Biostatistics 202: Logistic regression analysis

Y H Chan

In our last article on linear regression[1], we modeled the relationship between the systolic blood pressure, which was a continuous quantitative outcome, with age, race and smoking status of 55 subjects. If our interest now is to model the predictors for SBP ≥180 mmHg, a categorical dichotomous outcome (Table I), then the appropriate multivariate analysis is a logistic regression.

Table I. Frequency distribution of SBP ≥180 mmHg.

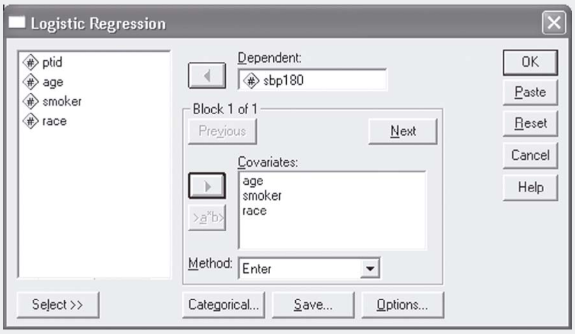
sbp ≥180

		Frequency	Percent	Valid percent	Cumulative percent
Valid	no	40	72.7	72.7	72.7
	yes	15	27.3	27.3	100.00
	Total	55	100.0	100.0	

Since our interest is to determine the predictors for SBP ≥180 mmHg, then the numerical coding for SBP ≥180 mmHg must be "bigger" than that of SBP <180 mmHg, say 1 & 0, respectively. SPSS will use the "higher coded" category to be the predicted outcome.

To perform the logistic regression using SPSS, go to ***Analyze, Regression, Binary Logistic*** to get template I.

Template I. Logistic regression.



Logistic Regression

Variables: ptid, age, smoker, race

Dependent: sbp180

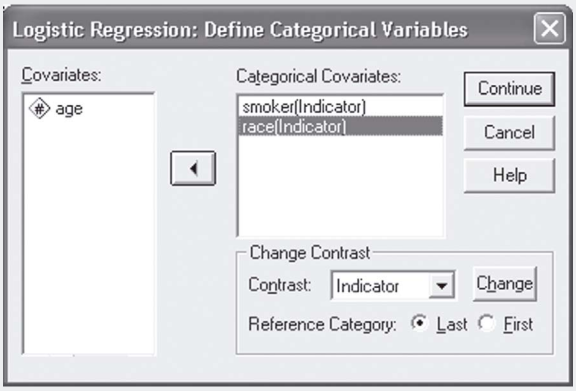
Block 1 of 1

Covariates: age, smoker, race

Method: Enter

Put sbp180 (the categorized SBP ≥180 mmHg & SBP <180 mmHg) in the Dependent box. Put age, race and smoker in the Covariates box. Click on the Categorical folder (in template I) to declare smoker and race as categorical variables (Template II).

Template II. Defining categorical variables.



Logistic Regression: Define Categorical Variables

Covariates: age

Categorical Covariates: smoker(Indicator), race(Indicator)

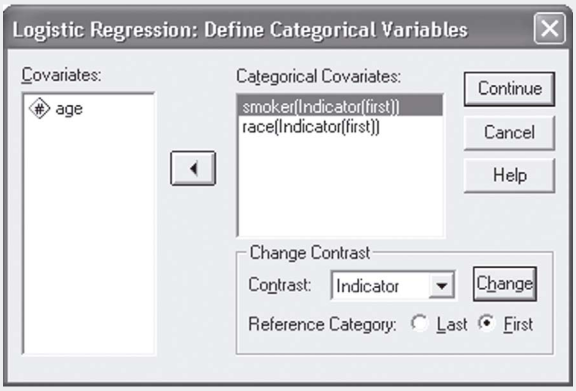
Change Contrast

Contrast: Indicator

Reference Category: ● Last ○ First

Since smoker and race are categorical, we will need a reference group (the default is the "highest coded" Last category). For race, usually we want the Chinese to be the reference and our standard coding is 1 = Chinese, 2 = Indian, 3 = Malay, 4 = Others, then we got to change the Reference Category (at the bottom of template II) to First and click on the Change button (Template III).

Template III. Changing the reference category.



