Cross-validation and Prediction with Logistic Regression

```
/* mathlogreg3.sas  */
%include 'readmath2.sas'; /* created mathex and mathrep */
title2 'How good is the prediction of passing the course?';
options pagesize=900;

proc logistic descending order=internal data=mathex;
     title3 'Exploratory sample, cutpoint=1/2';
     model passed = hsgpa hscalc precalc / ctable pprob = 0.50;

proc logistic descending order=internal data=mathex;
     title3 'Exploratory sample, many cutpoints';
     model passed = hsgpa hscalc precalc / ctable;

/* Goal here is to make as many correct predictions as possible. Note correct
percentage correctly classified is greatest at cutpoint of 0.58, and overall
proportion correct is 234/375 = 0.624, so set cutpoint at 0.60. */

/* Develop prediction for those with incomplete data. */

data ex2;
     set mathex;
     /* Define indicators for missing information */
     hsutil = hsgpa+hscalc;
     if hsutil = . then hsmiss=1; else hsmiss=0;
     if precalc = . then dtmiss=1; else dtmiss=0;
     mcombo = 10*hsmiss+dtmiss;
     label mcombo = 'HS Missing? DT Missing?';

proc freq;
     tables (hsmiss dtmiss mcombo) * passed  / nocol nopercent chisq missing;

/* Because our cutpoint is 0.60, we predict that anyone with missing
information will not pass. */

/************* Cross-validation ***************/

proc logistic descending order=internal data=mathrep;
     title3 'Replicate 3 tests: Bonferroni alpha = 0.05/3 = 0.0167';
     model passed = hsgpa hscalc precalc;

data rep2;
     set mathrep;
     /* Define indicators for missing information */
     hsutil = hsgpa+hscalc;
     if hsutil = . then hsmiss=1; else hsmiss=0;
     if precalc = . then dtmiss=1; else dtmiss=0;
     mcombo = 10*hsmiss+dtmiss;
     label mcombo = 'HS Missing? DT Missing?';
     if mcombo=0 then anymiss=0; else anymiss=1;
     label anymiss = 'Any Missing Info?';
     format hsmiss dtmiss anymiss ynfmt.;

     /* Predicted probability of Success */
     b0 = -14.7970;  b1 = 0.1173;  b2 =  0.0638;  b3 = 0.2989 ;
     lcombo = b0 + b1*hsgpa + b2*hscalc + b3*precalc  ;
```

```
     probsucc = exp(lcombo)/(1+exp(lcombo));
     label probsucc = 'Estimated probability of success';
     if mcombo = 01 then probsucc = 0.4032;
        else if mcombo = 10 then probsucc = 0.3429;
        else if mcombo = 11 then probsucc = 0.2703;
     cutpoint = 0.60;
     if 0 <= probsucc < cutpoint then predpass = 'No ';
        else if cutpoint <= probsucc <= 1 then predpass = 'Yes';

proc freq;
     title3 'Computation Check';
     tables anymiss*mcombo / norow nocol nopercent missing;

proc sort;
     by probsucc;

proc print;
     var hsgpa hscalc precalc probsucc anymiss predpass passed;

proc freq;
     title3 'Assess Accuracy';
     tables predpass * passed;
     tables anymiss * predpass * passed;
```

The LOGISTIC Procedure

Model Information

Data Set	WORK.MATHEX	
Response Variable	passed	Passed the course
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	579
Number of Observations Used	375

Response Profile

Ordered Value	passed	Total Frequency
1	Yes	234
2	No	141

Probability modeled is passed='Yes'.

