



The Logic and Logistics of Logistic Regression Including New Features in SAS® 9.2

Lawrence Rasouliyan, Dave P. Miller

ICON Late Phase & Outcomes Research

PhUSE 2012/Health Outcomes Research & Epidemiology/16 Oct 2012

- Logistic regression is an important method of multivariable analysis in observational studies.
- Methods apply when the outcome is dichotomous (i.e., yes/no, on/off, 1/0, diseased/not-diseased)
- Unless study is randomized, only limited conclusions can be drawn at the univariable level.
- Need to assess the association between exposure and outcome while controlling for confounders.
- Ultimate goal: To establish a model where you can predict the odds of experiencing the outcome among patients for any given set of underlying characteristics.

The Logic – Odds Ratio (OR)

- The measure of association you want in logistic regression.

$$OR = \frac{[Odds\ of\ the\ outcome\ occurring\ in\ Group\ A]}{[Odds\ of\ the\ outcome\ occurring\ in\ Group\ B]}$$

Associated Risk: (Decreased) (None) (Increased)

Odds Ratio: 0 1 ∞

Associated Risk: (Decreased) (None) (Increased)

Ln (Odds Ratio): $-\infty$ 0 ∞

- Unlike linear regression, the value of the outcome itself (0 or 1) is not regressed upon as the dependent variable...why not?
- Model the log odds of the outcome occurring (also called logit) as a linear function of the independent variables:

$$\ln \left[\frac{P(\text{Outcome})}{1-P(\text{Outcome})} \right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots$$

- Odds Ratios are then calculated by exponentiating different combinations of the β s.

- TENOR = The Epidemiology and Natural History of Asthma: Outcomes and Treatment Regimens
- Prospective, observational, multi-center, large scale study of patients with severe or difficult-to-treat asthma
- Outcome of interest (we've cleverly disguised): Poor Lung Function Outcome (PLFO)
 - **PLFO = 1** if patient experiences this outcome (N=612)
 - **PLFO = 0** if the patient does not experience the outcome (N=405)
- Objective of analysis: To determine which factors are associated with experiencing PLFO.

Significant Variables from Univariable Analysis - 1

/* Age per decade	*/ agebl_dec
/* Gender	*/ sexmf
/* Black race	*/ raceblack
/* Hispanic ethnicity	*/ racehisp
/* IgE >30, IgE >100	*/ ige>30 ige>100
/* Present smoker, Past smoker	*/ presentsmk pastsmk
/* Skn test done, Skn test pos.	*/ skntstdone skntstpos
/* Hx of allergic rhinitis	*/ hxar
/* Diabetic	*/ idiabet
/* Fam hx of allergy	*/ ihxallergy_fam
/* Fam hx of al. rhinitis	*/ ihxar_fam
/* Fam hx of atopic dermatitis	*/ ihxad_fam
/* Moisture at home	*/ imoist

Significant Variables from Univariable Analysis - 2

/* Mold at home	*/	imold
/* Has at least 1 pet	*/	ipet
/* Has a cat	*/	icat
/* Has a dog	*/	idog
/* Has multiple pets	*/	imultpet
/* Emotional stress triggers sx*/		sxemyn
/* Sinus infection triggers sx	*/	sxsiyn
/* Dust triggers sx	*/	sxduyn
/* Aspirin triggers sx	*/	sxasyn
/* Asthma duration per decade	*/	amdurbl_dec
/* College grad or higher	*/	collgrad
/* Unemployed	*/	unemployed
/* Medicaid insurance	*/	medicaid

Logistic Model – All Significant Univariable Terms

* Logistic model - all significant univariable terms;

```
proc logistic data=saslib.sad_PLFO;  
  where PLFO in (0,1);  
  class sexmf (ref='1') hxr (ref='0') / param=ref;  
  model PLFO (event='1') = agebl_dec sexmf raceblack  
    racehisp igegt30 igegt100 presentsmk pastsmk  
    skntstdone skntstpos hxr idiabet ihxallergy_fam  
    ihxr_fam ihxad_fam imoist imold ipet icat idog  
    imultipet sxemyn sxsiyn sxduyn sxasyn amdurbl_dec  
    collgrad unemployed medicaid;  
run;
```

Model Output – Response Profile/Class Info

Response Profile

Ordered Value	PLFO	Total Frequency
1	0	393
2	1	598

Probability modeled is PLFO=1.