Logistic Regression: Define Categorical Variables

Covariates: age

Categorical Covariates: smoker(Indicator(first)), race(Indicator(first))

Change Contrast

Contrast: Indicator

Reference Category: ○ Last ● First

Likewise, we have also changed the reference category for smoking to First as the coding is 1 = smoker and 0 = non smoker. The idea is to prepare the output for "easy interpretation"; that is, comparing the smoker with the non-smoker of having SBP ≥180. Tables IIa – IIe (only those of interest) are the output generated by SPSS when a logistic regression is performed.

Clinical Trials and Epidemiology Research Unit
226 Outram Road
Blk B #02-02
Singapore 169039

Y H Chan, PhD
Head of Biostatistics

Correspondence to:
Dr Y H Chan
Tel: (65) 6325 7070
Fax: (65) 6324 2700
Email: chanyh@cteru.com.sg

Table IIa. Number of cases in model.

Case processing summary

Unweighted Cases[a]		N	Percent
Selected cases	Included in analysis	55	100.0
	Missing cases	0	.0
	Total	55	100.0
Unselected cases		0	.0
Total		55	100.0

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

All 55 cases were included in the analysis. A subject will be omitted from the analysis if any one of his data point (for example, age) is missing, regardless of the availability of the others.

Table IIb. Predicted outcome coding.

Dependent variable encoding

Original value	Internal value
No	0
Yes	1

Table IIb is very important. It tells us which category SPSS is using as the predicted outcome, the higher coded category (having SBP ≥180 mmHg).

Table IIc. Amount of variation explained by the model.

Model summary

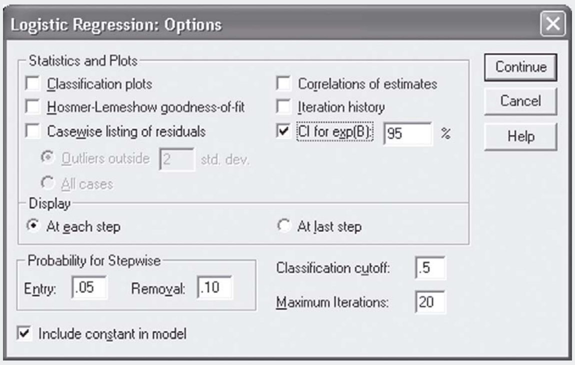
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	40.819	.349	.506

The Nagelkerke R Square shows that about 50% of the variation in the outcome variable (SBP ≥180) is explained by this logistic model.

How do we interpret the results in Table IId? Firstly, the **Wald** estimates give the “importance” of the contribution of each variable in the model. The higher the value, the more “important” it is.

If we are interested in a **predictor-model**, then both age and smoking status are important risk factors to having SBP ≥180, with p-values of 0.001 and 0.020 (given by the Sig column), respectively. The **Exp(B)** gives the **Odds Ratios**. Since age is a quantitative numerical variable, an increase in one-year in age has a 23.3% (95% CI 8.9% to 39.5%) increase in odds of having SBP ≥180. This 23.3% is obtained by taking Exp(B) for age – 1. To get the 95% CI, in Template I, click on the Options folder to get Template IV.

Template IV. Getting the 95% CI for the odds ratios.



Statistics and Plots

☐ Classification plots ☐ Correlations of estimates

☐ Hosmer-Lemeshow goodness-of-fit ☐ Iteration history

☐ Casewise listing of residuals ☑ CI for exp(B): 95 %

● Outliers outside 2 std. dev.

○ All cases

Display

● At each step ○ At last step

Probability for Stepwise

Entry: .05 Removal: .10

Classification cutoff: .5

Maximum Iterations: 20

☑ Include constant in model

Tick on CI for exp(B) for the 95% CI of the estimate.

In Table IId, what is SMOKER(1)? Table IIe shows the coding for the categorical variables. The reference group for a particular variable is given by the row of zeros. Thus for Smoker, the reference group is the non-smoker (as setup in Template III). A smoker compared to a non-smoker is 9.9 (95% CI 1.4 to 68.4) times more likely to have SBP ≥180.

Table IId. Estimates of the logistic regression model.

