Multiple logistic regression

Dr Wan Nor Arifin

Unit of Biostatistics and Research Methodology,
Universiti Sains Malaysia.
E-mail: wnarifin@usm.my

Contents

1 Objectives

1. Extend the knowledge of simple logistic regression to multiple logistic regression.
2. Understand and apply model-building steps of multiple logistic regression for independent variables (dichotomous, polytomous and continuous).
3. Fit the logistic regression model on an example data in SPSS software.

2 Multiple logistic regression model

A simple logistic regression model is given as

$$z = \alpha + \beta x$$

$$E(Y|x) = P(Y = 1|x) = p = \frac{e^z}{1 + e^z} = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}$$

In case of multiple logistic regression, it can be extended as,

$$z = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p = \alpha + \sum \beta_i x_i$$

$$E(Y|\mathbf{x}) = P(Y = 1|\mathbf{x}) = p = \frac{e^z}{1 + e^z} = \frac{e^{\alpha+\sum \beta_i x_i}}{1 + e^{\alpha+\sum \beta_i x_i}}$$

notice the bold $\mathbf{x}$ to indicate collection of xs.

In effect,

$$ln \left(\frac{p}{1-p} \right) = logit(p) = \alpha + \sum \beta_i x_i$$

and the OR for specific x_i,

$$OR = e^{\beta i}$$

$$OR = e^{\beta i \Delta}$$

i.e. the OR for x_i while keeping other xs fixed. Or in standard words, while controlling for other variables.

3 Independent variables

In our previous lecture, we only discussed about fitting the simple logistic regression for binary categorical and continuous variables. Categorical variable with k>2 i.e more than 2 categories was skipped as it is easier to explain in context of multiple logistic regression.

Recall in multiple linear regression, MLR, the need to create $k - 1$ dummy variables. Similarly in logistic regression, $k - 1$ dummy variables (a.k.a design variables) have to be created. For example,

$$race \text{ (0: Malay, 1: Chinese, 2: Indian)}, k = 3 \text{ categories}$$

into $k - 1 = 2$ dummy variables, while treating Indian as reference category.

$$race1 \text{ (1: Malay, 0: Indian \& Chinese)}$$

$$race2 \text{ (1: Chinese, 0: Indian \& Malay)}$$

thus our model becomes,

$$logit(p) = \alpha + \beta_{race1} race1 + \beta_{race2} race2$$

* Unfortunately in SPSS, the dummy variables are automatically generated for you.

4 Determining the significance of the variables

4.1 Likelihood ratio test, *G*

$$G = -2[log\,likelihood\,model\,without\,x\,variable - log\,likelihood\,model\,with\,x\,variable]$$

$$G = -2(L_0 - L_1)$$

then the P-value is $P[\chi^2(1) > G]$, as G follows chi-square distribution. The degrees-of-freedom, $df = v$ i.e. difference in number of parameters between the models.

Alternatively, as it is given as -2 log likelihood in SPSS, or deviance D,

$$G = D(model\,without\,x\,variable) - D(model\,with\,x\,variable)$$

$$G = D_0 - D_1$$

LR test is preferred over Wald test for multiple logistic regression. In case of simple logistic regression, we used the LR test to determine the significance of a variable by comparing the deviance of model with the variable (D_1) and model containing constant only (D_0). For multiple logistic regression we can test whether a variable or variables significantly contribute to the model or not in similar way,

$$G = D(model\,without\,x\,variables) - D(model\,with\,x\,variables)$$

$$G = D_B - D_A$$

4.2 Wald test, *W*

$$W = \frac{\hat{\beta}}{\hat{SE}(\hat{\beta})}$$

then the two-tailed P-value is $P(|z| > W)$, as W follows standard normal distribution. It is more suitable for testing a single variable. In multiple logistic regression, judgment on importance of single variable can be made, but the final decision is best made by LR test.

5 Model-building steps

The following steps are based on purposeful selection steps by Hosmer, Lemeshow and Sturdivant (2013). The model building steps for the logistic regression basically consists of:

1. Variable selection.

   (a) **Univariable analysis.**

      i. Categorical variables: Chi-square test.

      ii. Numerical variables: Simple logistic regression. Independent-t/ANOVA not recommended.

   (b) **Multivariable analysis.**

      i. Fit selected variables.

         - All variables P-value $< .25$.
         - Clinically important variables

      ii. Fit a smaller model by removing non-significant variables.

   (c) **Comparison of larger to smaller model.**

      i. Check change in coefficients, $\Delta\hat{\beta} > 20\%$.

$$\Delta\hat{\beta}\% = 100\frac{(\hat{\beta_{small}} - \hat{\beta_{large}})}{\hat{\beta_{large}}}$$

      i. Identify excluded variables that cause the change.

      ii. Add back important variables (clinically important and confounders).

