Categorical Data Analysis - Logistic Regression

Shengping Yang PhD ,Gilbert Berdine MD

I am planning a case-control study on lung cancer and body mass index (BMI). I think that this information would not fit a normal distribution. I would like to know more about how to analyze data from such a study.

Case-control studies are widely used in investigating the potential relationship of a suspected risk factor and a disease or outcome of interest. By looking retrospectively and comparing how frequently the exposure to a risk factor is present in subjects who have that disease (case) with those who do not have that disease (control), the relationship can be evaluated. The outcome variable can only take exactly two values, conventionally labeled as "case" and "control". In fact, this type of variable is called a **categorical/nominal** variable[1] (data that have two or more categories, but there is no intrinsic ordering to the categories).

Since a categorical outcome variable can take only a few (two in case-control studies) possible values, its distribution can be very different from normal. Thus many of the statistical methods developed for analyzing data with normally-distributed outcome variables are not suitable for analyzing data with categorical outcomes. *Note that those methods are also not suitable for analyzing data with* ***ordinal*** *(a statistical data type consisting of numerical scores that exist on a rank scale) or* ***cardinal*** *(a type of data in which observations can take only the non-negative integer values {0, 1, 2, 3, ...}, and where these integers arise from counting rather than ranking) outcome variables. Binary logistic regression (we will drop "binary" for simplification purpose) is widely used in case-control study data analyses. In this column, we will provide some details on the application, assumption, interpretation, and pitfalls of logistic regression.*

Corresponding author: Shengping Yang
Contact Information: Shengping.yang@ttuhsc.edu.
DOI: 10.12746/swrccc2014.0207.094

1.The basics of logistic regression

In the previous article, we showed how linear regression "fits" data point pairs of a continuous dependent variable x and a continuous variable y to the linear function $\hat{y}=mx+b$. Our case-control study cannot use this method, because our outcome variable y can take only values of 'case' or 'control'. Logistic regression solves this problem by transforming a non-linear equation into a linear form.

The first step is the use of the logistic function:

$$F(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}$$

The variable *t* can take any value from -∞ to +∞. The variable *t* will be 'fit' using regression methods to a linear function of our explanatory variable *x*. The explanatory variable in our example would be *BMI*. The linear model is:

$$t = \beta_0 + \beta_1 x$$

The logistic function becomes:

$$g(x) = \log\left(\frac{F(x)}{1 - F(x)}\right) = t = \beta_0 + \beta_1 x$$

The physical meaning of β_0 is the 'intercept' or log-odds of being a 'case' when the explanatory variable has a value of 0, if 0 is achievable. The physical

meaning of β_1 is the parameter which defines the rate of change in the log-odds with changes in the explanatory variable *(BMI).*

In order to estimate the regression coefficients, numeric methods, such as the Newton-Raphson iteration, are usually used because it is not possible to find a closed-form expression for the coefficient values. The Newton-Raphson iteration takes the form:

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

where x_1 is the new estimate, x_0 is the previous estimate, $f(x_0)$ is the value of the function for the previous estimate, and $f'(x_0)$ is the value of the first derivative for the previous estimate. The Newton method is well suited to automated computing provided the function is differentiable and the estimates converge to a single defined value.

2. Application of logistic regression in case-control studies.

In the example of a lung cancer study, the objective is to assess whether lung cancer is significantly associated with *BMI.* The two possible outcomes are: developed lung cancer and no lung cancer, respectively; and we want to evaluate the effect of *BMI* on lung cancer, while controlling for smoking and other risk factors.

A variety of software can be used for performing logistic regression analysis, such as SAS, Stata, SPSS, S-Plus/R, and Minitab. Since SAS is one of the most widely used software in statistics, below we provide the SAS code example for analyzing the lung cancer study data.

```
proc logistic descending;
class smoking;
model disease = BMI smoking <other risk
factors>;
run;
```

The *proc logistic* procedure is used for modeling the probability of developing lung cancer. The outcome variable Disease is a categorical variable, coded as "1" for subjects who developed cancer and "0" for those who did not. While *BMI* is treated as a continuous variable (we can later treat *BMI* as a categorical variable as well to see how it is associated with lung cancer), the class statement tells SAS that smoking is a categorical variable. The option *descending* is used by default to be consistent with how the outcome variable is coded.

3. Assumptions of logistic regression.

There are several assumptions underlying a logistic regression model. Since some of them are quite technical, we will skip them and focus only on the following three that are particularly relevant to a case-control study.

(a) No important variables are omitted.