Variables in the equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I. for EXP(B)	
								Lower	Upper
Step 1[a]	AGE	.209	.063	11.007	1	.001	1.233	1.089	1.395
	SMOKER(1)	2.292	.986	5.401	1	.020	9.896	1.432	68.380
	RACE			1.627	3	.653			
	RACE(1)	.640	1.009	.402	1	.526	1.896	.263	13.696
	RACE(2)	1.303	1.136	1.316	1	.251	3.681	.397	34.101
	RACE(3)	-.097	1.230	.006	1	.937	.908	.081	10.113
	Constant	-14..462	4.005	13.041	1	.000	.000		

Table IIe. Categorical variables coding.

Categorical variables codings

			Parameter coding		
		Frequency	(1)	(2)	(3)
RACE	Chinese	23	.000	.000	.000
	Indian	13	1.000	.000	.000
	Malay	10	.000	1.000	.000
	Others	9	.000	.000	1.000
Smoker	No	23	.000		
	Yes	32	1.000		

For Race, Chinese is the reference category. In Table IId, Race(1) refers to comparing the Indian with Chinese, Race(2) refers to comparing the Malay with Chinese and lastly, Race(3) for Others comparing with Chinese. In Template III, observe that we can only declare either the first or last as the reference. If we want Malay to be the reference, a recode to make Malay having the smallest or largest coding is required.

CHECKING MULTICOLINEARITY

How to check for multicolinearity? To get the correlations between any two variables, in Template IV, tick on the Correlations of estimates option to obtain table III.

Table III. Correlation matrix for SBP model.

Correlation matrix

		Constant	SMOKER(1)	RACE(1)	RACE(2)	RACE(3)	AGE
Step 1	Constant	1.000	.345	-.326	-.265	-.415	-.953
	SMOKER(1)	.345	1.000	.073	.081	-.122	-.450
	RACE(1)	-.326	.073	1.000	.700	.652	.068
	RACE(2)	-.265	.081	.700	1.000	.585	.030
	RACE(3)	-.415	-.122	.652	.585	1.000	.215
	AGE	-.953	-.450	.068	.030	.215	1.000

Apart from the expected moderate to high correlations within Race, the correlation values among age, smoker and race are low. The correlation between age and the constant is rather high (r = -0.953) which shows some multicolinearity. What should be done? Before we answer this question, let us look at another example which quite commonly happens in a many-variables study. Table IV shows a 8-variable model with the correlation matrix between any two variables given in Table V.

Table IV. An 8-variable logistic model with multicolinearity.

Variables in the Equation

		B	S.E.	Wald	df	Sig.
Step 1	V1	-1062.640	56906.272	.000	1	.985
	V2	-2033.243	107665.309	.000	1	.985
	V3	-2282.536	121116.943	.000	1	.985
	V4	-462.334	26296.043	.000	1	.986
	V5	1000.935	53615.449	.000	1	.985
	V6	65.543	5358.046	.000	1	.990
	V7	764.889	40207.609	.000	1	.985
	V8	-62.261	4286.793	.000	1	.988
	Constant	-829.405	44003.539	.000	1	.985

In the correlation matrix for this case, it is not so easy to spot where the multicolinearity is! Another drawback with the correlation matrix is that multicolinearity between one variable with a combination of variables will not be shown.

A simple but sometimes subjective technique is to inspect the magnitude of the standard error (SE) of each variable. The SEs in Table IV are very large implying multicolinearity exists and the model is not statistically stable. To "solve" this issue, start omitting the variable with largest SE, continue the process until the magnitude of the SEs hover around 0.001 – 5.0. There is no fixed criterion on how small the SE should be but a matter of judgment.

In Table IId, the SEs are within the acceptable criterion but there was a high correlation between age and the constant – should one of them be omitted? The recommendation is to keep the constant term in the model as it acts as a "garbage bin", collecting all unexplained variance in the model (recall from Table IIc that the variables only explains 50%). How to omit the constant? In template IV, at the left hand corner, uncheck the "Include constant in model".

A PREDICTION MODEL

Frequently our interest is to use the logistic model to predict the outcome for a new subject. How good is this model for prediction?

Table V. Correlation matrix of the 8-variable model.

