3.3 Back to Logistic Regression

SPSS also permits to fit a logistic regression model using a regression interface. By using this interface we can get extra information on the model to asses its fit and predictive strength and use built in model selection strategies.
Unfortunately the reference category is different than the default used in the GLM user interface and can not be changed within the regression interface, requiring to calculate a new response variable.

Example 1
High school and beyond – data

SPSS commands to calculate new response

1. Transform >Compute Variable
2. Target Variable: `program_reverse`
   Numeric Expression: 1-program
3. OK

Now use the new variable to recreate the analysis done before with the GLM interface:

1. Analyze>Regression>Binary Logistic
   Dependent: `program_reverse`
   Covariates: x
   Method: Enter (not changed)
2. Option Tab: Check "CI for exp(B)"
3. Save Tab: Check "Probabilities" to get predicted values for $\pi(x)$

The output includes some descriptive information and the following (Start reading the information on Block 1)

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	139.082	1	.000
	Block	139.082	1	.000
	Model	139.082	1	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	692.268[a]	.207	.276

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Classification Table[a]

Observed			Predicted program_reverse .00	Predicted program_reverse 1.00	Percentage Correct
Step 1	program_reverse	.00	198	94	67.8
		1.00	81	227	73.7
	Overall Percentage				70.8

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
Step 1[a]	x	.027	.003	106.411	1	.000	1.028	1.022	1.033
	Constant	-7.055	.695	103.104	1	.000	.001		

a. Variable(s) entered on step 1: x.

In the Omnibus Tests of Model Coefficients table the row labelled Model is the information on the test comparing the current model with the intercept-only model, the same as we have seen using the GLM interface.

The Model Summary table provides extra information on model fit. Cox&Snell R^2 and Nagelkerke R^2 are measures of model fit and try to mimick the R^2 used in multiple linear regression. The disadvantage of Cox&Snell is, that it has an upper bound which can be much smaller than 1, making it really hard to judge. This has been corrected with Nagelkerke's R^2. See a discussion of measures of model fit related to R^2 at `http://statisticalhorizons.com/r2logistic`

The Classification table provides information on the predictive strength of the model. Classifying an observation as success when the estimated probability for success based on the model exceeds 0.5 (more likely to be success than failure), for every observation one can check if the prediction matches the observed category. The classification model provided this information. In this example 67.8% of the students in an academic program (`program_reverse`=0) and 73.7% of students in non-academic programs are correctly classified. This pretty good given we on use their total score on a number of exams.

The Variables in Equation table is similar to the one produced with the GLM user interface. Giving estimate, standard errors, test statistics, P-values, exponentiated estimates, and confidence intervals for the different parameters in the model.

Unfortunately the output does not include the deviance and other measures of model fit.

The advantage of the Regression user interface is that it has built in model selection tools like forward and backward selection.

Example 2

We will try to determine the best parsimonious model for predicting if a student attends an academic program.

The predictor variables that will be considered for inclusion in the model are:

sex, ses, sctype (school type), HSP (hours of sleep), Locus, concept, motivation, reading, writing, math, science, civics, x, math*sex interaction, ses*x interaction

It is x = reading + writing + math + science + civics, so the predictors are collinear and backward selection will not work.

1. Analyze>Regression>Binary Logistic
   Dependent: `program_reverse`
   Covariates: move all the variables named above into the covariate box
   interaction are created by clicking one of the variables, then press Ctrl and simultaneously click the second variable, move the interaction into the box. Method: Forward Wald

2. Categorical: Move ses, sex, and sctype onto the box to declare them categorical predictors.
   Continue.

3. Option Tab: Check "CI for exp(B)"
   OK

4. Save Tab: Check "Standardized" and "Deviance" residuals.
   Continue.

5. OK

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	139.082	1	.000
	Block	139.082	1	.000
	Model	139.082	1	.000
Step 2	Step	28.161	1	.000
	Block	167.243	2	.000
	Model	167.243	2	.000
Step 3	Step	14.341	1	.000
	Block	181.584	3	.000
	Model	181.584	3	.000
Step 4	Step	13.248	1	.000
	Block	194.832	4	.000
	Model	194.832	4	.000
Step 5	Step	9.837	2	.007
	Block	204.669	6	.000
	Model	204.669	6	.000
Step 6	Step	4.619	1	.032
	Block	209.289	7	.000
	Model	209.289	7	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	692.268[a]	.207	.276
2	664.107[a]	.243	.324
3	649.766[b]	.261	.348
4	636.518[b]	.277	.370
5	626.680[b]	.289	.385
6	622.061[b]	.294	.393

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

b. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

The output shows that the model was built in 6 steps, at each step adding a predictor. For each step the full results are provided.

