A Tutorial on Logistic Regression in Dentistry

[1]Dr. Ramya Shenoy, [2]Dr. Harsh Priya

ABSTRACT

The present paper introduces the application of binary logistic regression to data pertaining to dentistry for use in publishing article. This article provides a brief overview of the type of logistic regression tests that are available to analyze research data and determine association and predict the model of concern in detail.

Key words: Logistic Regression, Binary Logistics, Multinomial Logistics, Statistical Tests

The present article describes the application of binary logistic regression[1] to data pertaining to dentistry. The study objective was to compare deft with two groups, one group was given health education and other group was control (data is enclosed).

Table 1: Frequency distribution of cases and controls

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Controls	275	51.9	51.9	51.9
	Intervention	255	48.1	48.1	100.0
	Total	530	100.0	100.0	

Since our interest was to determine the predictors for health education, then the numerical coding for cases will be bigger than controls say 2 & 1, respectively. SPSS will use the "higher coded" category to be the predicted outcome.

To perform the logistic regression[2] using SPSS, go to Analyse, Regression, Multinomial Regression.

Put cases and controls into dependent box. Put deft into covariates box and mother and father education, mother and father occupation into factor box. Since, mother and father education, mother and father occupation are categorical, we will need a reference group. E.g. For mother education it was coded like this 1=illiterate, 2=not upto high school, 3= upto high school, 4=Gradute. In this code 4 will be the reference category. This is required for easy interpretation of the results.

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	530	100.0
	Missing Cases	0	.0
	Total	530	100.0
Unselected Cases		0	.0
Total		530	100.0

(a) If weight is in effect, see classification table for the total number of cases.

Here 530 subjects are included in the analysis. A subject will be omitted from the analysis if any of the data point is missing, regardless of the availability of the others.

Dependent Variable Encoding

Original Value	Internal Value
Control group	0
intervention group	1

Control is the reference category.

Amount of variation explained by the model - Model Summary

Step	–2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	476.953[a]	.384	.513

(a) imation terminated at iteration number 20 because maximum iterations have been reached. Final solution cannot be found.

The Nagelkerke R Square[3] shows about 50% of the variation in the outcome variable (Health education) is explained by this logistic model.

[1]*Reader,* [2]*Assistant Professor, Dept. Public Health Dentistry, Manipal College of Dental Sciences, Mangalore, Manipal University*

Hosmer and Lemeshow Test

Step	Chi-square	Df	Sig.
1	43.383	8	.000

This shows how closely the observed and predicted probabilities match. If $p<0.05$ then it shows that the given model is not a good fit i.e. the sample size taken is not a true representative of the population. But this is true for small sample size; in case of large samples this may fail. This model is not a good fit[4].

Categorical Variables Codings

		Frequency	Parameter coding		
			(1)	(2)	(3)
paternal literacy	illiterate	25	1.000	.000	.000
	Not finished high school	375	.000	1.000	.000
	Not finished high school	125	.000	.000	1.000
	Graduate and higher	5	.000	.000	.000
paternal occupation	non skilled	295	1.000	.000	
	skilled	230	.000	1.000	
	professional	5	.000	.000	

The following table shows the reference category

The Final Output

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	ml			.002	3	1.000			
	ml(1)	1.623	1.150E4	.000	1	1.000	5.066	.000	.
	ml(2)	– 19.818	9.634E3	.000	1	.998	.000	.000	.
	ml(3)	– 19.833	9.634E3	.000	1	.998	.000	.000	.
	mo(1)	20.944	6.005E3	.000	1	.997	1.247E9	.000	.
	pl			2.563	3	.464			
	pl(1)	1.986	1.968E4	.000	1	1.000	7.283	.000	.
	pl(2)	22.447	1.797E4	.000	1	.999	5.608E9	.000	.
	pl(3)	22.873	1.797E4	.000	1	.999	8.583E9	.000	.
	deft	– .083	.025	11.265	1	.001	.920	.877	.966
	po			51.014	1	.000			
	po(1)	– 1.862	.261	51.014	1	.000	.155	.093	.259
	Constant	– 22.314	2.126E4	.000	1	.999	.000		

[a]Variable(s) entered on step 1: ml, mo, pl, deft, po.

