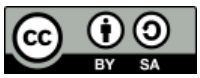
Simple logistic regression

Dr Wan Nor Arifin

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



Contents

1 Objectives

1. Understand the reasoning behind the move from linear regression model to logistic regression model for dichotomous outcome.
2. Identify logistic regression model formula.
3. Understand the concept of odds ratio.
4. Identify logit transformation function formula.
5. Understand how odds ratio is calculated from a logistic regression model and interpret the odds ratio.
6. Understand how a logistic regression model is fit by maximum likelihood estimation method.
7. Understand how to test significance of the coefficients.
8. Fit the logistic regression model on an example data in SPSS and R software – dichotomous and continuous independent variables.

2 Linear vs logistic for dichotomous outcome

In place of a numerical outcome, y in linear regression model, now we have a categorical outcome with two levels (yes/no, disease/no disease) coded as 0/1. Let us review back the linear regression model. Expected value of y, or conditional mean of y given x is

$$E(Y|x) = \alpha + \beta x$$

where the conditional mean should be $0 \leq E(Y|x) \leq 1$. But as x ranges between $-\infty$ to $+\infty$, it is impossible to keep both sides equal. Thus we need a different model for dichotomous outcome, an alternative is **logistic regression model.**

3 Logistic regression model

A logistic regression model is given as

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

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

A good thing about this model is that, now we can keep $0 \leq E(Y|x) \leq 1$, i.e the probability of having a disease, p is between 0 to 1. Before delving further into the logit transformation function that has a desirable use for epidemiologist, we need to understand the concept of odds and odds ratio.

4 Relative risk vs odds ratio

Consider Table 1 for a cohort study of relationship between smoking (exposure) and lung cancer (outcome).

Table 1: Lung cancer vs smoking

	Lung cancer	No lung cancer
Smoking	20	12
No smoking	95	73

Compare it to Table 2 for a case-control or cross-sectional study between lung cancer and smoking.

Table 2: Smoking vs lung cancer

	Smoking	No smoking
Lung cancer	20	95
No lung cancer	12	73

For Table 1, we are able to calculate relative risk of developing lung cancer for smoker vs non-smoker for the cohort study.

$$Risk_{smoking} = \frac{20}{20+12} = \frac{20}{32} = 0.625$$

$$Risk_{no\,smoking} = \frac{95}{95+73} = \frac{95}{168} = 0.565$$

$$Relative\,risk_{\frac{smoking}{no\,smoking}} = \frac{0.625}{0.565} = 1.106$$

However, for Table 2, it is inappropriate to calculate relative risk due to the study design (revise back your knowledge in epidemiology on causality). In this situation, we may calculate odds and odds ratio.

$$Odds_{lung\,cancer\,being\,smoker} = \frac{20}{95} = 0.211$$

$$Odds_{no\,lung\,cancer\,being\,smoker} = \frac{12}{73} = 0.164$$

$$Odds\,ratio_{\frac{lung\,cancer\,being\,smoker}{no\,lung\,cancer\,being\,smoker}} = \frac{0.211}{0.164} = 1.287 \approx 1.106$$

We notice here that this value approximates the relative risk. If we calculate the odds ratio for the cohort study as below

$$Odds\,ratio_{\frac{smoker\,with\,lung\,cancer}{non\,smoker\,with\,lung\,cancer}} = \frac{20}{12} \div \frac{95}{73} = \frac{20 \cdot 73}{12 \cdot 95} = 1.281$$

which is the odds ratio calculated for Table 2. Note that the value is slightly different due to rounding error.

The use of odds ratio is appealing because it is easily applicable for all the study designs. In addition, in relation to the logistic regression model, odds ratio can be obtained from the model based on the relevant coefficients.

