C.7 Logistic Regression

The aim of this study is to explore the predictors of malnutrition in children from a deprived section of society. One hundred malnourished and 100 normal children of age between 1 to 5 years − matched for age within ±6 months and sex – were considered for the study. The potential predictors were decided either from the previous knowledge or significant predictors from univariate analysis. The dependent variable is group (malnutrition=1, normal=0) and the predictors are maternal education (illiterate=1, literate=0); daily income of parents (low=1, not low=0), immunization (no=1, yes=0), colostrum (not given=1, given=0), breast feeding till 6 months (no=1, yes=0), mode of feeding (bottle=1, others=0). Consider only these six predictors for this exercise.

Backward elimination method was applied with criteria $P < 0.05$ to enter and $P > 0.10$ to remove a predictor from the model. Hosmer-Lemeshow goodness-of-fit was applied to find appropriateness of model. Pesudo R-square was also calculated to assess the variability explained by the predictors (Pesudo R-square is analog to the R-square used in multiple linear regression). Contrasts define the indicator variables. Note that all the predictors are binary in this example.

The SPSS command to run backward elimination logistic regression:

```
LOGISTIC REGRESSION VARIABLES  Group
 /METHOD = BSTEP(LR) maternal_education Breast_feed_6months mode_feeding
  immunization_status daily_income colostrumgiven_code
 /CONTRAST (maternal_education)=Indicator(1)
 /CONTRAST  (Breast_feed_6months)=Indicator(1)
 /CONTRAST (mode_feeding)=Indicator(1)
 /CONTRAST (immunization_status)=Indicator(1)
 /CONTRAST (daily_income )=Indicator(1)
 /CONTRAST (colostrumgiven)=Indicator(1)
 /SAVE=PRED PGROUP
 /PRINT = GOODFIT SUMMARY CI(95)
 /CRITERIA = PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .
```

This command runs the backward elimination stepwise (based on likelihood ratio) procedure. CI(95) displays the 95% confidence interval of odds ratio. PIN and POUT specify the probability to enter and remove a predictor from the model. Logistic will predict the probability of being malnourished or normal on the basis of the given predictors. 0.5 is the cut-off point of probability chosen to classify children as malnourished or normal on the basis of the fitted model. The following confirms the coding for the dependent variable.

Dependent Variable Encoding

Original Value	Internal Value
0	0
1	1

The following table displays variables, their categories, frequency, and coding scheme. For example daily income of parents is divided into two categories: Low category has120 children

and Not low has 80 children. Category perceived as risk for malnourishment is coded as 1 and the other as 0.

Categorical Variables Codings

		Frequency	Parameter coding (1)
daily_income	Not low	80	.000
	Low	120	1.000
Breast_feed_6months	Yes	126	.000
	No	74	1.000
colostrumgiven_code	Given	140	.000
	Not given	60	1.000
mode_feeding	Other types	119	.000
	Bottle feeding	81	1.000
immunization_status	Immunized	150	.000
	Not immunized	50	1.000
maternal_education	Literate	110	.000
	Illiterate	90	1.000

The following table gives result when no predictor is in the model.

Block 0: Beginning Block

Classification Table[a,b]

			Predicted		
			Group		Percentage Correct
	Observed		0	1	
Step 0	Group	0	0	100	.0
		1	0	100	100.0
	Overall Percentage				50.0

a. Constant is included in the model.

b. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.000	.141	.000	1	1.000	1.000

The following table displays the initial univariate significance of each predictor with dependent variable. Score in this table is similar to chi-square. Mode_feeding is marginally significant and all others are highly significant when considered one at a time.

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	maternal_education(1)	13.657	1	.000
		Breast_feed_6months(1)	21.965	1	.000

	mode_feeding(1)	3.507	1	.061
	immunization_status(1)	24.000	1	.000
	daily_income(1)	10.083	1	.001
	colostrumgiven_code(1)	18.667	1	.000
Overall Statistics		65.333	6	.000

The following table displays the steps of backward elimination method and corresponding significance. In this method, since the number of predictors serially reduce, chi-square value at each subsequent step either decreases or remains same.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	77.950	6	.000
	Block	77.950	6	.000
	Model	77.950	6	.000
Step 2(a)	Step	-.666	1	.414
	Block	77.284	5	.000
	Model	77.284	5	.000
Step 3(a)	Step	-2.110	1	.146
	Block	75.174	4	.000
	Model	75.174	4	.000

a A negative Chi-squares value indicates that the Chi-squares value has decreased from the previous step.

