

Validation Methods and ROC Curve

Marina Yogendran
Winnipeg SAS Users Group meeting
April 29th 2010



Validation Methods

When you have observational or survey data for a cohort, but would like to get more information from a different source (a secondary data) that has more information that you are interested in and would like to use this data for future analysis you might want to validate the secondary data using your primary data first - this process is called **Validation**.

There are several validation measures you could use to validate your data. They are

- Kappa statistic
- Sensitivity
- Specificity
- Negative Predictive Value (NPV)
- Positive Predictive Value (PPV)
- Youden's index

Kappa Statistic(κ) is a measure of agreement between two sources, which is measured on a binary scale (i.e. condition present/absent).

κ statistic can take values between 0 and 1.

- **Poor agreement : $\kappa < 0.20$**
- **Fair agreement : $\kappa = 0.20$ to 0.39**
- **Moderate agreement : $\kappa = 0.40$ to 0.59**
- **Good agreement : $\kappa = 0.60$ to 0.79**
- **Very good agreement : $\kappa = 0.80$ to 1.00**

Calculation of Validation Indices

Secondary Data
e.g. Administrative Data

Primary Data
(e.g. CCHS)
Gold standard

	Condition (False)	Condition (True)
Condition (False)	A (TN)	B (FP)
Condition (True)	C (FN)	D (TP)

$$\text{Sensitivity} = D/(C+D)$$

$$\text{Specificity} = A/(A+B)$$

$$\text{PPV} = D/(B+D)$$

$$\text{NPV} = A/(A+C)$$

Youden's(1950) Index combines information on sensitivity and specificity.

Youden's Index

$$= \text{sensitivity} + \text{specificity} - 1$$

Where sensitivity and specificity are calculated as proportions.

Youden's Index ranges between -1 and +1 (+1 being the optimal value)

e.g. ICU admissions during hospitalization.

Gold standard: ICU database

Test data: Hospital abstract ICU service code

code

ccsfg (ICU database CCS indicator) <u>Gold standard</u>	Frequency		scufg (Hospital abstract CCS indicator)		
	Percent Row Pct Col Pct		0	1	Total
0		285794	2156	287950	
		92.11	0.69	92.81	
	Specificity	99.25	0.75		
	PPV	99.74	9.08		
1		738	21578	22316	
		0.24	6.95	7.19	
		3.31	96.69		Sensitivity
		0.26	90.92		NPV
Total	286532	23734	310266		
	92.35	7.65	100.00		

```
proc freq data=test;  
tables ccsfg * scufg /  
kappa;  
run;
```

Instead of the option **kappa**
you could use **agree** and you
would get the same results

Statistics for Table of ccsfg by scufg

McNemar's Test

Statistic (S)	694.7906
DF	1
Pr > S	<.0001

Simple Kappa Coefficient

Kappa	0.9321
ASE	0.0013
95% Lower Conf Limit	0.9297
95% Upper Conf Limit	0.9346

Sample Size = 310266

Compute Confidence intervals

```
proc freq data=test;  
  tables ccsfg * scufg /list noprint out=datval  
  (drop=percent);  
run;
```

For Sensitivity:

```
ods output binomialprop=sens;  
proc freq data=datval order=freq;  
  where ccsfg=1;  
  weight count;  
  tables scufg / noprint;  
  exact binomial;  
  title 'Sensitivity:';  
run;  
ods output close;
```

Binomial Proportion for scufg = 1

Proportion (P)	0.9669
ASE	0.0012
95% Lower Conf Limit	0.9646
95% Upper Conf Limit	0.9693
Exact Conf Limits	
95% Lower Conf Limit	0.9645
95% Upper Conf Limit	0.9692

Wrote a macro to do the calculations and here is the output

validation	Label	value
Sensitivity	Proportion (P)	96.69
Sensitivity	95% Lower Conf Limit	96.46
Sensitivity	95% Upper Conf Limit	96.93
Specificity	Proportion (P)	99.25
Specificity	95% Lower Conf Limit	99.22
Specificity	95% Upper Conf Limit	99.28
PPV	Proportion (P)	90.92
PPV	95% Lower Conf Limit	90.55
PPV	95% Upper Conf Limit	91.28
NPV	Proportion (P)	99.74
NPV	95% Lower Conf Limit	99.72
NPV	95% Upper Conf Limit	99.76

```
ods trace on / listing;
proc logistic descending data=test;
  model ccsfg = scufg / ctable pprob=0.5;
run;
ods trace off;
```

Name: Classification

Label: Classification Table

Classification Table									
	Correct		Incorrect		Percentages				
Prob	Non-	Non-	Non-	Non-	Sensi-	Speci-	False	False	
Level	Event	Event	Event	Event	Correct	tivity	ficity	POS	NEG
0.500	21578	286E3	2156	738	99.1	96.7	99.3	9.1	0.3

```
proc logistic descending data=test;
  model ccsfg = scufg / ctable pprob = (0.5 to 1.0 by 1.0) ;
run;
```

