Logistic Regression

Wan Nor Arifin

Biostatistics and Research Methodology Unit, Universiti Sains Malaysia.

email: wnarifin@usm.my

Introduction

1. Statistical method to model relationship between:
   - outcome: binary categorical variable.
   - predictors/independent variables: numerical, categorical variables.
2. A type of Generalized Linear Models (GLMs).

Introduction

3. Basically, the relationship is structured as follows,

$$binary\ \ outcome = numerical\ \ predictors + categorical\ \ predictors$$

Introduction

4. More accurately, the *logistic* relationship structure,

$$log_e\left(\frac{proportion}{1 - proportion}\right) = numerical\ \ predictors + categorical\ \ predictors$$

We turned the binary outcome into proportion (p) of having the outcome. log_e is the *natural log*, sometimes written as *ln*.

The part, $\frac{p}{1-p}$ is known as *odds*.

Odds ratio vs relative risk

Association analysis for cross-tabulation of a binary factor and its outcome can be expressed as odds ratio.

- Odds is a measure of chance of disease occurence in a specified group,

$$Odds = \frac{n_{disease}}{n_{no\ disease}}$$

Odds ratio vs relative risk

- Odds ratio, OR is the ratio between the odds of two groups; the group with the risk factor and the group without the risk factor,

$$Odds\ ratio, OR = \frac{Odds_{factor}}{Odds_{no\ factor}}$$

- Odds ratio can be calculated for cohort, cross-sectional and case-control studied because it does not imply a cause-effect association, but only plain association.

Odds ratio vs relative risk

In epidemiology, it is common to describe the association between a risk factor and a disease in term of risk and relative risk.

- Risk is a measure of chance of disease occurence in a specified group, calculated as

$$Risk = \frac{n_{disease}}{n_{group}}$$

Odds ratio vs relative risk

- Relative risk is the ration between the risk in the group with the factor and the risk in the group without the risk factor,

  $$Relative\ risk, RR = \frac{Risk_{factor}}{Risk_{no\ \ factor}}$$

  It is only approriate to calculate risk and relative risk for cohort studies, because the cause-effect relationship is well defined.

OR is a good approximation of RR whenever the disease is rare. Rare diseases are commonly studied using case-control studies, thus the use of ORs are justified.

Odds ratio vs relative risk

As an example, we can calculate odds, OR, risk and RR from the following table.

Table 1: Smoker vs lung cancer

	Lung cancer	No lung cancer	Marginal total	Odds	Risk
Smoker	20	12	32	$20/12 = 1.667$	$20/32 = 0.625$
Non smoker	95	73	168	$95/73 = 1.301$	$95/168 = 0.565$

Thus OR and RR equal,

$$OR = 1.667/1.301 = 1.281$$

$$RR = 0.625/0.565 = 1.106$$

Simple logistic regression (SLogR)

1. Model relatioship between:
   - outcome: binary categorical variable.
   - a predictor: numerical or binary categorical variable.
2. Formula,

$$log_e\left(\frac{p}{1-p}\right) = intercept + coefficient \times numerical/binary\ \ predictor$$

Simple logistic regression (SLogR)

or in a proper equation form,

$$log_e\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1$$

Simple logistic regression (SLogR)

3. Odds ratio is easily obtained from a logistic regression,

$$OR_1 = e^{\beta_1}$$

4. p – proportion/probability. To obtain p,

$$p = \frac{e^{\beta_0 + \beta_1 x_1}}{1 + e^{\beta_0 + \beta_1 x_1}}$$

But as we will see later, this can be easily obtained in R.

Multiple logistic regression (MLogR)

1. Model relatioship between:
   - outcome: binary categorical variable.
   - predictors: numerical, categorical variables.
2. Formula,

$$log_e\left(\frac{p}{1-p}\right) = intercept + coefficients \times numerical\ \ predictors$$
$$+ coefficients \times categorical\ \ predictors$$

Multiple logistic regression (MLogR)

or in a nicer form,

$$log_e\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

where we have k predictors.

Whenever the predictor is a categorical variable with more than two levels, remember to consider dummy (binary) variable(s).

Analysis steps

1. Library
2. Load data
3. Data exploration
    - descriptive
4. Univariable
5. Multivariable
    - all selected
    - stepwise
    - confounder

Analysis steps

6. Multicollinearity, MC
7. Interaction
8. Model fit
    - Hosmer-Lemeshow
    - Classification table
    - AUC/C-stat
9. Interpretation
10. Equation
11. Prediction