Application of Odds Ratio and Logistic Models in Epidemiology and Health Research

By: Min Qi Wang, PhD; James M. Eddy, DEd; Eugene C. Fitzhugh, PhD

Wang, M.Q., Eddy, J.M., & Fitzhugh, E.C.* (1995). Application of odds ratios and logistic models in epidemiology and medical research. *Health Values, 19,* 1, 59-62.

Article:

M any researchers in the health field use the chi-square statistic to identify associations between variables. This edition of research notes will demonstrate that the odds ratio may be a preferred analysis to yield more useful and meaningful results. In epidemiological and health contexts, the outcome variable is often discrete, taking on two (or more) possible scores. For example, the outcome for cardiovascular diseases may be the presence or absence of the disease. Suppose we are interested in whether smoking is associated with cardiovascular diseases in a hypothetical group. We can construct a 2 x 2 crosstabulation table to examine the problem (Table 1).

TABLE 1
Frequencies of Smoking and Cardiovascular Diseases

		Cardiovascular Diseases		
		No(1)	Yes(2)	Total
Smoking	No (1)	2232 (A)	221 (B)	2453
	Yes (2)	321 (C)	64 (D)	385

The chance for cardiovascular diseases occurring in smokers compared to non-smokers is obtained by computing the cross product ratio:

$$\frac{2232 \times 64}{221 \times 321} = 2.01$$

which suggests that odds or risk for developing cardiovascular disease is twice as high among smokers than among non-smokers in the study population. This cross-product ratio is also called the odds ratio. If we conduct a chi-square analysis using the same data in Table 1, we get a chi-square of 21.36 and $p<.01$. The chi-square statistic only helps us decide whether smoking and cardiovascular disease are related or not. It does not tell us how much more likely cardiovascular disease would be expected to occur among smokers than among nonsmokers, which the odds ratio does. It seems obvious that the odds ratio is a more interpretable statistic than the chi-square. This is especially true with large samples where even a minuscule difference in proportion may result a statistically significant chi-square. The interpretation of such results under this condition can be difficult.

The odds ratio, a measure of association in a 2 x 2 crosstabulation context, has wide applications in epidemiology and health studies, as it approximates how much more likely it is for the outcome to be present among those exposed to a risk than among those not exposed to a risk. The odds ratio has several desirable

properties. First, the odds ratio has a clearer interpretation than that of other statistics such as the chi-square. Looking at the Table 1, A/B is the odds with no cardiovascular disease for the smokers, and C/D is the corresponding odds for the smokers. The odds ratio for the two odds is then:

$$\frac{A/B}{C/D} = \frac{A/D}{B/C} = \text{odds ratio} = 2.01$$

A second desired property is that the odds ratio can be applied to rnultidimensional crosstabulations (i.e., greater than 2 x 2 tables) either through a series of 2 x 2 partitionings or by examining several 2 x 2 subtables. Third, the odds ratio is invariant if the rows and columns (but not both) are interchanged, only the reference category changes. For example, if the first row and the second row are interchanged in Table 1, the crosstabulation will appear like that in Table 2.

TABLE 2
Frequencies of Smoking and Cardiovascular Diseases

		Cardiovascular Disease	
		No(1)	Yes(2)
Smoking	Yes (1)	321	64
	No (2)	2232	221

The odds ratio becomes: (321 x 221)/ (64 x 2232) = 0.50. In this case, we may say that the odds for cardiovascular disease to occur among nonsmokers is half that of smokers in the study population.

In addition, most statistical packages provide a 95% confidence interval of the odds ratio, which can further help researchers interpret the association of two variables. The 95% confidence interval indicates that if we were to repeat the same study, the odds ratio would fall in the range of the confidence interval in 95 out of 100 times. The confidence interval also indicates whether the odds ratio is significant or not, which we will describe later in the article.

The odds ratio can be obtained with the crosstabulation procedure using SPSS[1] or SAS.[2] Let's denote letter E as the risk exposure factor and the letter D as a disease factor. When it is assumed that both factors are binary measures, the SPSS command for obtaining the odds ratio and the 95% confidence interval is:

```
CROSSTABS  E BY D/STATISTIC = RISK.
```

The equivalent SAS command is:

```
PROC FREQ; TABLES E*D / CMH;
```

Readers familiar with the SPSS and the SAS may see that the subcommands RISK and CMH are the options for obtaining the acids ratios.