   (d) **Add unselected variables.**

      Identify variables that become significant.

      →*Preliminary main effects model.*

   (e) Close check on the selected variables.

      i. Linearity in logit for continuous variables.

      ii. **Numerical problems.**

         Cause very large coefficients and SEs.

         - Small cell counts – should be screened in 1(a).
         - Multicollinearity.
           - Between variables.
           - Indicate that the variables are redundant.
           - e.g. *Age* with *Age categories*, *Dead/Not dead* with *Pulse present/Pulse absent* etc.
           - Use appropriate correlation statistics.

   →*Main effects model.*

(f) **Interactions among variables.**
Among clinically plausible pairs – added to *Main effects model*.
→*Preliminary final model.*

2. Model fit assessment.

   (a) **Goodness-of-fit summary measures.**

      i. **Hosmer-Lemeshow test.**

         A. P-value > 0.05.

      ii. **Classification table.**

         A. Correctly classified $> 70\%$.

         B. Also calculate Specificity and Sensitivity.

      iii. **Area under Receiver Operating Characteristics (ROC) curve.**

         - AUC > 0.7

   (b) Regression diagnostics.

   (c) Cross-validation.

   → *Final model.*

We are going to cover only parts highlighted in **bold** only as the rests will be covered in Advanced Statistics in Semester 2.

6 Hands on in SPSS

Dataset: slog.sav (modified from a dataset, courtesy of AP Dr. Kamarul Imran Musa)

Dependent variable (DV): *cad* (1: Yes, 0: No)

Independent variables (IV): categorical – *race* (0: Malay, 1: Chinese, 2: Indian), *gender* (0: Female, 1: Male); numerical – *sbp* (systolic blood pressure), *dbp* (diastolic blood pressure), *chol* (serum cholestrol in mmol/L), *age* (age in years), *bmi* (Body Mass Index).

General SPSS steps:

1. Univariable analysis.

   (a) From the menu, **Analyze** → **Regression** → **Binary Logistic...**

   (b) In **Logistic Regression** window, assign **Dependent:** *cad*, **Covariates:** *sbp*.

   (c) Click on **Options...** button. In the window, choose **Iteration history** and **CI for exp(B)**. Click on **Continue** button. Click **OK** button.

(d) Repeat for the rest of numerical variables on by one.

(e) For categorical variables, perform chi-square test. from the menu, **Analyze** → **Descriptive Statistics** → **Crosstabs...**

(f) Assuming the data is from a cross-sectional study, assign **Rows:** *cad*, **Columns:** *gender*. Click on **Statistics...** button and choose **Chi-square**. Click on **Cells...** button and choose **Column** under **Percentages**. Click on **Continue** button. Click **OK** button.

(g) Repeat for *race*.

2. Multivariable analysis.

   (a) Following the general step in univariable analysis, assign all selected variables in **Covariates**.

   (b) For categorical variables, click on **Categorical...** button. In the window, place gender under **Categorical Covariates:**. Under **Change Contrast**, choose **First** (or **Last**) as **Reference Category:** and click on **Change** button. Click on **Continue** button.

   (c) Make sure the **Method** selected is **Enter**.

   (d) Click **OK** button.

3. Model fit assessment.

   (a) Hosmer-Lemeshow test & Classification Table – Click on **Options...** button. In the window, choose **Hosmer-Lemeshow goodness-of-fit**. Click on **Continue** button. Click **OK** button.

   (b) Area under ROC curve –

      i. To obtain *Predicted probability*, based on our *preliminary final model*, click on **Save...** followed by choosing **Probabilities** under **Predicted Values**. A new variable (usually *PRE_1*) will be created.

      ii. From the menu, **Analyze** → **ROC curve...** Assign **Test Variable:** *Predicted probability*, **State Variable:** *cad*, **Value of State Variable:** 0. Under **Display** choose **ROC Curve**, **With diagonal reference line** and **Standard Error and confidence interval**.

Perform the model building steps as outlined above in 5.

References

Hosmer, D. and Lemeshow, S. (2000). *Applied logistic regression (2nd eds).* Wiley Series in Probability and Statistics. USA: John Wiley & Sons, Inc.

Hosmer, D., Lemeshow, S., and Sturdivant, R. (2013). *Applied Logistic Regression*. Wiley Series in Probability and Statistics. Wiley.

Kleinbaum, D. and Klein, M. (2002). *Logistic regression: A self-learning text (2nd eds)*. Statistics for Biology and Health. USA: Springer New York.

Bartholomew, D. J., Steele, F., Moustaki, I., and Galbraith, J. I. (2008). *Analysis of multivariate social science data*. USA: CRC Press.