NOTE: 26 observations were deleted due to missing values for the response or explanatory variables.

Class Level Information

Class	Value	Design Variables	
sexmf	1	0	
	2	1	
hxr	0	0	0
	1	1	0
	2	0	1

Model Output – Maximum Likelihood Estimates (Abbreviated)

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.5256	0.6217	6.0213	0.0141
agebl_dec		1	0.3716	0.0657	32.0322	<.0001
sexmf	2	1	0.9011	0.1811	24.7661	<.0001
raceblack		1	0.6612	0.3026	4.7733	0.0289
racehisp		1	-1.0497	0.3767	7.7647	0.0053
igegt30		1	0.1306	0.2156	0.3671	0.5446
igegt100		1	0.1455	0.1913	0.5781	0.4471
presentsmk		1	1.3993	0.4224	10.9734	0.0009
pastsmk		1	0.4757	0.1767	7.2427	0.0071
skntstdone		1	0.2927	0.4264	0.4713	0.4924
skntstpos		1	-0.3947	0.3744	1.1112	0.2918
hxr	1	1	-0.5996	0.3921	2.3383	0.1262
hxr	2	1	-0.6863	0.3742	3.3626	0.0667
idiabet		1	0.1964	0.3336	0.3467	0.5560
...						

Model Output – Odds Ratio Estimates (Abbreviated)

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits	
agebl_dec		1.450	1.275	1.649
sexmf	2 vs 1	2.462	1.727	3.511
raceblack		1.937	1.070	3.506
racehisp		0.350	0.167	0.732
igegt30		1.140	0.747	1.739
igegt100		1.157	0.795	1.683
presentsmk		4.052	1.771	9.274
pastsmk		1.609	1.138	2.275
skntstdone		1.340	0.581	3.091
skntstpos		0.674	0.323	1.404
hxr	1 vs 0	0.549	0.255	1.184
hxr	2 vs 0	0.503	0.242	1.048
idiabet		1.217	0.633	2.340
...				

- As much of an art as it is a science
- Should take into account several things:
 - Statistical significance
 - Clinical plausibility
 - Collinearity
 - Effect modification
- Several statistically-driven algorithms exist:
 - **Forward Selection** – Adding variables one at a time and assessing the significance of each term
 - **Backward Elimination** – Starting with all variables in the model and removing them one at a time
 - **Stepwise Selection/Elimination** – Allowing variables to be added or removed upon successive revisions

* Stepwise Logistic Regression;

```
proc logistic data=saslib.sad_PLFO;  
  where PLFO in (0,1);  
  class sexmf (ref='1') hxr (ref='0') / param=ref;  
  model PLFO (event='1') = agebl_dec sexmf raceblack  
    racehisp igegt30 igegt100 presentsmk pastsmk  
    skntstdone skntstpos hxr idiabet ihxallergy_fam  
    ihxr_fam ihxad_fam imoist imold ipet icat idog  
    imultpet sxemyn sxsiyn sxduyn sxasyn amdurbl_dec  
    collgrad unemployed medicaid /  
  selection=stepwise slentry=0.05 slstay=0.05  
  details;  
run;
```

Summary of Stepwise Selection

Summary of Stepwise Selection

Step	Entered	Effect		DF	Number		Score		Wald	
		Removed			In		Chi-Square		Chi-Square	Pr > ChiSq
1	amdurbl_dec			1	1		92.9297			<.0001
2	agebl_dec			1	2		57.0683			<.0001
3	sexmf			1	3		37.8539			<.0001
4	raceblack			1	4		17.8072			<.0001
5	presentsmk			1	5		10.8102			0.0010
6	pastsmk			1	6		8.5910			0.0034
7	ihxad_fam			1	7		7.7448			0.0054
8	sxdun			1	8		6.5230			0.0106
9	collgrad			1	9		7.0821			0.0078
10	ipet			1	10		5.4084			0.0200
11	racehisp			1	11		5.8737			0.0154

