Statistical modelling Using SAS

A short course

By

Hasan Mytkolli*, PhD

Dragos Calitoiu*, PhD

June 22, 2009

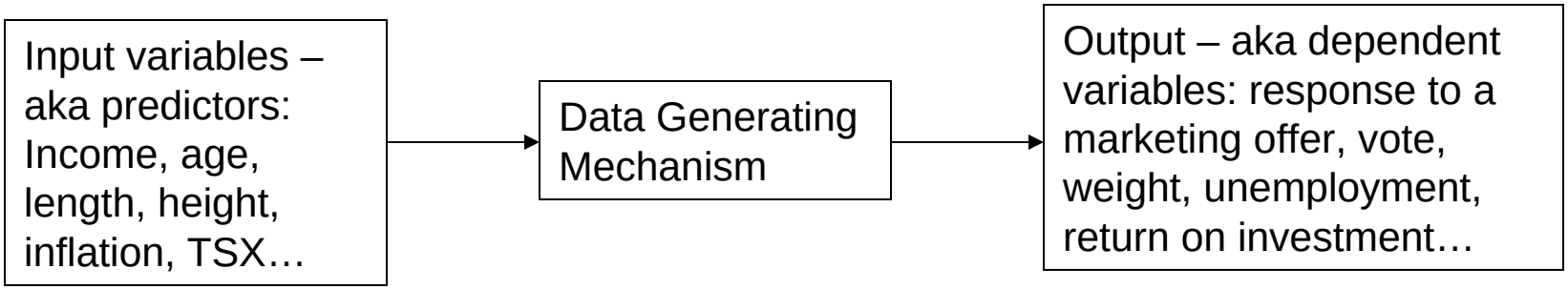
*) The authors of this presentation are the modellers for Bank of America – Canada Card Services. Opinions, ideas and everything said in this presentation are of the authors and should not be confounded with their position in the Bank.

The authors are also members of CORS (Canadian Operational Research Society) and OPTIMOD Research Institute.

Contact: Hasan.Mytkolli@mbna.com and Dragos.Calitoiu@mbna.com

Statistical modelling – two cultures

- Statistical analysis and modelling involves the appropriate application of statistical analysis techniques, each requiring certain assumptions be met, to perform hypothesis tests, interpret the data, and reach valid conclusions.
- Basically, there are two goals when analyzing data:
  - Prediction. By analyzing the past, one assumes that conclusions drawn can be used to predict the future.
  - Inferential. In this case, one may be interested to investigate the nature of the relationship between different sides of a complex phenomenon.
- No matter what the goals of the analysis are, we assume that we have a set of output variables, a set of input variables and an unknown mechanism that relates the output with the input. We prefer to call this the Data Generating Mechanism:



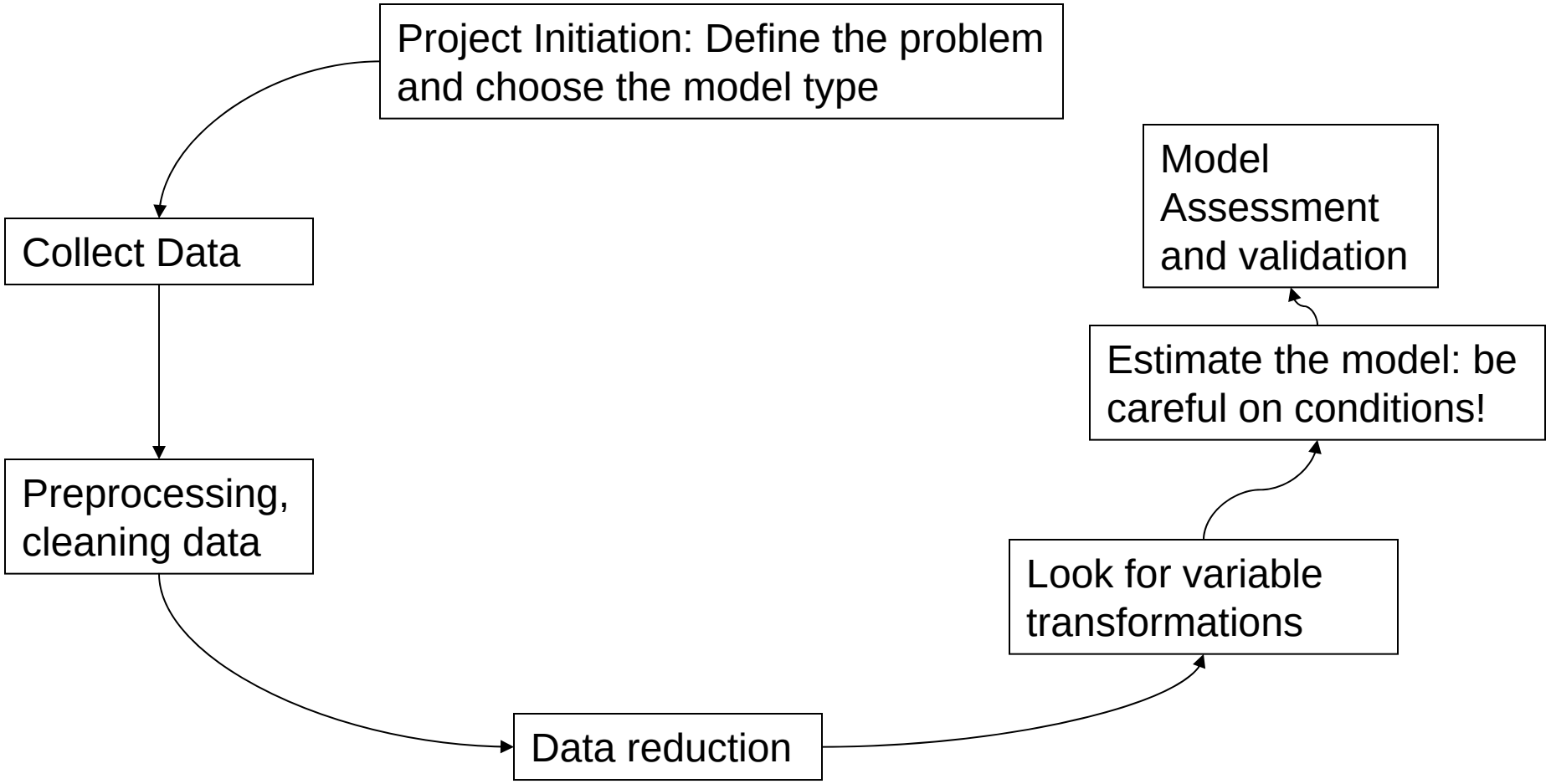
Statistical modelling – two cultures

- Speaking about the middle box – DGM, statisticians are divided into two main camps:
    - Stochastic modelling. One assumes that the relationship between the input and output is driven from a stochastic process: linear regression, logistic, regression, Cox model, …
    - The algorithmic approach. According to this approach, the relationship is too complex and unknown. Instead of a concrete equation, this approach looks for an algorithm that can help predict the future of the output from the input: Decision Trees, Neural nets, Genetic models, …

modelling and Statistical Software

- No matter which camp you decide to sit in, statistical analysis and modelling requires careful selection of analytic techniques, verification of assumptions, and verification of data. Descriptive statistics, graphs, and relational plots of the data should first be examined to evaluate the legitimacy of the data, identify possible outliers and assumption violations, and form preliminary ideas on variable relationships for modelling.
- The vast amount of data, information contained in the data, the complex nature of the calculations related to any of the above mentionned models requires appropriate software: SAS, SPSS, MINITAB, SPLUS, Statistica…
- The authors of this presentation consider SAS being in the forefront, especially when big and complex problems are to be analyzed.

Stages of a typical modelling process



Don't forget: the key to a successful model is to feel the data, to live with your data!

Project Initiation: Define the problem and the type of model

- “Well begun is halfway done”. This does not work very well in statistical modelling!
- “Get off on the wrong foot”. Yes, this is absolutely true in statistical modelling as well! A bad start is a guarantee for failure!
- The problems a modeller faces could come from different sources: from a business area, from a partner, from outside the organization, it could be a research topic of interest...
- With the exception of the last case, we can easily call the requester a customer.
- It is the duty of the modeller to clarify the customer’s needs.
- An example. You are assigned to build an customer attrition model. What will be the definition of attrition?