NOTE: 204 observations were deleted due to missing values for the response or explanatory variables.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	498.554	375.618
SC	502.481	391.326
-2 Log L	496.554	367.618

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	128.9358	3	<.0001
Score	107.7971	3	<.0001
Wald	79.6583	3	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-14.7970	2.2683	42.5550	<.0001
hsgpa	1	0.1173	0.0310	14.3281	0.0002
hscalc	1	0.0638	0.0132	23.3346	<.0001
precalc	1	0.2989	0.0844	12.5464	0.0004

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
hsgpa	1.124	1.058	1.195
hscalc	1.066	1.039	1.094
precalc	1.348	1.143	1.591

Association of Predicted Probabilities and Observed Responses

Percent Concordant	83.1	Somers' D	0.663
Percent Discordant	16.7	Gamma	0.665
Percent Tied	0.2	Tau-a	0.312
Pairs	32994	c	0.832

Classification Table

Prob Level	Correct Event	Correct Non-Event	Incorrect Event	Incorrect Non-Event	Percentages Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.500	202	93	48	32	78.7	86.3	66.0	19.2	25.6

Pred. Pass	Passed No	Passed Yes		
				Correct = (93+202)/375 = 0.7866667
				Sens = 202/234 = 0.8632479
No	93	32	125	Spec = 93/141 = 0.6595745
Yes	48	202	250	FalsePos = 48/250 = 0.192
				FalseNeg = 32/125 = 0.256
	141	234	375	

"The accuracy of the classification is measured by its sensitivity (the ability to predict an event correctly) and specificity (the ability to predict a nonevent correctly). Sensitivity is the proportion of event responses that were predicted to be events. Specificity is the proportion of nonevent responses that were predicted to be nonevents."

Same output up to this point: `Exploratory sample, many cutpoints`

Classification Table

Prob Level	Correct Event	Correct Non-Event	Incorrect Event	Incorrect Non-Event	Percentages Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.020	234	0	141	0	62.4	100.0	0.0	37.6	.
0.040	234	1	140	0	62.7	100.0	0.7	37.4	0.0
0.060	234	2	139	0	62.9	100.0	1.4	37.3	0.0
0.080	234	3	138	0	63.2	100.0	2.1	37.1	0.0
0.100	234	5	136	0	63.7	100.0	3.5	36.8	0.0
0.120	234	8	133	0	64.5	100.0	5.7	36.2	0.0
0.140	233	13	128	1	65.6	99.6	9.2	35.5	7.1
0.160	232	15	126	2	65.9	99.1	10.6	35.2	11.8
0.180	228	19	122	6	65.9	97.4	13.5	34.9	24.0
0.200	227	23	118	7	66.7	97.0	16.3	34.2	23.3
0.220	226	29	112	8	68.0	96.6	20.6	33.1	21.6
0.240	226	31	110	8	68.5	96.6	22.0	32.7	20.5
0.260	224	38	103	10	69.9	95.7	27.0	31.5	20.8
0.280	221	40	101	13	69.6	94.4	28.4	31.4	24.5
0.300	217	47	94	17	70.4	92.7	33.3	30.2	26.6
0.320	216	53	88	18	71.7	92.3	37.6	28.9	25.4
0.340	215	58	83	19	72.8	91.9	41.1	27.9	24.7
0.360	212	65	76	22	73.9	90.6	46.1	26.4	25.3
0.380	211	72	69	23	75.5	90.2	51.1	24.6	24.2
0.400	210	76	65	24	76.3	89.7	53.9	23.6	24.0
0.420	208	77	64	26	76.0	88.9	54.6	23.5	25.2
0.440	208	82	59	26	77.3	88.9	58.2	22.1	24.1
0.460	207	87	54	27	78.4	88.5	61.7	20.7	23.7
0.480	203	90	51	31	78.1	86.8	63.8	20.1	25.6
0.500	202	93	48	32	78.7	86.3	66.0	19.2	25.6
0.520	198	99	42	36	79.2	84.6	70.2	17.5	26.7
0.540	195	101	40	39	78.9	83.3	71.6	17.0	27.9
0.560	190	104	37	44	78.4	81.2	73.8	16.3	29.7
0.580	188	107	34	46	78.7	80.3	75.9	15.3	30.1
0.600	182	107	34	52	77.1	77.8	75.9	15.7	32.7
0.620	177	109	32	57	76.3	75.6	77.3	15.3	34.3
0.640	175	111	30	59	76.3	74.8	78.7	14.6	34.7
0.660	165	115	26	69	74.7	70.5	81.6	13.6	37.5
0.680	161	116	25	73	73.9	68.8	82.3	13.4	38.6
0.700	156	116	25	78	72.5	66.7	82.3	13.8	40.2
0.720	146	120	21	88	70.9	62.4	85.1	12.6	42.3
0.740	141	124	17	93	70.7	60.3	87.9	10.8	42.9
0.760	134	124	17	100	68.8	57.3	87.9	11.3	44.6
0.780	123	126	15	111	66.4	52.6	89.4	10.9	46.8
0.800	115	126	15	119	64.3	49.1	89.4	11.5	48.6
0.820	106	129	12	128	62.7	45.3	91.5	10.2	49.8
0.840	95	132	9	139	60.5	40.6	93.6	8.7	51.3
0.860	84	133	8	150	57.9	35.9	94.3	8.7	53.0
0.880	74	133	8	160	55.2	31.6	94.3	9.8	54.6
0.900	67	134	7	167	53.6	28.6	95.0	9.5	55.5
0.920	55	136	5	179	50.9	23.5	96.5	8.3	56.8
0.940	40	136	5	194	46.9	17.1	96.5	11.1	58.8
0.960	32	139	2	202	45.6	13.7	98.6	5.9	59.2
0.980	14	140	1	220	41.1	6.0	99.3	6.7	61.1
1.000	0	141	0	234	37.6	0.0	100.0	.	62.4

