Missing Data, Imputation and Variable Selection in Multivariate Modeling

Shankar Srinivasan, Arlene Swern, Pavel Kiselev and Albert Elion-Mboussa

2015 BASS Conference, 2nd November

Rockville, Maryland

Motivation and Objective

- Registries – non-interventional studies
  - collect data on procedures and assessments which are considered part of standard practice.
- Leads to a lot of missing data.
- Even if 10-15% is missing per covariate, easily leads to only 40-50% of complete cases.

Objective:

Present a multivariate modeling approach involving variable selection that uses multiple imputation, model presentation and internal and external validation.

Outline

- **Example – Discrete Case: Analysis of deaths within 180 days. (Note that the actual results are replaced with dummy data in data presentation). Logistic regression Modeling**
  - **Analysis with missing data, imputation and variable selection**
  - **Model Representation**
  - **Internal Validation**
  - **External Validation**
- **Example - Continuous Case: Analysis of MM Disease Registry baseline QOL endpoints. Clinical baseline factors as predictive of EQ5D index.**
  - **Introduction**
  - **Multivariate model selection**
  - **Results**
- **Conclusions**

General Approach to Multivariate Logistic Analyses: Endpoint=Death within 180 days (yes, no)

Univariate Logistic Regression analysis to determine significant variables to be entered into Multivariate analysis.

10 datasets created using multiple imputation.

Datasets stacked and variable selection done using underweighted observations.

10 Unstacked multivariate logistic analyses done for the variables selected.

Inferences combined using Rubin's method to obtain estimates and p-values.

Multiple Imputation Framework

1a. Imputation Model

$$\text{Data}: \boldsymbol{Y} = (\boldsymbol{Y}_{mis}, \boldsymbol{Y}_{obs})$$

$$= (\boldsymbol{Y}_{1obs}, \boldsymbol{Y}_{1mis}, \boldsymbol{Y}_{2obs}, \boldsymbol{Y}_{2mis}, \boldsymbol{Y}_{3obs}, \boldsymbol{Y}_{3mis})$$

$$\text{Distribution}: f(y \mid \theta)$$

1b. Impute the Missing Values

$$f(\hat{\boldsymbol{Y}}_{mis} \mid \boldsymbol{Y}_{obs}, \hat{\theta})$$

1c. Produce Multiple Imputations

$$\boldsymbol{Y}_{(1)} = (\hat{\boldsymbol{Y}}_{mis(1)}, \boldsymbol{Y}_{obs}) \qquad \boldsymbol{Y}_{(m)} = (\hat{\boldsymbol{Y}}_{mis(m)}, \boldsymbol{Y}_{obs}) \qquad \boldsymbol{Y}_{(M)} = (\hat{\boldsymbol{Y}}_{mis(M)}, \boldsymbol{Y}_{obs})$$

Reproduced From: Berglund and Heeringa (2014) MI using SAS. SAS Publ.

SAS Code: Multiple Imputation

Default: MCMC Multivariate Normal

- **proc mi data =Edeath nimpute = 10 seed = 651467 out=Edeathm ;**
  var age issstage ecog IMWG_risk mhdiabn Beta2_M mhhyn mobility VTE plat_ct…… d180; *(partial list of Variables);
  run;

Recoding of discrete variables (Can also use round option of the MI procedure)

- if mhdiabn le 0.5 then mhdiabn = 0;
- if mhdiabn gt 0.5 then mhdiabn = 1;
- if mhhyn le 0.5 then mhhyn = 0;
- if mhhyn gt 0.5 then mhhyn = 1;

Bernaards et. al. (2007) Statistics in Medicine 26:1368–1382

Procedure: Model Selection For Logistic Regression

Selection = score in SAS provides the score statistic for all possible models.

Difference in score statistic - a chi-squared distribution, with degrees of freedom given by the difference in the number of variables in the model.

Starting with best 1 variable model,
move in 1 variable increments to the best k variable model,
till the incremental score statistic is less than the critical value.

Select that best k variable model.