The following table is showing the model summary. Last two columns are two versions of pseudo R-square. The value of pseudo R-square is decreasing because of backward elimination method. This output does not say which variables have been eliminated.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	199.309(a)	.323	.430
2	199.975(a)	.321	.427
3	202.085(a)	.313	.418

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

The following table shows the goodness of fit of the model at each step. At final steps the *P*-value 0.091 which is not significant. So we cannot reject the null hypothesis. In other words, the model can be considered adequate.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	6.001	8	.647
2	8.319	7	.305
3	12.306	7	.091

The following table shows the observed and expected frequencies at each step of elimination in normal and malnutrition categories. These are divided into 10 strata as required for Hosmer-

Lemeshow test. These start with nearly 20 subjects in each stratum. Some frequencies are less than 5 and that raises questions about validity of this test for these data.

Contingency Table for Hosmer -Lemeshow Test

		Group = Normal		Group = Malnutrition		Total
		Observed	Expected	Observed	Expected	Observed
Step 1	1	19	17.244	0	1.756	19
	2	21	20.765	3	3.235	24
	3	15	15.451	5	4.549	20
	4	11	12.959	9	7.041	20
	5	11	10.897	9	9.103	20
	6	10	9.019	12	12.981	22
	7	6	6.558	14	13.442	20
	8	2	4.095	17	14.905	19
	9	4	2.521	17	18.479	21
	10	1	.492	14	14.508	15
Step 2	1	31	27.784	0	3.216	31
	2	13	13.403	3	2.597	16
	3	14	16.790	9	6.210	23
	4	11	14.091	12	8.909	23
	5	11	9.469	9	10.531	20
	6	7	6.838	11	11.162	18
	7	6	5.864	14	14.136	20
	8	4	3.874	16	16.126	20
	9	3	1.887	26	27.113	29
Step 3	1	44	41.108	3	5.892	47
	2	14	16.914	9	6.086	23
	3	5	7.895	8	5.105	13
	4	14	11.178	7	9.822	21
	5	11	8.686	9	11.314	20
	6	3	6.239	17	13.761	20
	7	5	5.141	16	15.859	21
	8	3	2.427	19	19.573	22
	9	1	.412	12	12.588	13

The following table displays the classification of subjects as observed and predicted by the model. Logistic regression is used to calculate probability of each subject based on the model and subjects with probability more than 0.5 were classified as malnourished. The final logistic model was able to correctly predict 75% of the subjects. This is not high but that is the best achievable by these data when linear combination of these six predictors is considered.

Classification Table (a)

Observed	Predicted		
	Group		Percentage Correct
	Normal	Malnutrition	

Step 1	Group	Normal	72	28	72.0
		Malnutrition	25	75	75.0
	Overall Percentage				73.5
		Normal	72	28	72.0
		Malnutrition	27	73	73.0
	Overall Percentage				72.5
		Normal	77	23	77.0
		Malnutrition	27	73	73.0
	Overall Percentage				75.0

a The cut value is .500

Following table displays the coefficient (B=log-odds ratio), their standard error, Wald test, P-value, odds ratio = exp(B), and their 95% confidence interval.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	maternal_education(1)	.848	.362	5.477	1	.019	2.335	1.148	4.749
	Breast_feed_6months(1)	1.779	.381	21.802	1	.000	5.925	2.808	12.504
	mode_feeding(1)	.530	.358	2.195	1	.138	1.698	.843	3.423
	immunization_status(1)	1.833	.461	15.800	1	.000	6.256	2.533	15.449
	daily_income(1)	.312	.382	.669	1	.413	1.366	.647	2.886
	colostrumgiven_code(1)	1.529	.405	14.252	1	.000	4.616	2.086	10.212
	Constant	-2.284	.411	30.850	1	.000	.102		
Step 2[a]	maternal_education(1)	.923	.351	6.921	1	.009	2.517	1.265	5.006
	Breast_feed_6months(1)	1.825	.378	23.324	1	.000	6.202	2.957	13.007
	mode_feeding(1)	.515	.356	2.090	1	.148	1.674	.833	3.364
	immunization_status(1)	1.926	.449	18.381	1	.000	6.859	2.844	16.541
	colostrumgiven_code(1)	1.527	.405	14.233	1	.000	4.603	2.083	10.176
	Constant	-2.156	.375	33.068	1	.000	.116		
Step 3[a]	maternal_education(1)	.920	.349	6.956	1	.008	2.510	1.267	4.975
	Breast_feed_6months(1)	1.813	.375	23.420	1	.000	6.130	2.941	12.775
	immunization_status(1)	1.946	.450	18.691	1	.000	7.001	2.897	16.917
	colostrumgiven_code(1)	1.506	.399	14.262	1	.000	4.511	2.064	9.858
	Constant	-1.943	.335	33.711	1	.000	.143		