The FREQ Procedure

Table of hsmiss by passed

hsmiss
passed(Passed the course)

Frequency| Row Pct	No	Yes	Total
0	178	259	437
	40.73	59.27	
1	96	46	142
	67.61	32.39	
Total	274	305	579

Statistics for Table of hsmiss by passed

Statistic	DF	Value	Prob
Chi-Square	1	31.0486	<.0001
Likelihood Ratio Chi-Square	1	31.4291	<.0001
Continuity Adj. Chi-Square	1	29.9799	<.0001
Mantel-Haenszel Chi-Square	1	30.9950	<.0001
Phi Coefficient		-0.2316	
Contingency Coefficient		0.2256	
Cramer's V		-0.2316	

Fisher's Exact Test

Cell (1,1) Frequency (F)	178
Left-sided Pr <= F	1.835E-08
Right-sided Pr >= F	1.0000
Table Probability (P)	1.250E-08
Two-sided Pr <= P	2.670E-08

Sample Size = 579

Table of dtmiss by passed

dtmiss
passed(Passed the course)

Frequency Row Pct	No	Yes	Total
0	210 43.75	270 56.25	480
1	64 64.65	35 35.35	99
Total	274	305	579

Statistics for Table of dtmiss by passed

Statistic	DF	Value	Prob
Chi-Square	1	14.3764	0.0001
Likelihood Ratio Chi-Square	1	14.4799	0.0001
Continuity Adj. Chi-Square	1	13.5504	0.0002
Mantel-Haenszel Chi-Square	1	14.3516	0.0002
Phi Coefficient		-0.1576	
Contingency Coefficient		0.1557	
Cramer's V		-0.1576	

Fisher's Exact Test

Cell (1,1) Frequency (F)	210
Left-sided Pr <= F	1.112E-04
Right-sided Pr >= F	1.0000
Table Probability (P)	6.631E-05
Two-sided Pr <= P	1.619E-04

Sample Size = 579

Table of mcombo by passed

mcombo(HS Missing? DT Missing?)
passed(Passed the course)

Frequency Row Pct	No	Yes	Total
0	141 37.60	234 62.40	375
1	37 59.68	25 40.32	62
10	69 65.71	36 34.29	105
11	27 72.97	10 27.03	37
Total	274	305	579

Statistics for Table of mcombo by passed

Statistic	DF	Value	Prob
Chi-Square	3	42.0295	<.0001
Likelihood Ratio Chi-Square	3	42.6443	<.0001
Mantel-Haenszel Chi-Square	1	33.7370	<.0001
Phi Coefficient		0.2694	
Contingency Coefficient		0.2601	
Cramer's V		0.2694	

Sample Size = 579

The LOGISTIC Procedure

Model Information

Data Set	WORK.MATHREP	
Response Variable	passed	Passed the course
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	579
Number of Observations Used	393

Response Profile

Ordered Value	passed	Total Frequency
1	Yes	229
2	No	164

Probability modeled is passed='Yes'.