Classification Table

Prob Level	Correct		Incorrect		Correct	Percentages			
	Event	Non- Event	Event	Non- Event		Sensi- tivity	Speci- ficity	FALSE POS	FALSE NEG
0.5	21578	2.86E+05	2156	738	99.1	96.7	99.3	9.1	0.3
0.6	21578	2.86E+05	2156	738	99.1	96.7	99.3	9.1	0.3
0.7	21578	2.86E+05	2156	738	99.1	96.7	99.3	9.1	0.3
0.8	21578	2.86E+05	2156	738	99.1	96.7	99.3	9.1	0.3
0.9	21578	2.86E+05	2156	738	99.1	96.7	99.3	9.1	0.3
1	0	2.88E+05	0	22316	92.8	0	100.		7.2

ROC (Receiver Operating Characteristics)

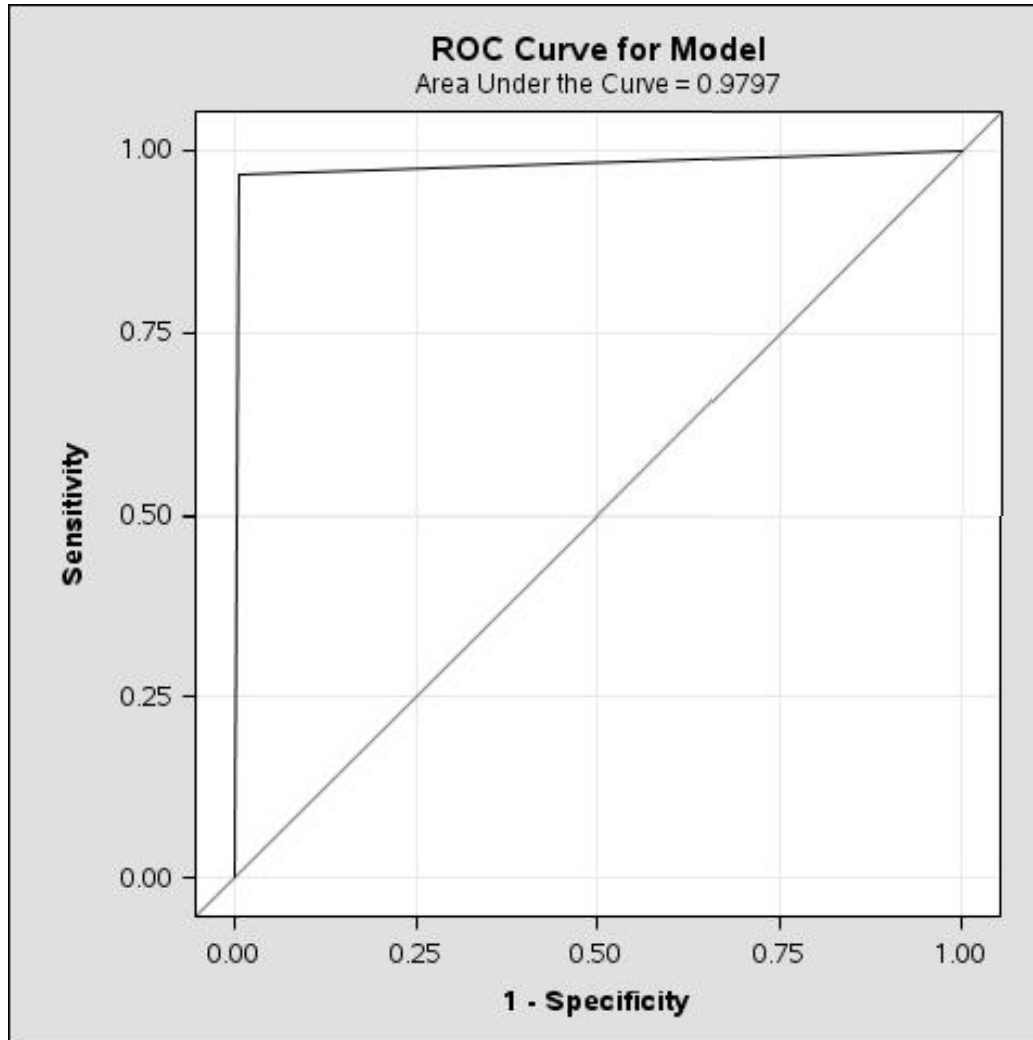
Curves:

ROC curve is a graphical plot of Sensitivity against (1 – specificity).

i.e. a plot of proportion of true positives (events predicted to be events) versus the proportion of false positives (nonevents predicted to be events)

In SAS you can use `outroc=<filename>` to get the values for each predictive probability to plot the ROC curve or in SAS 9.2 you could use an option in the proc statement, `plots=roc` to get the ROC curve.

ROC curve



```
ods graphics on;  
proc logistic descending  
data=test plots=roc;  
  model ccsfg = scufg /  
  outroc=rocdata ;  
run;  
ods graphics off;
```

Proc logistics also gives you association statistics, which contains the area under the ROC curve, i.e. the c-statistic

Name: Association
Label: Association Statistics

Association of Predicted Probabilities and Observed Responses

Percent Concordant	96.0	Somers' D	0.959
Percent Discordant	0.0	Gamma	0.999
Percent Tied	4.0	Tau-a	0.128
Pairs	6425892200	c	0.980

References

- Receiver Operating Characteristic (ROC) Curves by Mithat Gonen, Memorial Sloan-Kettering Cancer Center, Paper 21-31, Statistics and Data Analysis
- Some Issues in Using Proc LOGISTIC for Binary Logistic Regression by David D. Schlotzhauer