- For this particular asthma outcome, investigators thought that the disease might manifest itself differently in males and females
- Effect modification was assessed by refitting the model multiple times with the main effects terms and a single interaction term between gender and one of the main effects
- The only interaction that attained significance was that between gender and duration of asthma (per decade)

```
* Model with gender*duration of asthma interaction;

proc logistic data=saslib.sad_PLFO;
  where PLFO in (0,1);
  class sexmf (ref='1') / param=ref;
  model PLFO (event='1') = agebl_dec sexmf raceblack
    racehisp presentsmk pastsmk ihxad_fam ipet sxduyn
    amdurbl_dec collgrad sexmf*amdurbl_dec;
run;
```

Output – Maximum Likelihood Estimates: Main Effects w/ Interaction

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.2035	0.4056	29.5126	<.0001
agebl_dec		1	0.3608	0.0603	35.7844	<.0001
sexmf	2	1	1.4646	0.3267	20.0966	<.0001
raceblack		1	0.8101	0.2821	8.2434	0.0041
racehisp		1	-0.8205	0.3555	5.3284	0.0210
presentsmk		1	1.3501	0.3957	11.6429	0.0006
pastsmk		1	0.4827	0.1710	7.9643	0.0048
ihxad_fam		1	-0.5862	0.1936	9.1690	0.0025
ipet		1	-0.3841	0.1530	6.3020	0.0121
sxduyn		1	-0.4603	0.1907	5.8285	0.0158
amdurbl_dec		1	0.4661	0.0622	56.1684	<.0001
collgrad		1	-0.3761	0.1568	5.7548	0.0164
amdurbl_dec*sexmf	2	1	-0.2194	0.1119	3.8461	0.0499

- Investigators wanted to consider a few additional variables which were marginally insignificant in the main effects but had clinical relevance:
 - Hx of allergic rhinitis
 - Family hx of atopic dermatitis
 - Sinus infection triggers
 - Aspirin sensitivity
- Variables were individually added then removed from the main effects plus interactions model to assess significance.
- Aspirin sensitivity (sxasyn) was the only variable that attained significance – permanently added to the model.

Code for Final Logistic Model



Late Phase & Outcomes Research

[illegible]

Final Logistic Model – Odds Ratios

Odds Ratio Estimates

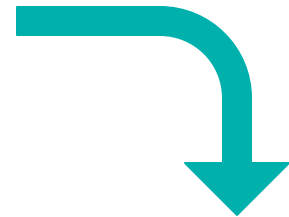
Effect	Point Estimate	95% Wald Confidence Limits	
agebl_dec	1.437	1.277	1.618
raceblack	2.199	1.265	3.822
racehisp	0.443	0.220	0.893
presentsmk	3.936	1.808	8.569
pastsmk	1.640	1.172	2.294
ihxad_fam	0.563	0.385	0.824
ipet	0.694	0.514	0.939
sxduyn	0.626	0.430	0.910
sxasyn	1.538	1.003	2.360
collgrad	0.699	0.514	0.952

Where's gender? Where's asthma duration?

ODDSRATIO Statement – **New to SAS 9.2**

* Final logistic regression model with ODDSRATIO statements;