Correlation matrix

		Constant	V1	V2	V3	V4	V5	V6	V7	V8
Step 1	Constant	1.000	-.878	-.892	.965	-.920	-.924	-.917	-.523	-.412
	V1	-.878	1.000	.659	-.831	.743	.938	.766	.144	.389
	V2	-.892	.659	1.000	-.887	.866	.746	.809	.679	.222
	V3	.965	-.831	-.887	1.000	-.980	-.917	-.887	-.555	-.374
	V4	-.920	.743	.866	-.980	1.000	.877	.832	.598	.342
	V5	-.924	.938	.746	-.917	.877	1.000	.799	.280	.378
	V6	-.917	.766	.809	-.887	.832	.799	1.000	.620	.150
	V7	-.523	.144	.679	-.555	.598	.280	.620	1.000	-.155
	V8	-.412	.389	.222	-.374	.342	.378	.150	-.155	1.000

Table VI. Model discrimination.

Classification table[a]

			Predicted		
			SBP ≥180		Percentage
	Observed		no	yes	correct
Step 1	SBP ≥180	no	38	2	95.0
		yes	6	9	60.0
	Overall percentage				85.5

[a] The cut value is .500

The overall accuracy of this model to predict subjects having SBP ≥180 (with a predicted probability of 0.5 or greater) is 85.5% (Table VI). The sensitivity is given by 9/15 = 60% and the specificity is 38/40 = 95%. Positive predictive value (PPV) = 9/11 = 81.8% and negative predictive value (NPV) = 38/44 = 86.4%. How to use this information?

When we have a new subject, we can use the logistic model to predict his probability of having SBP ≥180. Let us say we have a black box where we input the age, smoking status and race of a subject and the output is a number between 0 to 1 which denotes the probability of the subject having SBP ≥180 (see Fig. 1).

Fig. 1 The logistic regression prediction model.

Age, race, smoking status of subject
Black box
Probability of having SBP ≥180

In the black box, we have the equation for calculating the probability of having SBP ≥180 which is given by

$$\text{Prob (SBP} \geq 180) = \frac{1}{1+e^{-z}}$$ where e denotes the exponential function

with z = -14.462 + 0.209 * Age + 2.292 * Smoker(1) + 0.640 * Race(1) +1.303 * Race(2) - 0.097 * Race(3)

The numerical values are obtained from the B estimates in Table IId.

For example, we have a 45-year-old non-smoking Chinese, then Smoker(1) = Race(1) = Race(2) = Race(3) = 0, and

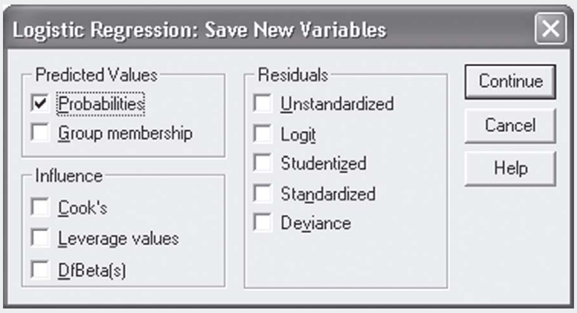
z = -14.462 + 0.209 * 45 = -5.057 and e^{-z} = 157.1 which gives the Prob (SBP ≥ 180) = 1/ (1 + 157.1) = 0.006; very unlikely that this subject has SBP ≥180 and the NPV tells me that I am 86.4% confident.

Let us take another example, a 65-year-old Indian smoker, then Smoker(1) = 1, Race(2) = Race(3) = 0 but Race(1) = 1. Hence z = -14.462 + 0.209 * 65 + 2.292 * 1 + 0.64 * 1 = 2.055 and e^{-z} = 0.128 which gives the Prob (SBP ≥180) = 1/(1 + 0.128) = 0.89; very likely that this subject has SBP ≥ 180 and the PPV gives a 81.8% confidence.

The default cut-off probability is 0.5 (and for this model, it seems that this cut-off gives quite good results). We can generate different probability cutoffs, by changing the 'Classification cutoff' in Template IV, and tabulate the respective sensitivity, specificity, PPV and NPV, then decide which is the best cut-off for optimal results.

The **area under the ROC curve**, which ranges from 0 to 1, could also be used to assess the model discrimination. A value of 0.5 means that the model is useless for discrimination (equivalent to tossing a coin) and values near 1 means that higher probabilities will be assigned to cases with the outcome of interest compared to cases without the outcome. To generate the ROC, we have to save the predicted probabilities from the model. In Template I, click on the Save button to get Template V.