Theory: Stacked Weighted analysis. Wood et al

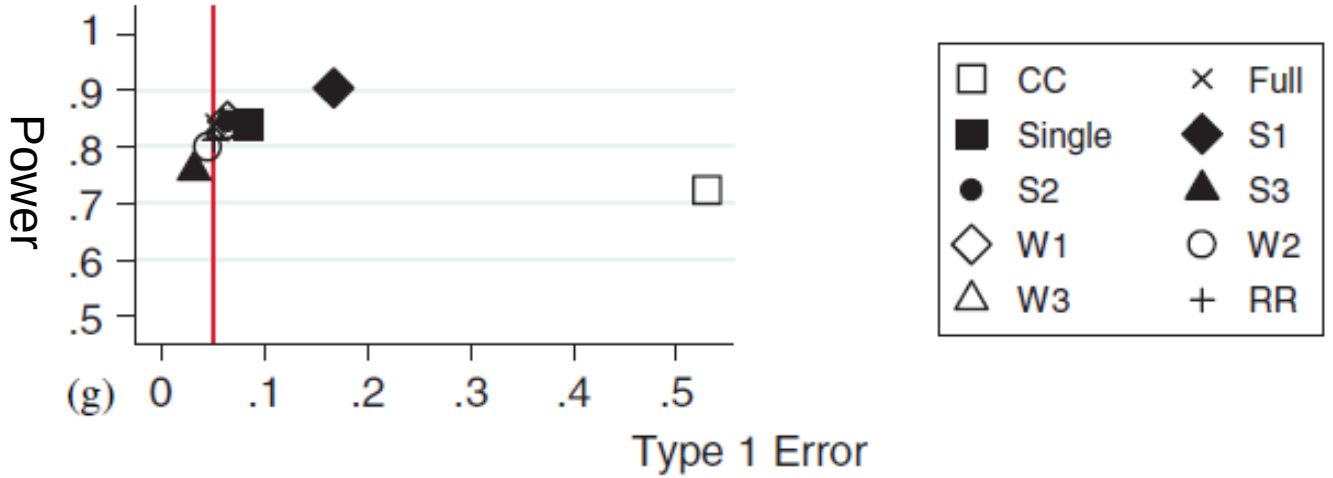
They Considered:

- Complete Case (CC)
- Single Stochastic Imputation (Single)
- Separate Imputations
  - S1: select predictors that appear in any model;
  - S2: select predictors that appear in at least half of the models;
  - S3: select predictors that appear in all models.
- Weighted Analysis
  - W1: $w_i = 1/M.$
  - W2: $w_i = (1 - f)/M$
  - W3: $w_i = (1 - f_i)/M$
- Rubin's Method (RR)

Wood et. al (2008). ***Statist. Med. 2008; 27:3227–3246***

Theory: Stacked Weighted analysis. Wood et al

❖Power defined as the probability of correctly selecting each of the variables in true model.
❖Type 1 error defined as the probability of wrongly selecting variables not in the true model.
❖Simulations find that W2 and W3 are close in type I error and power to the ideal but difficult to implement RR method.



SAS Code: Variable Selection

- Datasets Stacked and the following CODE. Weight = **0.9137/10**

- **proc logistic data = Edeathm2 ;**

```
    model d180 = age issstage  mobility IMWG_risk  ecog
       Beta2_M VTE bm calcium creat plat_ct clcr
       /selection = score

       details lackfit ;

       weight wt;

run;
```

SAS Code: Logistic regression by Imputation

- **Obtain parameter estimates separately for reduced model obtained by the stacked weighted logistic regression (and clinically meaningful).**

- **proc logistic data=Edeathm2;**

  model dthbf180 (event = 'Yes') = age ecog VTE mobility ISS plat_ct

  /risklimits details lackfit covb;

  by _Imputation_;

  ods output ParameterEstimates=lgparms CovB=lgcovb;

  run;

Rubin's Inferences behind PROC MI

The MI estimates of the vector of parameters

$$\bar{\theta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m$$

Within Imputation Variance

$$\overline{W} = \frac{1}{M}\sum_{m=1}^{M} W_M$$

Between- imputation Variance

$$B = \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\theta}_M - \bar{\theta}\right)^2$$

Total variability of $\bar{\theta}$

$$T = \overline{W} + \frac{M+1}{M}B$$

SAS Code: Combine Using MIANALYSE

- **Combine the estimates from the separate imputations.**

- **proc mianalyze parms=lgparms**

  covb=lgcovb;

  modeleffects Intercept age ecog VTE mobility ISS plat_ct;

  ods output ParameterEstimates=est1;

  run;