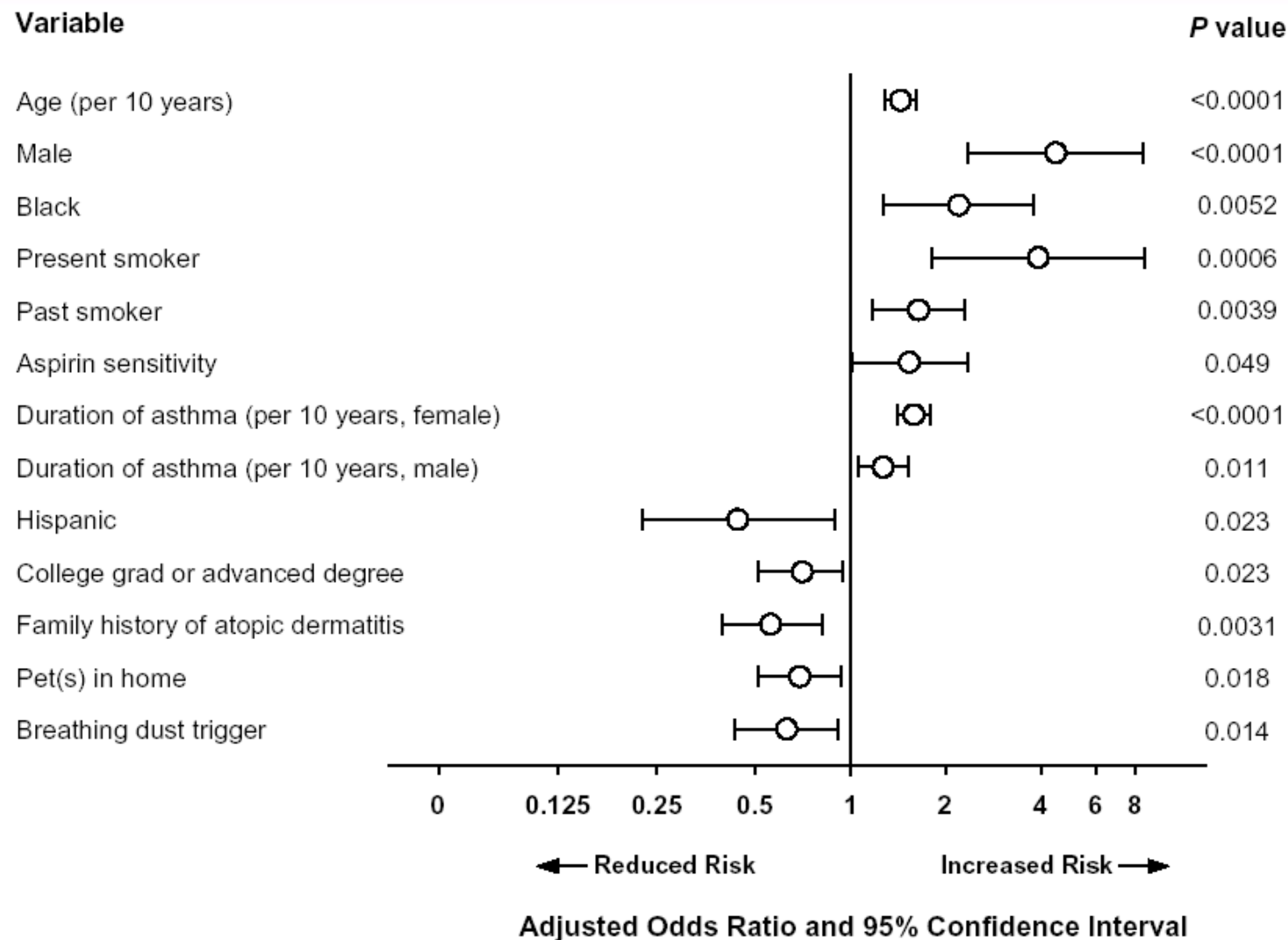
```
proc logistic data=saslib.sad_PLFO;
  where PLFO in (0,1);
  class sexmf (ref='1') / param=ref;
  model PLFO (event='1') = agebl_dec sexmf raceblack racehisp
    presentsmk pastsmk ihxad_fam ipet sxduyn
    amdurbl_dec collgrad sexmf*amdurbl_dec sxasyn;
  oddsratio "Asthma Duration by Gender" amdurbl_dec;
  oddsratio "M vs F at Asthma Dur=20y" sexmf /
    at (amdurbl_dec=2.0) diff=ref;
run;
```



Wald Confidence Interval for Odds Ratios

	Label	Estimate	95% Confidence Limits	
Asthma Duration by Gender	amdurbl_dec at sexmf=1	1.588	1.406	1.793
Asthma Duration by Gender 2	amdurbl_dec at sexmf=2	1.273	1.057	1.533
M vs F at Asthma Dur=20y	sexmf 1 vs 2 at amdurbl_dec=2.0	2.867	2.024	4.059

Interpreting the Final Model



In case you're interested...