Define the problem – an example

- The dictionary definition for customer attrition – “a business term used to describe loss of clients or customers” is too generic and almost useless when it comes to modelling:
  - It could perfectly fit the situation in telecommunications when the customers switch service providers or close their accounts, but what about the customers who reduce their usage?
  - The situation is alike in banking industry. The above definition does not tell us anything for the lost opportunities from the “silent” attrition.
  - The situation is even more confused when considering the retail industry.
- The above arguments show that building a model is more than an “one person” job. It is the customer or the client who posses the problem, but it is modeller’s responsibility to clarify upfront what the objective of the modelling process is.
- Unless you are doing some “pure” scientific research:
  - You cannot build a good model without understanding the business requirements.

Identify the model type

- Most of the time, the problem identification will pave the way for model identification.
- The model could be from the simplest to the most advanced one:
  - Decision Trees
  - Cluster analysis
  - Regression (linear, logistic, survival…)
  - Neural nets
  - Genetic models
  - Pattern recognition
  - …
- SAS has many resources you can choose to use:
  - Base SAS (for those who like to do programming or have their say in the model
  - SAS Enterprise Miner
  - SAS OR

Data collection

- Identify the variables with interest
  - Rule number 1: Do not take sides – include any variable that may effect the outcome.
  - Rule number 2: Consult your client – sometimes they have valuable experience that can help you when building the model and…
    - Don’t forget, you have to “sell” your model. Involving the client makes it an easy “sale”.

Data collection

- Most likely your data is in different tables, maybe even in different physical places (databases, servers…)
- You can use SAS SQL to build the development datasets.
- A typical query that queries a database has the following structure:

```
proc sql;
        connect to DBMS-name (connection statements);
        select
        column list
        from connection to DBMS-name
        (
        DBMS-query
        )
disconnect from DBMS-name;
quit;
```

Preprocessing, Cleaning data

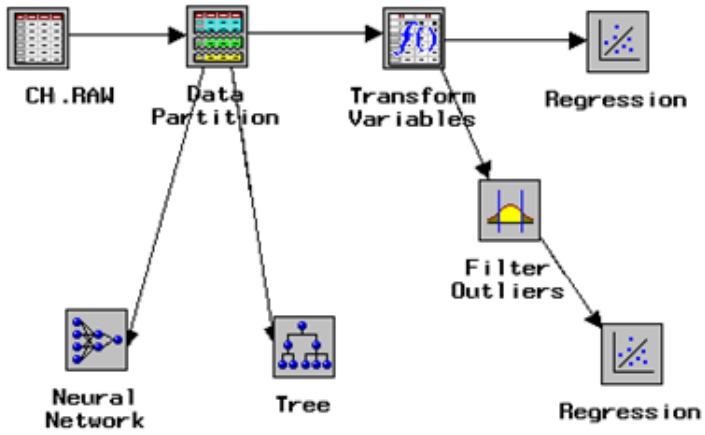
- So, you have a table, but you are still not ready to start the fun work of building the model.
- Probably the first thing to do is some checks:
  - Does your table have the number of observations you were expecting?
  - How about the missing? Where do they come from? Is it the information in the source tables that has missing or some potential mistakes while querying?
- Once you are sure you have the right table, you can do some preprocessing:
  - Decide what you will do with missing data. Your options are:
    - discard the variable – this mostly happens when there are too many “unjustified” missing.
    - data imputation. You can choose between building your own rule or use Proc MI.
  - Decide what you will do with the outliers. Your options are :
    - Remove outliers
    - Do nothing
    - Transform (you will hear more on that during Dragos Calitoiu presentation)

Data reduction and Variable transformations

- You will hear more for this two stages from Dragos Calitoiu.

Finalizing your Model

- Most of the models listed in slide 8 can be easily built using the Enterprise Miner. SAS Institute has dedicated courses on EM so we are not going to spend any time on that.



- Our presentation is focused on utilizing Base SAS in building the models (people with no access to EM or those who want to get their hands dirty).

Linear Regression

- We assume that you have to model the relationship between a continuous dependent variable and independent variables, the functional form being:

  Y=b0+b1*x1+b2*x2+…bk*xk+e

  We assume that error terms are normally iid.