Model Presentation

- Typical Model Presentation. Odds Ratios, Confidence Intervals and P-values.
  - data est1; set est1;
  - OR = exp(Estimate);
  - ORL = exp(LCLMean);
  - ORU = exp(UCLMean);
  - Run;
- Heat Map presentation to aid in physician-patient communication. Matrix presentation of predicted values

$$\theta_{hi} = \frac{\exp\left\{(\alpha + \sum_{k=1}^{l} \beta_k x_{hik})\right\}}{1 + \exp\left\{(\alpha + \sum_{k=1}^{l} \beta_k x_{hik})\right\}}$$

Matrix example: Vastesaeger N, *et al. Ann Rheum Dis (2011).*

Typical Model Presentation: Logistic Regression

Note: Data redacted with xx.x, estimates changed and variables switched

Celgene
Study MM-Connect

Table x.xx
Logistic Regression Analysis of Baseline Characteristics Associated with Deaths within 180 Days
All Registry Subjects Cohort-1

Characteristics	Univariate Analysis			Multivariate Analysis		
	Estimate	95% CI	P-value	Estimate	95% CI	P-value
Patient Specific						
Age (<=75 vs >75)	1.97	(1.26, 3.06)	0.003	1.69	(1.01, 2.74)	0.042
Body Mass Index	x.xx	(x.xx, x.xx)	x.xxx			
ECOG performance status score	x.xx	(x.xx, x.xx)	x.xxx	1.85	(1.12, 2.67)	0.041
History of diabetes	x.xx	(x.xx, x.xx)	x.xxx			
History of hypertension	x.xx	(x.xx, x.xx)	x.xxx			
History of venous thromboembolism (VTE)	x.xx	(x.xx, x.xx)	x.xxx	1.97	(1.06, 2.95)	0.022
Del(17P) from FISH and Cytogenetic forms	x.xx	(x.xx, x.xx)	x.xxx			
T(414) from FISH	x.xx	(x.xx, x.xx)	x.xxx			
Disease Specific						
Lactic Acid Hydrogena (<= 300 g/dL vs > 300 g/dL)	x.xx	(x.xx, x.xx)	x.xxx			
Extramedullary plasmacytoma (Yes/No)	x.xx	(x.xx, x.xx)	x.xxx			
Immunoglobulin IgG Class (< 5 g/dL vs >= 5 g/dL)	x.xx	(x.xx, x.xx)	x.xxx			
Albumin (<=3.5 g/dL vs >3.5 g/dL	x.xx	(x.xx, x.xx)	x.xxx			
ISS disease stage (calculated)	x.xx	(x.xx, x.xx)	x.xxx	2.19	(1.28, 3.51)	0.007
Myeloma bone involvement	x.xx	(x.xx, x.xx)	x.xxx			
Hypercalcemia (Serum Calcium >=11.5 mg/dL)	x.xx	(x.xx, x.xx)	x.xxx	1.68	(0.86, 3.08)	0.121
Renal Insufficiency (Serum Creatinine > 2 mg/dL)	x.xx	(x.xx, x.xx)	x.xxx			
Anemia (Hemoglobin < 10 g/dL or >2 below LLN)	x.xx	(x.xx, x.xx)	x.xxx			
Platelet Count (for 100 x 10^9L increments)	x.xx	(x.xx, x.xx)	x.xxx	1.96	(1.19, 2.90)	0.008
IMWG risk	x.xx	(x.xx, x.xx)	x.xxx			
Beta 2 Microglobulin (beta>=5.5 mg/L)	x.xx	(x.xx, x.xx)	x.xxx			
HRQol from Eq5d						
Self care from Eq5D	x.xx	(x.xx, x.xx)	x.xxx			
Mobility from Eq5D	x.xx	(x.xx, x.xx)	x.xxx	2.77	(1.65, 4.64)	<0.001
Novel Therapy						
Novel Therapy Use (0,1) vs >=2	x.xx	(x.xx, x.xx)	x.xxx			

Heat Map Presentation: Logistic Regression

Note: Estimates changed and variables switched and/or changed

Figure 1: Inputs to Prognostic Chart

Attributes	Select Value from Drop-Down List
1. Mobility	Some Problem in Walking About
2. Platelet Count in X10^9/L	<=150
3. ISS Stage	I, II
4. Age Group	<=75
5. ECOG Performance Poor	No
6. History of VTE	No

