=(line product);
        estimate p=(1) (12)  q=(1) input=(line product) method=ML;
        forecast lead=12 back=0 alpha=0.05  id=Txn_Month interval=Month;
        outlier;
        by price discount cost;
        run;
quit;
```

ESM

```
proc esm data=Work.preProcessedData back=0 lead=12 plot=(corr errors
                modelforecasts);
        by price discount cost;
        id Txn_Month interval=Month;
        forecast sale / alpha=0.05  model=simple transform=none;
run;
```

UCM

```
proc ucm data=Work.preProcessedData;
        id Txn_Month interval=Month;
        model sale;
        irregular;
        level;
        forecast lead=12 back=0 alpha=0.05;
        outlier;
        by price discount cost;
run;
```

Multivariate Analysis

Principal Component Analysis

```
proc princomp data=SASHELP.CARS plots(only)=(scree);
        var EngineSize Cylinders Horsepower Invoice Weight Length;
run;
```

Factor Analysis

```
proc factor data=SASHELP.CARS method=principal nfactors=7 plots=(scree);
        var MSRP Invoice EngineSize Cylinders Horsepower Weight Length;
run;
```

Canonical Correlation

```
proc cancorr data=SASHELP.CARS;
        /*** The VAR statement defines Variable set 1 ***/
        var EngineSize;

        /*** The WITH statement defines Variable set 2 ***/
        with Cylinders;
run;
```

Discriminant Analysis

```
ods noproctitle;

proc discrim data=SASHELP.CARS pool=yes;
        class Model;
        var Invoice EngineSize Cylinders Horsepower;
        priors prop;
run;
```

Correspondence Analysis

```
proc corresp data=SASHELP.CARS dimens=2 plots;
        tables Make Origin, Invoice EngineSize Cylinders Horsepower;
run;
```

Multidimensional Preference Analysis

```
proc prinqual data=SASHELP.CARS mdpref n=2 plots
out=Work.Prinqual_Scores
                replace;
        transform monotone(EngineSize Cylinders Horsepower);
run;
```

Cluster Analysis

Compute Similarities and Distances

```
proc distance data=SASHELP.CARS method=dgower out=work.Distance_dist;
        var interval(MPG_Highway / std=range) ordinal(EngineSize Cylinders
Horsepower / std=range) nominal(Origin Model);
run;
```

Cluster Variables

```
proc varclus data=SASHELP.CARS hierarchy plots;
        var EngineSize Cylinders Horsepower Weight Length;
run;
```

K-Means Clustering

```
proc stdize data=SASHELP.CARS out=Work._std_ method=range;
        var EngineSize Cylinders Horsepower Weight Length;
run;

proc fastclus data=Work._std_ maxclusters=100;
        var EngineSize Cylinders Horsepower Weight Length;
run;
```

Cluster Observations

```
proc distance data=SASHELP.CARS method=dgower
out=Work._tmp_distances;
        var interval(EngineSize Cylinders Horsepower / std=std)
ordinal(Invoice Weight
                Length / std=std) nominal(Make Model Type Origin);
run;

proc cluster data=Work._tmp_distances method=ward plots;
        var Dist:;
run;
```

Estimate Within-Cluster Covariances

```
proc aceclus data=SASHELP.CARS proportion=0.1;
        var EngineSize Cylinders Horsepower;
run;
```