a. Variable(s) entered on step 1: maternal_education, Breast_feed_6months, mode_feeding, immunization_status, daily_income, colostrumgiven_code.

The predictors remained significant after elimination are maternal education, breast feeding till 6 months, immunization, and colostrum given. The other two were eliminated as not significant (see table given below). When these four variables are in the model, income and feeding by bottle or otherwise do not make any significant contribution to the prediction of malnourishment. Since our coding is 1 for the risk present and 0 for absent for all the predictors, positive values of the coefficients indicate that these are positively associated. Odds ratio of maternal education represents illiterate women were 2.510 times likely to have malnourished child compared to literate women as far as this model is concerned. Similarly, those mothers who did not breast feed till 6 months has 6.130 times likely to have malnourished child compared to mother who had breast feed more than 6 months. Highest odds ratio is for immunization. These ORs are adjusted for other terms in the model and measure independent contribution of these predictors.

The coefficients stated in the above table can be used to construct the logistic model. This is

$\ln[p/(1-p)] = -1.943 + 1.506(\text{colostrum}) + 1.946(\text{immunization}) + 1.813(\text{breast feeding}) + 0.920(\text{maternal education})$

For example, when all risk factors are present (all predictors = 1),
$\ln[p/(1-p)] = -1.943 + 1.506 + 1.946 + 1.813 + 0.920 = 4.242$. Odds ratio for this child to be malnourished is 4.242 relative to the one with no risk factor. This equation gives $p/(1-p) = e^{4.242} = 69.547$, and $p = 69.542/(1 + 69.547) = 0.986$. The estimated probability of a child with these risk factors being malnourished is 0.986.

The following table display the status of variable excluded from the model.

Variables not in the Equation

			Score	df	Sig.
Step 2(a)	Variables	daily_income(1)	.671	1	.413
	Overall Statistics		.671	1	.413
Step 3(b)	Variables	mode_feeding(1)	2.109	1	.146
		daily_income(1)	.562	1	.453
	Overall Statistics		2.765	2	.251

a Variable(s) removed on step 2: daily_income.
b Variable(s) removed on step 3: mode_feeding.

Command to obtain ROC curve based on this logistic regression:

```
ROC PRE_2 BY Group (1)
  /PLOT=CURVE(REFERENCE)
  /PRINT=SE COORDINATES
  /CRITERIA=CUTOFF(INCLUDE) TESTPOS(LARGE) DISTRIBUTION(FREE) CI(95)
  /MISSING=EXCLUDE.
```

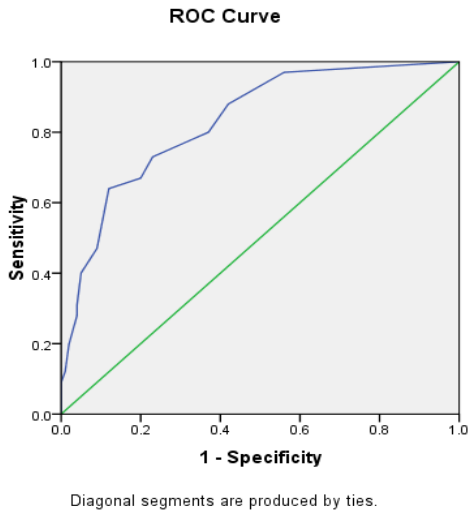

FIGURE C.7 ROC curve for predicted probabilities