Figure 2: Prognostic Chart with Estimated Probability of Death Before 180 Days

		Confined to bed					Some problem in walking about					No problem in walking about			
Platelet>150	ISS I,II	12%	7%	8%	5%	No	5%	3%	3%	2%	No	2%	1%	1%	1%
		19%	13%	13%	8%	Yes	9%	6%	6%	4%	Yes	4%	3%	3%	2%
	ISS III	21%	14%	14%	9%	No	10%	7%	7%	4%	No	5%	3%	3%	2%
		33%	23%	23%	16%	Yes	17%	11%	12%	7%	Yes	8%	5%	5%	3%
		Age>75		Age<=75			Age>75		Age<=75			Age>75		Age<=75	
Platelet<=150	ISS I,II	20%	13%	13%	9%	No	9%	6%	6%	***4%***	No	4%	3%	3%	2%
		31%	21%	22%	15%	Yes	16%	11%	11%	7%	Yes	8%	5%	5%	3%
	ISS III	33%	23%	24%	16%	No	18%	11%	12%	8%	No	8%	5%	5%	3%
		48%	36%	36%	26%	Yes	28%	19%	20%	13%	Yes	14%	9%	9%	6%
		Yes	No	Yes	No		Yes	No	Yes	No		Yes	No	Yes	No
		VTE		VTE			VTE		VTE			VTE		VTE	

Selected Attributes are highlighted in Blue and the estimated probability is ***Bold, italics and underlined***. For a prognostic chart without the blue highlights or the marked up probablility enter blanks for the first two inputs.

Reading the Matrix

- Larger Blocks in the matrix are the larger effects.
- We move to smaller blocks within the larger blocks with factors which have succeeding smaller effects.
- Traffic Color coded. Green for lower probability of dying within 180 days, through Yellow to Red for higher probabilities.
- Designed to show higher risk in the Bottom Left corner and lower risk towards the Top Right corner.

Data Altered Prediction Matrix

Exit from slide review mode and click excel attachment below

Data altered
diction Matric Exar

Internal Cross validation of the Model for Deaths Before 180 Days

- Internal validation of the model was done using the concordance index and internal bootstrap re-sampling methods. R rms package.

- Concordance probability (Harrell's C-Index) is the probability that a randomly selected pair of patients, one with a poorer survival outcome than the other, will be correctly differentially identified based on inputting the two patient's baseline prognostic characteristics in the fitted model.

- For the logistic model the concordance probability is identical to the area under the receiver operating characteristic (ROC) curve for the model.

Harrell Frank E (2001). Regression Modeling Strategies: with applications to linear models, logistic regression and Survival analysis.

Validation Steps – Logistic Analysis

- Uses R package “rms”.
- Import each of the 10 imputed datasets into R and run the following R code for each dataset
  - library("rms")
  - ## Imputatiom # 1
  - f <- lrm(dthbf180 ~ agen+iss+ecog+platcount+mobility+ VTE, data = impt1log, x=TRUE, y=TRUE)
  - validate(f,B=100, dxy = TRUE)
- Use Concordance probability, C-Index =0.5*|Dxy|+0.5 where Dxy is the Somer’s D statistic output by R.
- Find the mean of the Concordance probabilities for test and training over the 10 imputations.
- Obtain the percent reduction in Concordance Probability as

  100*{(C Avg for training) – (C Avg for Test)}/(C Avg for training)
- Calculate the bootstrap adjusted C or AUC as C Avg for C-index corrected.
- Calculate the 95% CI for this AUC using expressions from the reference below.

Hanley and McNeil (1982). Radiology Vol 143, No 1, Pg 29-36.

Validation Steps – Logistic Analysis

Confidence Interval for AUC

Let AUC denote the sample AUC value. For large samples, the distribution of AUC is approximately normal. Hence, a100(1 – α)% confidence interval for AUC may be computed using the standard normal distribution as follows

$$AUC \pm z_{\alpha/2} SE(AUC)$$

The width of the confidence interval is $2z_{\alpha/2} SE(AUC)$. One-sided limits may be obtained by replacing $\alpha/2$ by α.