CHEST[®]

Official publication of the American College of Chest Physicians



Risk Factors Associated With Persistent Airflow Limitation in Severe or Difficult-to-Treat Asthma: Insights From the TENOR Study

June H. Lee, Tahirah Haselkorn, Larry Borish, Lawrence Rasouliyan, Bradley E. Chipps and Sally E. Wenzel

Chest 2007;132:1882-1889
DOI 10.1378/ichest.07-0713

The online version of this article, along with updated information and services can be found online on the World Wide Web at:
<http://chestjournal.org/cgi/content/abstract/132/6/1882>

PLFO = PAFL = Persistent Airflow Limitation

Other topics covered in the paper

Late Phase & Outcomes Research

- Hosmer-Lemeshow Goodness of Fit test – Assessing how well the model fits the data
- Receiver-Operating Characteristic (ROC) analysis – Using the model as a tool to “diagnose” the outcome

Lawrence Rasouliyan, MPH
Senior Research Manager, Statistical Analysis
ICON Late Phase & Outcomes Research (ILPOR)
Email: Lawrence.Rasouliyan@iconplc.com



A Symbol of Excellence

Back up slides (time permitting)

Hosmer-Lemeshow Goodness of Fit

Late Phase & Outcomes Research

- Specified by the “lackfit” option at the end of the model statement
- Ranks observations by predicted value into deciles and stratifies by true outcome status.
- Evaluates the observed and expected number of observations in each decile and calculated an overall chi-square statistic with 8 degrees of freedom

Partition for the Hosmer and Lemeshow Test

Group	Total	PLFO = 1		PLFO = 0	
		Observed	Expected	Observed	Expected
1	101	12	16.93	89	84.07
2	101	30	29.27	71	71.73
3	101	32	40.58	69	60.42
4	101	61	50.40	40	50.60
5	101	62	59.38	39	41.62
6	101	69	67.77	32	33.23
7	101	81	75.40	20	25.60
8	101	81	82.73	20	18.27
9	101	88	89.07	13	11.93
10	99	89	93.47	10	5.53

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
15.3499	8	0.0527

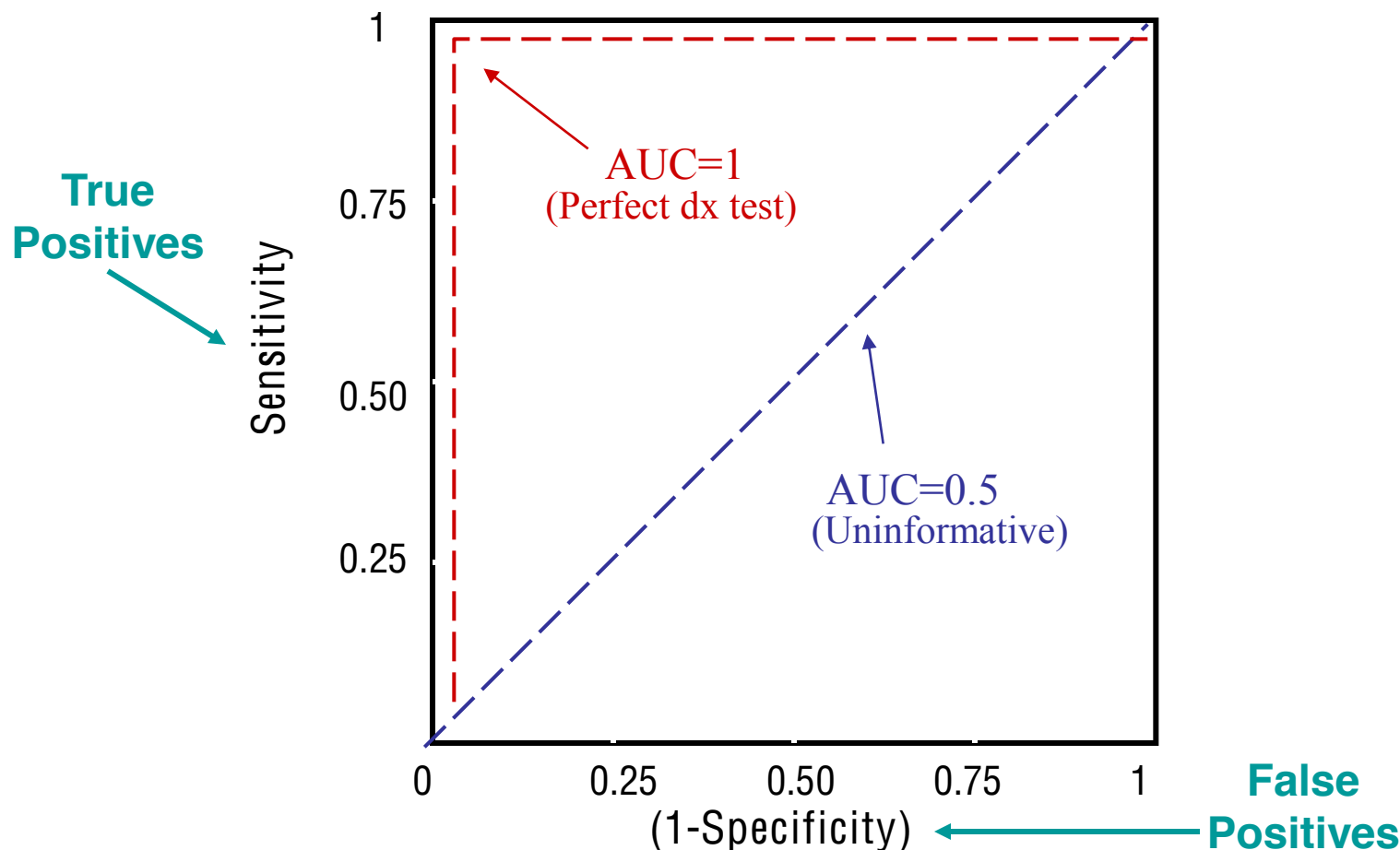
Not significant – Fail to reject goodness of fit.