NOTE: 186 observations were deleted due to missing values for the response or explanatory variables.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	536.013	439.295
SC	539.987	455.190
-2 Log L	534.013	431.295

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	102.7184	3	<.0001
Score	89.3400	3	<.0001
Wald	72.6624	3	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-13.2803	1.9422	46.7537	<.0001
hsgpa	1	0.1117	0.0276	16.4289	<.0001
hscalc	1	0.0509	0.0129	15.6476	<.0001
precalc	1	0.2285	0.0759	9.0611	0.0026

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
hsgpa	1.118	1.059	1.180
hscalc	1.052	1.026	1.079
precalc	1.257	1.083	1.458

Association of Predicted Probabilities and Observed Responses

Percent Concordant	78.4	Somers' D	0.570
Percent Discordant	21.4	Gamma	0.571
Percent Tied	0.2	Tau-a	0.278
Pairs	37556	c	0.785

The FREQ Procedure

Table of anymiss by mcombo

anymiss(Any Missing Info?)
mcombo(HS Missing? DT Missing?)

Frequency	0	1	10	11	Total
No	393	0	0	0	393
Yes	0	59	105	22	186
Total	393	59	105	22	579

Obs	hsgpa	hscalc	precalc	probsucc	anymiss	predpass	passed
1	68.7	51	2	0.05281	No	No	No
2	73.0	43	3	0.06953	No	No	Yes
3	74.0	50	2	0.08875	No	No	No
4	67.0	57	4	0.10854	No	No	No
5	70.0	54	4	0.12508	No	No	No
6	72.8	54	3	0.12835	No	No	No
7	69.7	56	4	0.13555	No	No	Yes
8	68.8	63	3	0.14056	No	No	No
9	73.8	54	3	0.14205	No	No	No
10	74.3	58	2	0.14388	No	No	No
11	81.2	50	1	0.14389	No	No	No
12	75.8	52	3	0.15560	No	No	No
13	71.2	56	4	0.15751	No	No	No
14	74.5	50	4	0.15808	No	No	No
15	71.8	55	4	0.15839	No	No	No
16	71.5	56	4	0.16224	No	No	No
17	74.5	51	4	0.16676	No	No	Yes
18	77.7	51	3	0.17766	No	No	No
19	75.7	50	4	0.17773	No	No	No
20	73.0	60	3	0.18103	No	No	No
21	73.0	60	3	0.18103	No	No	No
22	76.0	55	3	0.18596	No	No	No
23	71.2	64	3	0.18765	No	No	No
24	76.0	60	2	0.18902	No	No	Yes
25	79.0	50	3	0.19099	No	No	No
26	71.8	54	5	0.19230	No	No	Yes
27	72.3	58	4	0.19463	No	No	Yes
28	76.7	64	1	0.19498	No	No	No
29	69.3	64	4	0.19952	No	No	No

Skipping ...

Obs	hsgpa	hscalc	precalc	probsucc	anymiss	predpass	passed
51	74.8	63	3	0.24846	No	No	Yes
52	76.5	60	3	0.24996	No	No	No
53	77.0	50	5	0.25343	No	No	No
54	74.5	60	4	0.26220	No	No	No
55	71.8	70	3	0.26656	No	No	No
56	.	.	.	0.27030	Yes	No	No
57	.	.	.	0.27030	Yes	No	No
58	.	.	.	0.27030	Yes	No	No
59	.	.	.	0.27030	Yes	No	No
60	.	.	.	0.27030	Yes	No	Yes
61	.	.	.	0.27030	Yes	No	No
62	.	.	.	0.27030	Yes	No	No
63	.	.	.	0.27030	Yes	No	Yes
64	75.2	.	.	0.27030	Yes	No	No
65	.	.	.	0.27030	Yes	No	Yes
66	.	.	.	0.27030	Yes	No	No
67	74.8	.	.	0.27030	Yes	No	No
68	.	.	.	0.27030	Yes	No	Yes
69	.	.	.	0.27030	Yes	No	No
70	.	.	.	0.27030	Yes	No	Yes
71	76.0	.	.	0.27030	Yes	No	No
72	.	.	.	0.27030	Yes	No	No
73	.	.	.	0.27030	Yes	No	No
74	.	.	.	0.27030	Yes	No	No
75	.	83	.	0.27030	Yes	No	No
76	.	.	.	0.27030	Yes	No	No
77	.	.	.	0.27030	Yes	No	No
78	69.5	70	4	0.27229	No	No	No
79	69.2	57	7	0.27870	No	No	Yes
80	72.2	61	5	0.28057	No	No	No
81	75.3	60	4	0.28076	No	No	No
82	73.5	68	3	0.28083	No	No	Yes

Skipping ...