How do we interpret the results here?

The above table is statistically stable only if the standard error is within 0.001 to 0.05. So **our table is statistically not stable.**

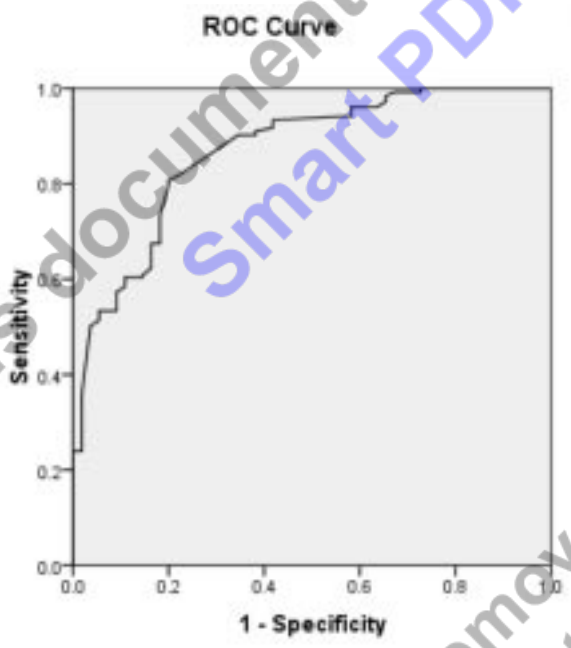
Here Wald[5,6] estimate gives the importance of contribution of each variable in the model. The **higher** the value the **more important** it becomes. The above output shows that father's occupation and deft plays important role in receptibility to health education given.

There is statistical significant result with deft, i.e. health education is the protective factor for deft.

Parental occupation i.e. non-skilled persons are more receptable to health education compared to skilled people.

Prediction model

How good is this model for prediction?



Diagonal segments are produced by ties

ROC curve is plotted to estimate the predictive ability.

The ROC[7,8] area is 0.869, which means almost 86% of all possible pairs of subjects which are in intervention group. This means excellent discrimination with case and control group.

Acknowledgements

My heartfelt Thanks to the H.O.D and Staff, The Department of Statistics, Manipal University.

REFERENCES

1. Agresti, A. (1990), Categorical Data Analysis. Wiley, New York.
2. SAS/STAT Software: Changes and Enhancements. Cary, NC. (The PHREG Procedure.) SAS Institute Inc. (1993), SAS Technical Report P-243.
3. Hosmer, D.W., Jr. and Lameshow, S. (1989), Applied Logistic Regression. Wiley, New York.
4. Albert A. and Anderson, J.A. (1984), "On the existence of maximum likelihood estimates in logistic regression models." Biometrika, 71, pp. 1-10.
5. SAS/STAT Software: The GENMOD Procedure. Cary, NC. Schlotzhauer, D.C (1993), "Some issues in using PROC LOGISTIC for binary logistic regression". Observations: The Technical Journal for SAS Software Users. Vol. 2, No. 4.
6. Santner T.J. and Duffy, E.D. (1986), "A note on A. Albert and J.A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models." Biometrika, **73**, pp. 755-758.
7. SAS Institute Inc. (1990), SAS/STAT User's Guide, Vol. 1 & 2, Version 6, Fourth Edition, Cary, NC. (The CATMOD, LOGISTIC, PROBIT procedures.) SAS Institute Inc. (1992), SAS Technical Report P-229.
8. Strauss, D. (1992), "The many faces of logistic regression."The American Statistician, Vol. 46, No. 4, pp. 321-326.