Not including known risk factor(s) in a logistic regression model creates estimation bias, because compensating for the missing risk factor(s) results in over- or underestimating the effect of other risk factors. Therefore, it is important for researchers to make sure that all known potential risk factor/confounder data are collected. For example, in the lung cancer study, while our objective is to investigate the association between lung cancer and BMI, we still need to simultaneously collect data on smoking, family history of cancer, exposure to pollution, and any other known confounding variables.

(b) The observations are independent.

When this assumption is violated, the estimated standard errors are incorrect, as are the inferences. To avoid this violation, the study design and sampling plan have to be developed properly.

(c) No severe collinearity among independent variables is present.

Collinearity occurs when two or more predictor variables in a multiple regression model are highly correlated. For example, gestational age and birth weight are highly correlated, i.e., low (high) gestational age is usually associated with low (high) birth weight. Including both variables in a logistic regression model will cause collinearity. Severe collinearity inflates the standard errors for the coefficients, which causes the estimated coefficients to be unreliable. Therefore, considerations need to be taken in the study planning stage to avoid causing collinearity problems.

4. Interpretation of logistic regression.

By definition, the odds of an event (disease) is the ratio of the probability that an event will occur to the probability that the event will not occur. In the lung cancer study, suppose that we have the following data:

	Developed lung cancer	No lung cancer
Smoker	n_{sc}	n_{sn}
Non-smoker	n_{nc}	n_{nn}

The odds of developing lung cancer for smokers is n_{sc}/n_{sn}, and for non-smokers is n_{nc}/n_{nn}. The odds ratio (OR) is the ratio of these two, thus,

$$OR = \frac{n_{sc}/n_{sn}}{n_{nc}/n_{nn}} = \frac{n_{sc}\,n_{nn}}{n_{sn}\,n_{nc}}$$

Numerically, suppose n_{sc}=400, n_{nc}=100, n_{sn}=300, and n_{nn}= 700, then OR=(400×700)/(300×100) = 9.33.

In the above example, there is only one risk factor (smoking), and the odds ratio calculated is called **raw** odds ratio. Logistic regression analysis can handle models with multiple risk factors, and provide odds ratio estimates for each risk factor while adjusting for all other risk factors (called **adjusted** odds ratio). Now suppose that the adjusted odds ratio for smoking is 8.55 (with *P* value less than a pre-specified significance level); then we can interpret it as: The odds of lung cancer is 8.55 times as high for smokers than for non-smokers given other risk factors equal.

5. Pitfalls in interpretation of logistic regression.

As one of the major limitations of an observational study, a logistic regression can be used only for detecting association, rather than causation. For example, supposing we found a significant association between lung cancer and smoking, we cannot conclude that smoking causes lung cancer because there are alternative explanations - "The same thing that causes people to smoke may predispose them to lung cancer[3]." Therefore, further studies have to be conducted to verify that a causal effect does exist.

Another issue associated with logistic regression is the interpretation of odds ratio. Clinicians think in probabilities, not odds. Although odds ratios are valid measurements of strength of an association, many times they are not good indications of relative risk (RR; the ratio of the probability of an event occurring in an exposed group to the probability of the event occurring in a non-exposed group). In fact, odds ratio can be used as a proxy for relative risk only when the assumption of "rare" event is met[2]. For a "rare" event, the probabilities of an event for both the exposed and non-exposed groups are very small, i.e., we have both $P(event \mid exposure) \approx 0$ and $P(event \mid non\text{-}exposure) \approx 0$. *Therefore,*

$$OR = \frac{P(event|exposure)/[1-P(event|exposure)]}{P(event|non\text{-}exposure)/[1-P(event|non\text{-}exposure)]}$$

$$\approx \frac{P(event|exposure)}{P(event|non\text{-}exposure)} = RR$$

Sample size calculation is critical to the success of a case-control study. In general, sample size increases with smaller effects and smaller pre-defined Type I and Type II errors. We will discuss sample size calculation issues in future articles.

Author affiliation : Shengping Yang is a biostatistician in the Department of Pathology at TTHUSC. Gilbert Berdine is a pulmonary physician in the Department of Internal Medicine at TTUHSC.
Submitted: 5/2/2014
Accepted: 6/1/2014
Published electronically: 7/13/2014

References

1. Agresti A. Categorical Data Analysis. John Wiley & Sons, Inc. 2013. 1-7; 163-191. Print.
2. Grimes DA, Schulz KF. Making sense of odds and odds ratios. Obstetrics & Gynecology 2008; 111(2): 423-426.
3. Milberger S, et al. Tobacco manufacturers' defense against plaintiffs' claims of cancer causation: throwing mud at the wall and hoping some of it will work. Tob Control 2006; 15(Suppl 4): iv17-iv26.