5 Logit transformation

By applying a logit link function on the basic logistic regression formula above, it allows calculation of the odds and subsequently the odds ratio. The function is linear and can range from $-\infty$ to $+\infty$.

$$logit(p) = ln\left(\frac{p}{1-p}\right)$$

$$p = \frac{e^z}{1+e^z}$$

$$1 - p = 1 - \frac{e^z}{1+e^z} = \frac{1+e^z-e^z}{1+e^z} = \frac{1}{1+e^z}$$

$$\frac{p}{1-p} = \frac{e^z}{1+e^z} \div \frac{1}{1+e^z} = \frac{e^z(1+e^z)}{1+e^z} = e^z$$

$$ln\left(\frac{p}{1-p}\right) = ln(e^z) = z = \alpha + \beta x$$

thus

$$logit(p) = \alpha + \beta x$$

is the $ln(odds)$. However, in the current form, it not really useful. What we need is the odds ratio.

Odds ratio, OR when $x = 0, 1$

$$OR = \frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}} = \frac{e^{z_1}}{e^{z_0}} = \frac{e^{\alpha+\beta(1)}}{e^{\alpha+\beta(0)}} = \frac{e^{\alpha+\beta}}{e^{\alpha}} = e^{\beta}$$

OR, when x is continuous numerical variable e.g. age

$$OR = \frac{e^{\alpha+\beta x_1}}{e^{\alpha+\beta x_0}} = e^{\beta(x_1-x_0)}$$

where x_1 and x_0 are any two values of x. As we are most interested in change or difference between the values, $\Delta = x_1 - x_0$

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

or for 1 unit change in x

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

which is same formula applied when x is binary.

6 Maximum likelihood estimation method

Linear regression model uses ordinary least squares (OLS) estimation method to obtain values of α and β that minimize sum of squared deviations of the observed values of Y from the ones predicted by the model.

However, for a nonlinear model like logistic regression model, OLS cannot be used. Instead, maximum likelihood (ML) estimation method can be used to estimate the unknown parameters α and β.

In ML estimation method, a likelihood function $l(\theta)$ that indicates the likelihood of observing the data for a set of unknown parameters $\theta = \alpha, \beta$, is specified.

$$l(\theta) = \prod_{i=1}^{n} p(x_i)^{y_i}[1 - p(x_i)]^{1-y_i}$$

or in form of log likelihood

$$L(\theta) = ln[l(\theta) = \sum_{i=1}^{n} \{y_i ln[p(x_i)] + (1 - y_i) ln[1 - p(x_i)]\}$$

ML estimators $\hat{\theta}$ that maximize $ln[l(\theta)]$ is then found iteratively by computer software as it is complicated to solve the equation by usual means.

As an example, using the data in Table 2,

$$l(\theta) = p(1)^{20} \times [1 - p(1)]^{12} \times p(0)^{95} \times [1 - p(0)]^{73}$$

7 Determining the significance of the variables

7.1 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,

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

$$G = D_0 - D_1$$

7.2 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.

8 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 – *gender* (1: Male, 0: Female), numerical – *sbp* (systolic blood pressure)

Steps:

1. From the menu, **Analyze** → **Regression** → **Binary Logistic...**
2. In **Logistic Regression** window, **Dependent:** *cad*, **Covariates:** *gender*.
3. Click on **Categorical...** button. In the window, place gender under **Categorical Covariates:**. Under **Change Contrast**, choose **First** as **Reference Category:** and click on **Change** button. Click on **Continue** button.
4. Click on **Options...** button. In the window, choose **Iteration history** and **CI for exp(B)**. Click on **Continue** button.
5. Click **OK** button.
6. Repeat the same steps for *sbp*.

9 Hands on in R

```
data = read.csv("slog.csv")

# gender, categorical
table(gender=2-data$gender, cad=2-data$cad)
lreg = glm(cad ~ gender, data = data, family = binomial(link = "logit"))
summary(lreg)
cbind(coef(lreg), confint(lreg))
exp(cbind(coef(lreg), confint(lreg)))

# sbp, numerical
lreg1 = glm(cad ~ sbp, data = data, family = binomial(link = "logit"))
summary(lreg1)
cbind(coef(lreg1), confint(lreg1))
exp(cbind(coef(lreg1), confint(lreg1)))

# an increase in 10mmHg in sbp
exp(cbind(coef(lreg1)[[2]]*10, confint(lreg1)[[2]]*10, confint(lreg1)[[4]]*10))
```

References

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

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

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