- You can use either proc reg or proc glm to build and test assumptions of your model.
  - Proc reg is probably the most known one. However, PROC REG has some limitations as to how the variables in your model must be set up, not handling the interactions being one of the most important.
  - Example:

```
PROC REG DATA=yourdata options;
MODEL depvar = indvar1 indvar2 …. /options;
Output out=residuals r=resid;
RUN;
```

    - There are lots of options, we would like to emphasize those that will help you to “calibrate” your model:
      - Selection: stepwise, forward, backward, adjrsq
      - Multicollinearity: VIF, COLLIN
      - Outliers: Influence R
      - Model Fit: Lackfit
      - Residual related: Dw spec – can be used to test for iid

```
PROC UNIVARIATE DATA=RESIDS
NORMAL PLOT;
VAR RES;
RUN;
```

    - If you want to introduce interactions, you have to do it prior through a data step.

Linear Regression continued

- Proc GLM is the other option and has some advantages compared to proc reg.
- Proc GLM allows you to write interaction terms and categorical variables (even if they are formatted as character) with more than two levels directly into the MODEL.
- The “negative” side of proc glm is that it does not offer you the opportunity to test the way you can do in proc reg.
- Example:

```
PROC GLM DATA=yourdata options;
MODEL depvar = indvar1 indvar2 ….indvark indvari*indvarj;
Output out=residuals r=resid;
RUN;
```

- Comparison between REG and GLM:
  - In some ways, proc glm is superior to proc reg because proc glm allows manipulations in the model statement (such as variable interactions) which are not allowed in proc reg.
  - However, proc reg allows certain automatic model selection features and a crude plotting feature not available in proc glm.
- Some words of wisdom:
  - If the interaction of two or more variables comes to be significant and you decide to keep it in the final model, then it is advised to keep the respective variables as well.
  - When doing real data modelling, especially those related to human activity, the variability of dependent variables could be huge. This is good from the modelling perspective (we model variability), but it is rare that you will get models that fit R-square at 90% and above. Sometimes the R-square could be even less then 50%. The decision to call it a good model or not depends on the objective of the model.
  - One way to improve the model fit is to consider piecewise regression – fitting different models for different parts of the data (where there is more data homogeneity).
- Finally, if you really want to enter the world of modelling, calculations, algorithms, SAS has a wonderful tool, Proc IML.

Logistic Regression

- We suppose we are trying to model the outcome of a Bernuli (yes –no) trial. Clearly, we can not apply the linear regression (the errors will be either 0 or 1).
- Instead, we model:

$$\log it = \log\left(\frac{p}{1-p}\right) = \sum \beta_i x_i$$

which yields the probability estimation:

$$p = \frac{\exp\left(\sum \beta_i x_i\right)}{1+\exp\left(\sum \beta_i x_i\right)}$$

- SAS has different options to estimate logistic regression: proc logistic, proc genmod, proc catmod, proc nlmixed.
- Each of the above procs can do much of the same thing, while
  - proc logistic can handle the ROC curve and can perform the exact logistic regression (the case of small sample datasets).
  - proc genmod presents a unified approach to the analysis of categorical data, including Poisson and Negative Binomial (for counts), gamma, and normally distributed data.
  - Proc catmod is better on discrete type of independent variables. If you have continuous independent variables then you better use proc logistic or proc genmod.
  - Proc nlmixed gives lots of options in programming.

Logistic regression – proc logistic

- Proc logistic has many options that gives you the opportunity to produce lots of output. A simple call of the proc logistic could look like:

```
Proc logistic data=yourdata options;
Class varclass;
Model depvar=indvar1 indvar2…indvark/options;
Output out=dataset;
Score yoursecondata out=scoreddata;
Run;
```

- The depvar could be related to single trials (bad, live, resp…) or multi trials. In this last case the depvar is comprised by two variables (k/n): k – the number of occurencies and n – the number of trials. Different from proc reg, proc logistic accepts the interactions, but no odds ratio is outputted in this case.

- The options that proceed the proc logistic are mostly related to the parameter estimations and the output related, the options at the model statement could be from the model selection (the same as at proc reg), to the model fit, model specification (logit, probit, mlogit), outroc…The new feature (SAS 9) is the option to score datasets within the estimation process.

