JOURNAL OF CLINICAL RESEARCH BEST PRACTICES

Vol. 6, No. 4, April 2010 “Can You Handle the Truth?”

Making Sense of Biostatistics: Logistic Regression (Part 1)

By Ronald E. Dechert

In previous columns, we have reviewed correlation, simple linear regression, and multivariate linear regression. In each of these columns, we were interested in the association between a dependent variable and one or more independent variables. These methods work well with dependent variables that are continuous (e.g., 1, 2, 3… or 1.27, 3.59, 2.15), but not with dependent variables that are categorical, such as clinical outcomes like life/death or the presence/absence of a disease.

To determine the influence of one or more independent variables on a binary clinical outcome, we need to employ a different type of analysis: logistic regression. With logistic regression, the outcome (dependent) variable equals 0 if the outcome is negative and 1 if the outcome is positive. The independent variables can still take many different forms, such as continuous (age, weight), binomial (gender), nominal (race), or ordinal (pain scale). Logistic regression can also be used when the outcome variable takes more than two responses, e.g., mild, moderate or severe. That form of logistic regression is referred to as “multinomial logistic regression” and is beyond the scope of this column.

Since the outcome variable can only take a value of 1 if the condition is present or 0 if absent, logistic regression is actually analyzing the probability that Y will be 1 or 0 when the various values of X are controlled for. In other words, logistic regression enables investigators to predict the likelihood that the outcome will be positive or negative given certain values of X. Logistic regression uses equations of the following form:

$$P(Y=1) = \frac{\exp^{(B0 + B1X1 + B2X2\ldots.+BnXn)}}{1 + \exp^{(B0 + B1X1 + B2X2\ldots.+BnXn)}}.$$

In this equation, P is the probability that the outcome will be positive, exp is the exponential function (which always takes the value of 2.72) used to derive probability, and the B’s are the coefficients for each independent variable. The following example derived from an actual logistic regression analysis will explain the equation.

Dr. Sam Munson has a patient registry that records patient age, gender (0=male, 1=female), cardiac risk index, whether the patient is following a nutritional program defined by the study protocol, and presence/absence of heart disease. Dr. Munson wants to use his registry to determine the probability of developing heart disease in patients who were on his nutritional program compared to those who were not, when controlling for age and cardiac risk index. Five-hundred subjects have participated in the registry for an average of 10 years.

Dr. Munson runs a logistic regression using heart disease as the outcome variable and age, gender, cardiac risk index, and on-program as the independent variables. The results of the analysis are shown in Table 1.

Table 1. Logistic Regression Results

Independent Variable	B	Beta (B)	Significance	Exp(B) *	Lower CI**	Upper CI**
Gender (0=male, 1=female)	B_1	-0.138	0.393	0.871	0.634	1.197
Cardiac Risk Index	B_2	-0.006	0.001	0.994	0.991	0.996
On-Protocol (0=No, 1=Yes)	B_3	-0.628	0.001	0.534	0.388	0.733
Age (Years)	B_4	0.043	0.001	1.044	1.034	1.054
Constant	B_0	-1.544	0.001	0.214	NA	NA

* Exp(B) = Odds Ratio
** CI = Confidence Interval

With these results, Dr. Munson can develop a predictive model of heart disease. He simply inserts the coefficients in the table (Beta) for each independent variable into the logistic equation.

Suppose Dr. Munson has a new patient, Jane Doe, and wants to determine the probability of heart disease for that patient. Ms. Doe is a 65-year-old female who has been on his nutrition program and has a cardiac risk index of 155. Her probability of developing heart disease is:

$$\text{Prob (Heart Disease)} = \frac{2.72^{(-1.544 + (-0.138*1) + (-0.006*155) + (-0.628*1) + (0.043*65))}}{1 + 2.72^{(-1.544 + (-0.138*1) + (-0.006*155) + (-0.628*1) + (0.043*65))}}$$

$$= \frac{2.72^{(-0.445)}}{= 1 + 2.72\ (-0.445)}$$

$$= 0.64 / 1.64$$

$$= 0.39$$

From the logistic regression, Dr. Munson is able to tell Ms. Doe that the probability for patients like her developing heart disease in the next 10 years is 39% (0.39).

This example demonstrates one of the chief attributes of logistic regression: its ability to predict outcomes like heart disease based on a patient's characteristics. Other attributes that can be derived from logistic regression analysis will be reviewed in the next column.

Author

Ronald E. Dechert, DPH, is Associate Director of the Mott Respiratory Care at the University of Michigan Medical Center. Contact him at 734.936.5237 or rdechert@umich.edu.