562	92.5	93	5	0.97015	No	Yes	Yes
563	89.2	95	6	0.97127	No	Yes	Yes
564	92.7	83	8	0.97732	No	Yes	Yes
565	91.5	90	7	0.97747	No	Yes	Yes
566	87.8	97	7	0.97775	No	Yes	Yes
567	89.2	90	8	0.97811	No	Yes	Yes
568	91.2	97	6	0.97982	No	Yes	Yes
569	87.8	97	8	0.98340	No	Yes	Yes
570	92.2	94	7	0.98382	No	Yes	Yes
571	91.5	92	8	0.98518	No	Yes	Yes
572	87.0	96	9	0.98556	No	Yes	Yes
573	94.2	96	7	0.98868	No	Yes	Yes
574	95.8	98	6	0.98886	No	Yes	Yes
575	96.7	97	6	0.98931	No	Yes	Yes
576	96.7	98	7	0.99254	No	Yes	Yes
577	93.3	95	9	0.99260	No	Yes	Yes
578	95.3	98	8	0.99347	No	Yes	Yes
579	95.8	96	9	0.99481	No	Yes	Yes

The FREQ Procedure

Table of predpass by passed

predpass
passed(Passed the course)

Frequency / Percent / Row Pct / Col Pct	No	Yes	Total
No	246	114	360
	42.49	19.69	62.18
	68.33	31.67	
	83.39	40.14	
Yes	49	170	219
	8.46	29.36	37.82
	22.37	77.63	
	16.61	59.86	
Total	295	284	579
	50.95	49.05	100.00

Percent correct = 42.49 + 29.36 = 71.85

Sensitivity = 59.86 (Col percent: 170/284)

Specificity = 83.39 (Col percent: 246/295)

False Pos = 22.37 (Row percent: 49/219)

False Neg = 31.67 (Row percent: 114/360)

From the exploratory data, had

Classification Table

	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.600	182	107	34	52	77.1	77.8	75.9	15.7	32.7

Breaking down the results by whether there is missing data,

Table 1 of predpass by passed
Controlling for anymiss=No

predpass
passed(Passed the course)

Frequency / Percent / Row Pct / Col Pct	No	Yes	Total
No	115	59	174
	29.26	15.01	44.27
	66.09	33.91	
	70.12	25.76	
Yes	49	170	219
	12.47	43.26	55.73
	22.37	77.63	
	29.88	74.24	
Total	164	229	393
	41.73	58.27	100.00

Table 2 of predpass by passed
Controlling for anymiss=Yes

predpass
passed(Passed the course)

Frequency / Percent / Row Pct / Col Pct	No	Yes	Total
No	131	55	186
	70.43	29.57	100.00
	70.43	29.57	
	100.00	100.00	
Yes	0	0	0
	0.00	0.00	0.00
	.	.	
	0.00	0.00	
Total	131	55	186
	70.43	29.57	100.00

All data pooled

Percent correct = 42.49 + 29.36 = 71.85
Sensitivity = 59.86 (Col percent: 170/284)
Specificity = 83.39 (Col percent: 246/295)
False Pos = 22.37 (Row percent: 49/219)
False Neg = 31.67 (Row percent: 114/360)

Complete data

Percent correct = 29.26 + 43.26 = 72.52
Sensitivity = 74.24 (Col percent)
Specificity = 70.12 (Col percent)
False Pos = 22.37   (Row percent)
False Neg = 33.91   (Row percent)

Incomplete data

Percent correct = 70.43 + 0
Sensitivity = 0   (Col percent)
Specificity = 100 (Col percent)
False Pos = 0/0   (Row percent)
False Neg = 29.57 (Row percent)