- In addition to the above statements and options proc logistic has quite a healthy bunch of other options/statements that can help in solving different situations and produce a rich output.
- Proc logistic produces the parameter estimations (interval estimations included), the fit statistics (loglike, AIC, SBC, goodness of fit -a Hosmer and Lemeshow Test, a deciling partition, odds ratio, classification table…).
- The success in building a good model using the logistic regression is determined by the ability to read and carefully utilize the output information.
- The interpretation of the odds ratio matrix is very important from the model usage perspective.

Proc logistic – the output

- The following is a simplified example of the output from a call to proc logistic:

Response Profile		
Ordered Value	cluster_0	Total Frequency
1	1	80544
2	0	45781

Model Fit Statistics		
Criterion	Intercept	Intercept and covariates
AIC	165434.8	164880.98
SC	165444.5	164919.97
-2 Log L	165432.76	164872.98

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard error	Wald Chi Square	Pr > ChiSq
Intercept	1	0.1145	0.0225	25.9888	<.0001
bcx_var1	1	0.8952	0.1427	39.3401	<.0001
bcx_var2	1	-0.2919	0.1577	3.4251	0.0642
bcx_var3	1	1.7137	0.0956	321.2591	<.0001

Proc logistic – the output

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald CL**	
bcx_var1	2.448	1.851	3.238
bcx_var2	0.747	0.548	1.017
bcx_var3	5.549	4.601	6.693

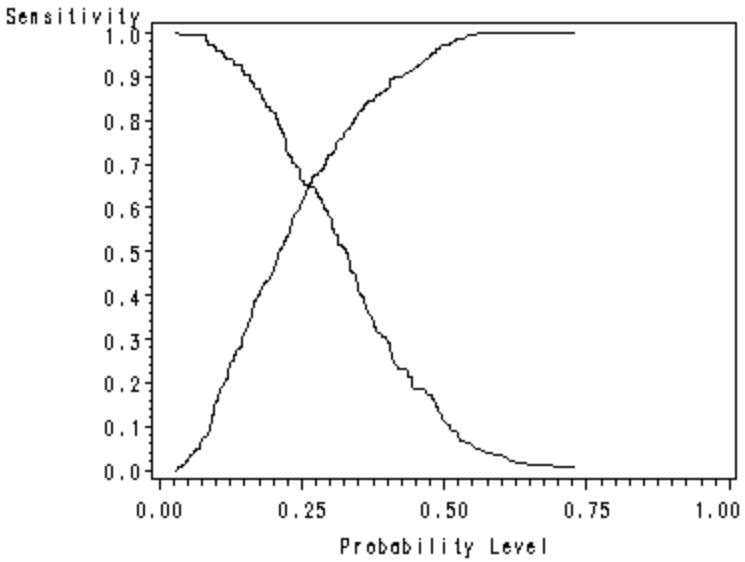
Partition for the Hosmer and Lemeshow Test					
Group	**Total**	**cluster_0 = 1**		**cluster_0 = 0**	
		Observed	**Expected**	**Observed**	**Expected**
1	12794	7197	7200.25	5597	5593.75
2	12613	8490	7973.25	4123	4639.75
3	5166	3332	3267.38	1834	1898.62
4	46831	29037	29619.69	17794	17211.31
5	12379	8506	8212.89	3873	4166.11
6	3284	2204	2179.83	1080	1104.17
7	22771	14902	15114.78	7869	7656.22
8	10487	6876	6975.84	3611	3511.16

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	**DF**	**Pr > ChiSq**
170.7696	6	<.0001

Classification table, ROC

- Proc logistic gives an opportunity to produce the classification table at different cut points. Each cut point defines a number of misclassified outcomes. Option outroc can produce the classification at the record level.
- Sensitivity and specificity are numbers from 0 to 1 that summarize the performance of a diagnostic test with a positive/negative outcome. Each subject has a true status which we might call responder or non-responder, and an estimated status, also responder or non-responder. Sensitivity and specificity are, respectively, the chance that a true responder and a true non-responder will be identified as such.
- You can plot sensitivity and specificity using proc plot:

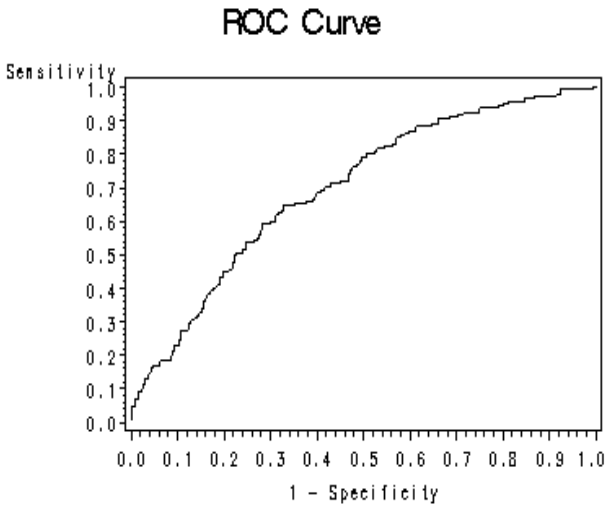
```
data roc; set roc;
spec = 1-_1mspec_;
run;
symbol1 i=join v=none ;
proc gplot data=roc2;
plot _sensit_*_PROB_=1 spec*_PROB_=1
/ overlay haxis=0 to 1 by .25 vaxis=0 to 1 by .1 ;
 run;
```



Classification table, ROC

- Receiver operating characteristic (ROC), or simply ROC curve, is a graphical plot of the sensitivity vs. (1 - specificity). The ROC can also be represented equivalently by plotting the fraction of true positives (TPR = true positive rate) vs. the fraction of false positives (FPR = false positive rate).
- You can plot the ROC using SAS:

```
title 'ROC Curve';
proc gplot data=roc1;
plot _sensit_*_1mspec_=1
/ vaxis=0 to 1 by .1 ;
run;
```



- The most common term in the business related applications for the ROC is Lift Curve or Gini Curve.
- The surface of the area between the ROC curve and the main diagonal multiplied by 2 is called Gini Index and is often used to measure the impact of the model.
- The most common way marketing researchers use to produce the Lift Curve is by deciling. The whole dataset is sorted by predicted probability and then equally split in ten groups – deciles. Next step is to count the cumulative percentage of the “good” for each decile.

Discrimination Power: K-S

- Another advantage of deciling is that it can be used to calculate another very important model performance indicator: k-s (kollmogorov-smirnov) statistics.
- Indeed, in this case you need to count for each decile cumulative percentage of “good” and “bad”. K-s is just the maximum distance between two cumulative distributions and measures how powerful the model is in discriminating the “good” from the “bad”.
- You can use SAS code to produce the k-s statistics:

```
Proc npar1way data=yourdata ks;
Class good;
Var p_1;
Run;
```

Model Validation, PSI

- A model is built either for academic/research purposes or for business usage.
- Most often a business oriented model is used to forecast. The goodness of the model can be measured via its ability to correctly predict the future.
- We believe that we have built a stable model via fulfilling all the statistical requirements.
- The problem is that often the sample errors could be interpreted as real trends and become part of the model. If this is the case, then the very next time we apply the model we may find big discrepancies between the performance of the model in the development and the new datasets.
- Validation of the model is advised to protect from such a situation.
- It is strongly advised that each time you build a model, you randomly split the data in two samples: development and validation datasets.

Model Validation, PSI (II)

- Once you have finalized your model you should produce the same reports and performance indicators for validation datasets and compare them with those from development datasets. Big differences should ring the alarm bell for potential problems with your model.

- Population Stability index is a well known indicator that measures the differences of the model performance in the development and validation datasets. Both datasets need to be sorted and deciled. Then the differencies between decile percentages for the “good” is multiplied by the natural logarithm of the differences ratio; the results are summed up to produce PSI:

$$PSI = \sum \left(f^{i}_{development} - f^{i}_{validation} \right) * \ln\left(f^{i}_{development} / f^{i}_{validation} \right)$$

So how good is your model?

- So you have done such a lot of work and you are happily coming to an end. But did you meet the objectives?
- So, “the ends meet”. At the end, you have to go back to beginning. The objective of your work is to meet the requirements of your client.
- You have to present the results in a language that your client understands.
- You have to make your model a business tool. It is your responsibility to care for the preparation of the model for application and to further monitor the way it is used.