The formula for SE(AUC) was given by Hanley and McNeil (1982) is

$$SE(AUC) = \sqrt{\frac{AUC(1-AUC) + (N_1 - 1)(Q_1 - AUC^2) + (N_2 - 1)(Q_2 - AUC^2)}{N_1 N_2}}$$

where

$$Q_1 = \frac{AUC}{2 - AUC}$$

$$Q_2 = \frac{2AUC^2}{1 + AUC}$$

N1 is the # of deaths in 180 days and N2 is the number alive at 180 days.

Results of Internal Validation

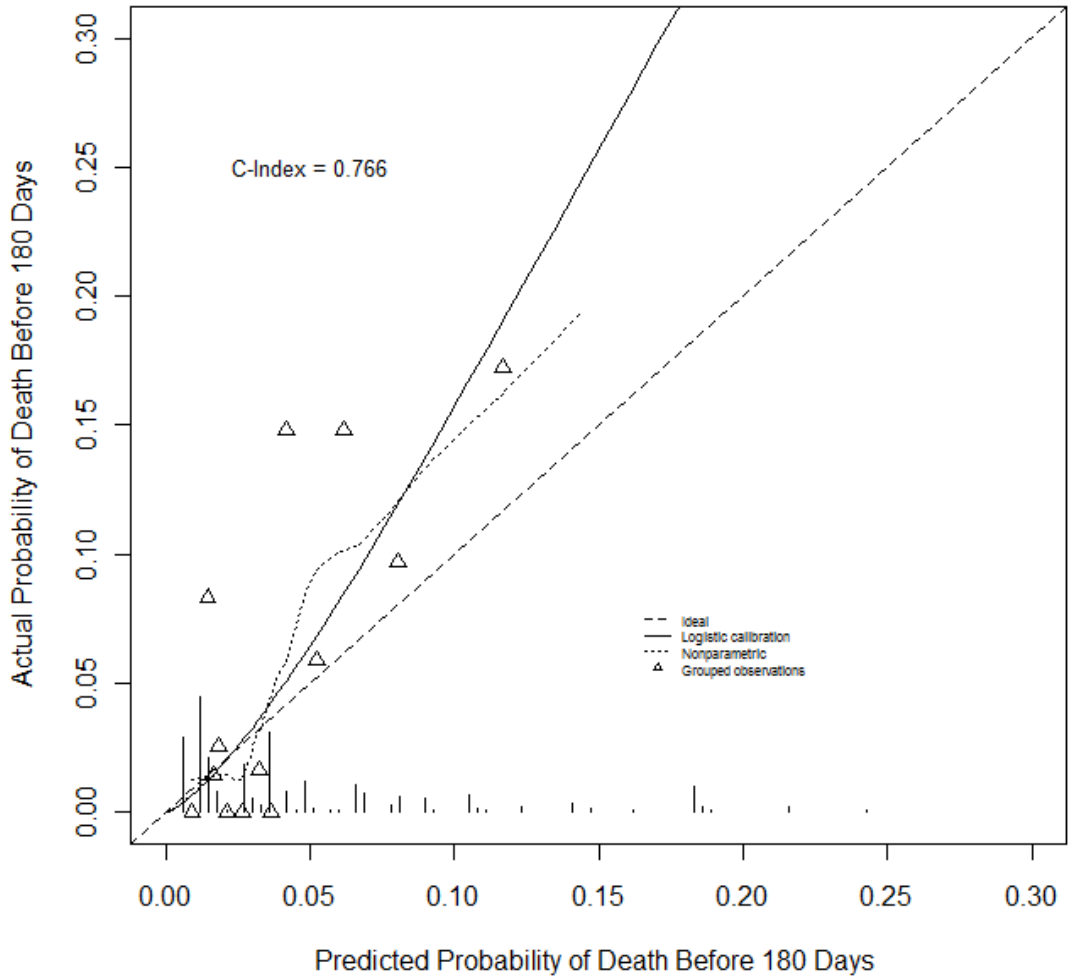
- The percent reduction in the concordance probability in the test bootstrap re-sampling estimate compared to the training bootstrap estimate was 2.53%.
- The training optimism adjusted concordance probability of the fitted logistic model was estimated as 73.00% with a 95% confidence interval of (67.29%, 78.69%). A concordance probability significantly greater than 50% is indicative of a good predictive model.