As we can see the omnibus test is for all models significant, they all show significant better fit than the intercept only model. Both Cox&Snell and Nagelkerke R^2 keep improving showing that with every step the model improves. On the otherhand we can not see big improvements in the classification table, the last model classified 74% of the observations correctly, but with the first model already almost 71% were properly classified.

Classification Table[a]

	Observed		Predicted program_reverse .00	Predicted program_reverse 1.00	Percentage Correct
Step 1	program_reverse	.00	198	94	67.8
		1.00	81	227	73.7
	Overall Percentage				70.8
Step 2	program_reverse	.00	208	84	71.2
		1.00	73	235	76.3
	Overall Percentage				73.8
Step 3	program_reverse	.00	201	91	68.8
		1.00	63	245	79.5
	Overall Percentage				74.3
Step 4	program_reverse	.00	205	87	70.2
		1.00	71	237	76.9
	Overall Percentage				73.7
Step 5	program_reverse	.00	208	84	71.2
		1.00	71	237	76.9
	Overall Percentage				74.2
Step 6	program_reverse	.00	209	83	71.6
		1.00	73	235	76.3
	Overall Percentage				74.0

a. The cut value is .500

The last table finally shows in which order variables have been added to the model, and the parameter estimates for each model. It is relevant to observe that the final model does not include all main effects for the included interactions.
At this point the subject expert together with the statistician should decide if the interaction should be removed or the main effects should be added for the final model. In this case I would recommend to add ses, and remove math by sex for a final model.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	x	.027	.003	106.411	1	.000	1.028	1.022	1.033
	Constant	-7.055	.695	103.104	1	.000	.001		
Step 2[b]	sctype(1)	-1.421	.288	24.337	1	.000	.241	.137	.425
	x	.027	.003	99.251	1	.000	1.027	1.022	1.033
	Constant	-5.738	.747	58.952	1	.000	.003		
Step 3[c]	sctype(1)	-1.386	.289	23.067	1	.000	.250	.142	.440
	locus	.612	.165	13.684	1	.000	1.843	1.333	2.549
	x	.024	.003	69.665	1	.000	1.024	1.018	1.029
	Constant	-4.954	.774	41.008	1	.000	.007		
Step 4[d]	sctype(1)	-1.412	.292	23.296	1	.000	.244	.137	.432
	locus	.619	.169	13.428	1	.000	1.857	1.334	2.586
	science	-.063	.018	12.646	1	.000	.939	.907	.972
	x	.037	.005	58.900	1	.000	1.037	1.028	1.047
	Constant	-5.072	.784	41.887	1	.000	.006		
Step 5[e]	sctype(1)	-1.362	.296	21.120	1	.000	.256	.143	.458
	locus	.561	.171	10.765	1	.001	1.753	1.254	2.452
	science	-.065	.018	13.065	1	.000	.937	.905	.971
	x	.037	.005	59.145	1	.000	1.038	1.028	1.048
	ses * x			9.514	2	.009			
	ses(1) by x	-.003	.001	7.190	1	.007	.997	.995	.999
	ses(2) by x	-.002	.001	7.596	1	.006	.998	.996	.999
	Constant	-4.743	.803	34.922	1	.000	.009		
Step 6[f]	sctype(1)	-1.384	.298	21.515	1	.000	.250	.140	.450
	locus	.600	.173	11.963	1	.001	1.822	1.297	2.560
	science	-.076	.019	16.292	1	.000	.927	.894	.962
	math by sex(1)	.008	.004	4.539	1	.033	1.008	1.001	1.016
	x	.039	.005	61.590	1	.000	1.040	1.030	1.050
	ses * x			8.195	2	.017			
	ses(1) by x	-.003	.001	5.856	1	.016	.997	.995	.999
	ses(2) by x	-.002	.001	6.847	1	.009	.998	.996	.999
	Constant	-4.784	.806	35.200	1	.000	.008		

a. Variable(s) entered on step 1: x.

b. Variable(s) entered on step 2: sctype.

c. Variable(s) entered on step 3: locus.

d. Variable(s) entered on step 4: science.