Template V. Saving the predicted probabilities.



Logistic Regression: Save New Variables

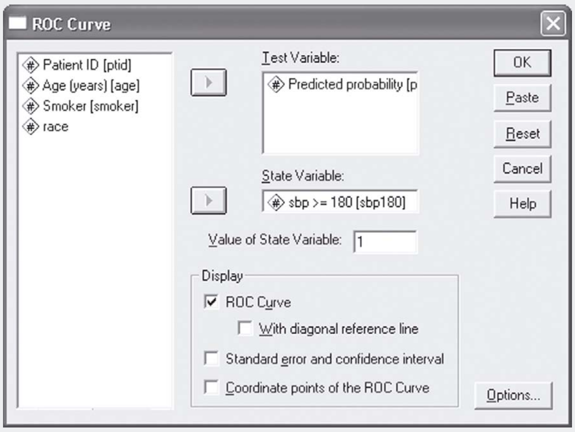
Predicted Values
- ☑ Probabilities
- □ Group membership

Influence
- □ Cook's
- □ Leverage values
- □ DfBeta(s)

Residuals
- □ Unstandardized
- □ Logit
- □ Studentized
- □ Standardized
- □ Deviance

Check the Predicted Values – Probabilities. A new variable, pre_1 (Predicted probability), will be created when the logistic regression is performed. Next go to ***Graphs, ROC curve*** – see Template VI.

Template VI. ROC curve.



ROC Curve

Patient ID [ptid]
Age (years) [age]
Smoker [smoker]
race

Test Variable: Predicted probability [p

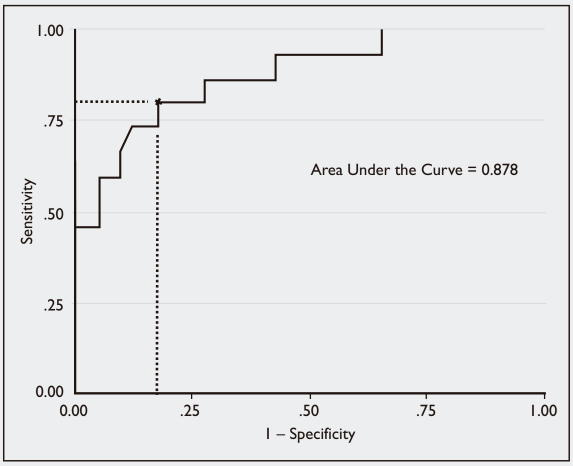
State Variable: sbp >= 180 [sbp180]

Value of State Variable: 1

Display
- ☑ ROC Curve
- □ With diagonal reference line
- □ Standard error and confidence interval
- □ Coordinate points of the ROC Curve

Put Predicted probability (pre_1) into the test Variable box, sbp180 in the State Variable and Value of State Variable = 1 (to predict SBP ≥180).

Fig. 2 ROC curve and area.



The ROC area is 0.878 (Fig. 2) which means that in almost 88% of all possible pairs of subjects in which one has SBP ≥180 and the other SBP <180, this model will assign a higher probability to the subject with SBP ≥180. The optimal sensitivity/ specificity is obtained from the point (*) nearest to the left upper corner of the box. Thus the optimal sensitivity = 78% and specificity = 1 - 0.18 = 82%.

Hosmer-Lemeshow goodness of fit (obtained by checking the relevant box in template IV) tells us how closely the observed and predicted probabilities match. The null hypothesis is "the model fits" and a p value >0.05 is expected (Table VII). Caution has to be exercised when using this test as it is dependent on the sample size of the data. For a small sample size, this test will likely indicate that the model fits and for a large dataset, even if the model fits, this test may "fail".

Table VII. Hosmer-Lemeshow test.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	5.869	7	.555

The above material covered the situation where the response outcome has only two levels. There are times when it is not possible to collapse the outcome of interest into two groups, for example stage of cancer. There are also situations where our study is a matched case-control. If in doubt, do seek help from a Biostatistician. The next article, Biostatistics 203, will be on Survival Analysis.

REFERENCE

1. Chan YH, Biostatistics 201: Linear regression analysis. Singapore Med J 2004; 45:55-61.