R External Validation Code – Logistic regression

- Getting Concordance Probability:
- From the combined miananalyze coefficient estimates of the multiple imputation analysis derive predictions in the test dataset as follows
  - logist <- read.csv("Z:/MedicalAffairs-MM/Connect-MM/Stats/Albert/Death within 180 Days/Validation/CC 4013 015/d_mm_015logit.csv", header =T)
  - library(rms)
  - phat <- 1/(1+exp(-(-.68+(0.64*logist$mobf+0.70*logist$ecog+ 0.49*logist$VTE+ 0.50*logist$issstage+0.49*logist$agen+0.81*logist$plat_ct))))
  - val.prob(phat, logist$dthbf180, xlab="Predicted Probability of Death Before 180 Days ", ylab=" Actual Probability of Death Before 180 Days ",lim=c(0,.3), m= 30, cex = 0.5)
  - text(0.05,0.25, "C-Index = 0.766", cex = 0.8)
- The C-Index (along with a host of other statistics) is provided when you run the analysis without “lim =c(0,0.3)” above.

Note: Parameter estimates changed to mask data.

Validation Using External Study: Logistic Regression



ANALYSIS OF THE MM DISEASE REGISTRY BASELINE CONTINUOUS QOL ENDPOINTS

General Approach to Multivariate Analyses: QOL Endpoints.

Univariate Regression Analysis to determine which variables to enter into Multivariate.

10 datasets created using multiple imputation.

Datasets stacked and variable selection was done using weighted SSE .

10 Unstacked multivariate regression analyses done for the variables selected.

Inferences combined using Rubin's method to obtain estimates and p-values.

Procedure : Model Selection for Continuous Case

Determine the best 1 variable model to best p variable model.

Compute scaled deviance = SSE/(sigma squared) for the best 1 variable to the best p variable model.

The scaled deviance = -2*LOG(Likelihood) is a chi-squared score statistic.

Difference in scaled deviance is chi-squared with d.f. given by the difference in the number of variables in the Model.

Starting with best 1 variable model, move in 1 variable increments to the best k variable model. Use 1 d.f chi-square critical value.

McCullagh and Nelder (1989). Generalized Linear Models, Chapman & Hall/CRC

SAS Analysis Details :Multivariate Model Selection

- **SAS PROC REG was run against the stacked dataset with option *selection = maxr.***
- **Analysis weighted by mean percent of non-missing observations divided by the number of imputations.**
- **SSE recorded for each best model**

SigmaSqrd=(Saturated SSE)/(Nobs-Nparm-1)

SAS Analysis Details: Model Selection cont'd

- **Predictor selected in the best 1-variable model was included by default**

- **For 2-variable model and up,**
  - **the scaled deviance *=SSE/SigmaSqrd***

- **The difference between incremental models**
  - $DEVdif_i = DEV_i - DEV_{i-1}$

SAS Analysis Details: Model Selection Conclusion

- **Difference in scaled deviance between two models differing by one variable has a chi-squared distribution with 1 d.f.**
- **Continue adding variables while**
  - ***{abs(DEVdif$_i$)>2.71}***
- **For list of variables identified,**
  - **SAS PROC GLM was run by imputation**
  - **Combined using SAS PROC MIANALYZE.**

EQ-5D final results after MIANALIZE: Imputed vs. Un-imputed Complete Case Analysis

Variable	Estimate Imputed model	P-val Imputed model	Estimate Un-imputed model	P-val Un-imputed model
ECOG	-0.08	<0.001	-0.11	<0.001
BONEINV	-0.06	<0.001	-0.08	<0.001
HISPAN	-0.08	<0.001	-0.11	<0.001
DSSTAGE	-0.01	0.067	-0.02	0.058
AGEGRPN	0.02	0.005		
ISS3	-0.02	0.117		
SEX	0.02	0.045	0.03	0.033
CALCIUM	-0.04	0.08		

Conclusions

- With Registries missing data analysis can be tied into multivariate analysis efficiently using a few SAS proc calls.

- Features
  - A weighted stacked analysis can be used for variable selection.
  - For the exponential family including the normal continuous case the deviance statistic (-2*LOG(Likelihood) can be used for variable selection as the difference in deviance is a chi-squared statistic.
  - For logistic regression the SAS score statistic can be used.
  - Ordinary regression one needs the scaled deviance obtained as the (stacked weighted SSE)/(Saturated Sigma Square).
  - Model can be presented in user friendly manner.
  - R package rms can be used for internal and external validation for logistic, regression and survival cases.