Very close though, some reason for concern...

- ***Continuous measure*** is used as a “diagnostic” tool in predicting a ***dichotomous outcome***
- Examples:
 - Chance of rain (*continuous*) to predict whether rain will actually occur (*dichotomous*)
 - Credit score (*continuous*) to predict whether card holder will have a delinquent payment (*dichotomous*)
- In our analysis, we want to evaluate how well our final model “diagnoses” PLFO
 - Calculate predicted possibility of PLFO for each patient (*continuous measure*) as specified by the model
 - Look to see how predictive this is in detecting whether the patient experienced PLFO (*dichotomous outcome*)

Receiver-Operating Characteristic (ROC) Analysis

- For each cutoff value, you can calculate the sensitivity and specificity (and 1-specificity):



ROC and ROCCONTRAST Statements

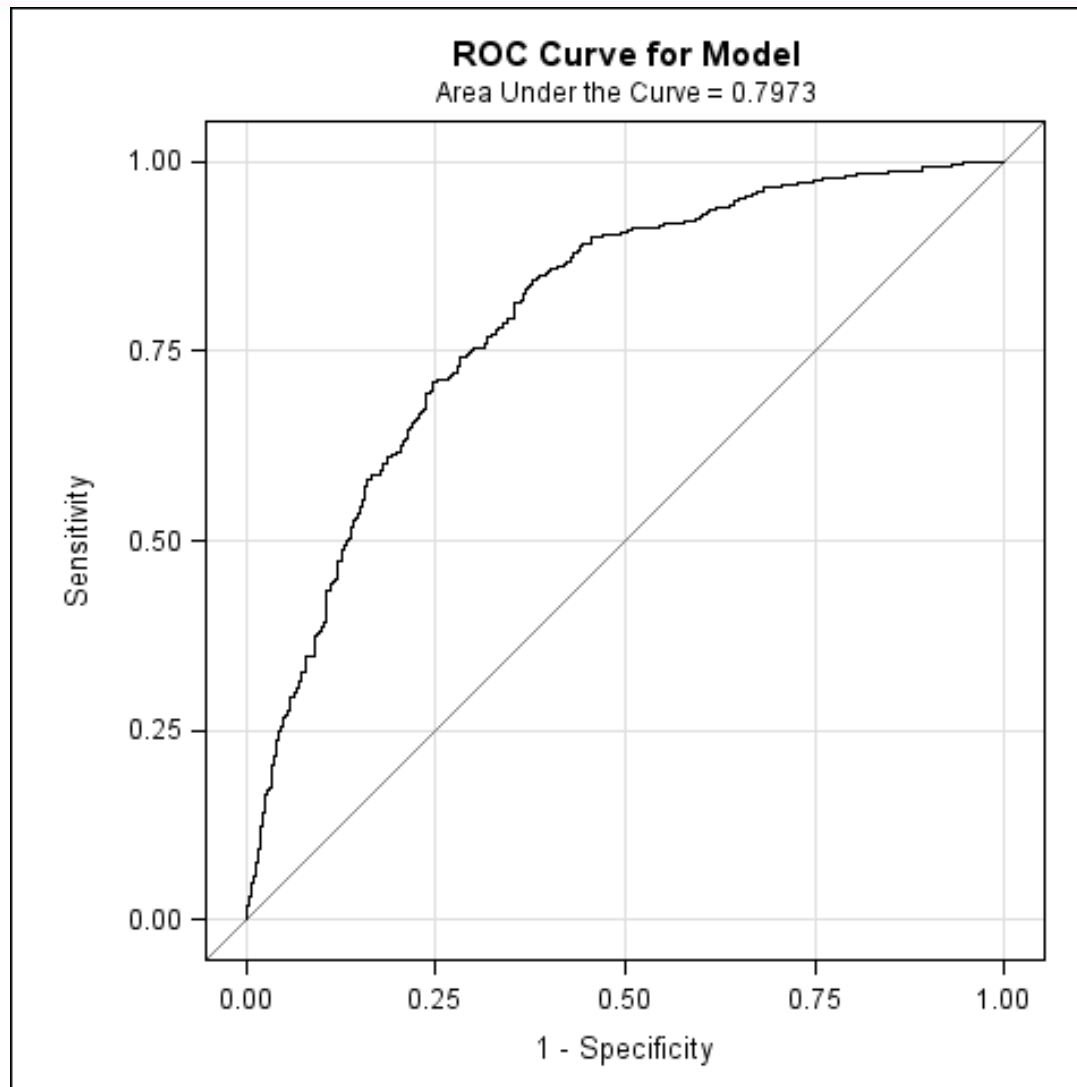
New to SAS 9.2

* Generating ROC data and comparing AUC values;

```
ods graphics on;
proc logistic data=saslib.sad_PLFO plots=roc;
  where PLFO in (0,1);
  class sexmf (ref='1') / param=ref;
  model PLFO (event='1') = agebl_dec sexmf raceblack racehisp
    presentsmk pastsmk ihxad_fam ipet sxduyn
    amdurbl_dec collgrad sexmf*amdurbl_dec sxasyn;
  roc "Diagonal";
  roccontrast reference("Diagonal") / estimate e;

run;
ods graphics off;
```

ROC Plot from ODS Graphics



ROC and ROCCONTRAST Output

ROC Association Statistics

----- Mann-Whitney -----

ROC	Area	Standard	95% Wald	
		Error	Confidence	Limits
Model	0.7973	0.0145	0.7689	0.8256
Diagonal	0.5000	0	0.5000	0.5000

ROC Contrast Test Results

Contrast	DF	Chi-Square	Pr > ChiSq
Reference = Diagonal	1	421.8615	<.0001

ROC Contrast Rows Estimation and Testing Results

Contrast	Estimate	Standard	95% Wald		Chi-Square	Pr >
		Error	Confidence	Limits		ChiSq
Model - Diagonal	0.2973	0.0145	0.2689	0.3256	421.8615	<.0001

- Logistic regression is a powerful method in determining factors associated with a dichotomous outcome.
- PROC LOGISTIC generates a plethora of statistics that can be very helpful tools in constructing models, assessing risk, identifying effect modification, determining goodness of fit, and characterizing diagnostic ability.