Even though the analysis of a 2 x 2 table is common, it is more than likely that a disease or a health problem may be associated with multiple factors in epidemiology and health studies. Perhaps the most frequently

adopted statistic in such condition is the logistic regression. It turned out that when performing logistic regression with more than one independent variable, the odds ratio is the preferred parameter of interest, which we will discuss next.

The Odds Ratio and the Logistic Model

Let us use an example to elaborate the odds ratio as a parameter in a logistic regression. Suppose we have one dependent variable (denoted by letter D: where 1= no disease, 2= disease) and two independent variables, which may be the exposure to risk factors (denoted by letters El and E2: where 1= exposure to risk, 2= nonexposure to the same risk). Data are presented as the following:

D	E1	E2
2	2	1
2	2	1
1	1	1
2	2	1
2	2	1
1	1	2
2	2	2
2	2	2
1	1	2
2	2	1
1	2	2
2	1	1
2	2	2
2	2	1
1	2	1
2	2	2
1	2	2
2	2	2
1	1	1
1	1	2

Our goal is to examine how the two independent variables may predict the dependent variable. Because of the dichotomous scales, the obvious choice is to apply the logistic regression.

The SPSS logistic regression[3] command is :

LOGISTIC REGRESSION D WITH El E2.

The SAS logistic regression[2] command is :

PROC LOGISTIC; MODEL D = El E2;

The logistic regression output from the SPSS package looks like Table 3. Note we have only presented the relevant statistics for our discussion. The complete out-put can be found in the SPSS manual.[3]

In Table 3, the coefficient of 3.1901 in the logistic regression represents the odds change in the dependent variable (D) for a change of one unit in the independent variable (El). This interpretation is similar to that of the linear regression coefficient. The only difference is that the change in the dependent variable is the change of log odds in the logistic regression.

The odds ratio in the logistic regression is defined as the ratio of the odds for El = 2 to the odds for El = 1 and can be obtained from the coefficient (for a dichotomous independent variable) as:

Exp(B)

where Exp is the base of the natural logarithms (approximately 2.7183), and B is the coefficient. The odds ratios for variable E1 and E2 are

For El: $Exp^{3.1901} = 24.2901$

For E2: $Exp^{-1.3541} = 0.2582$

The computation of the odds ratio is not necessary because the current version of SPSS provides this parameter under the heading Exp(B) (the last column in Table 3). The odds ratio of 24.2901 suggests that the odds for the disease to occur is approximately 24 times for those exposed to the risk (El) than for those not exposed to the risk. This fact concerning the interpretability of the coefficients by the odds ratios is the fundamental reason why logistic regression has proven to be a powerful analytic tool for epidemiological and health research.

From these two examples, we can see that when the coefficient is greater than zero, the odds ratio will be greater than 1, whereas when the coefficient is smaller than 0, the odds ratio will be smaller than 1 (see Table 3). In case the software used provides only the coefficient, but not the odds ratio for logistic regression, one may obtain the odds ratio with most calculators that has an Exp function.

It is worth mentioning that the logistic regression from the SAS output is slightly different from that of the SPSS output (see Table 4)

TABLE 3
Partial SPSS Logistic Regression Output

Variable	B	S.E.	df	Sig	Exp(B)
E1	3.1901	1.4113	1	0.0238	24.29
E2	-1.3541	1.2619	1	0.2833	0.2582

The coefficients of El and E2 are - 3.1901 and 1.3541, which have the reversed signs compared to the SPSS out-put. The reason for this is that the SPSS package adopts the smaller value of the independent variable (1 in this case) as the reference category whereas the SAS package adopts the opposite, with the greater value of the independent variable (2 in this case) as the reference category. Consequently, the Exp(3.1901) of El in SPSS estimates the odds of those ex-posed to the risk (2) relative to those unexposed to the risk (1), whereas in SAS, the Exp(-3.1901) of El estimates the odds of those not exposed to the risk (1) relative to those exposed to the risk (2). As a general practice, we seem to favor the estimate of the odds of an "exposed" group relative to that of an "unexposed" group. Therefore, the unexposed group is considered the reference category, a smaller value of a dichotomous code.

TABLE 4
Partial SAS Logistic Regression Output

Variable	B	S.E.	df	Sig	Odds Ratio
E1	-3.1901	1.4113	1	0.0238	0.041
E2	1.3541	1.2619	1	0.2833	3.873

Because of the importance of the odds ratio as a measure of association, 95% confidence interval estimates are often presented in conjunction with the odds ratio scores. Unfortunately, both the SPSS and the SAS currently do not provide such interval estimates. The estimates, how-ever, can be obtained without much difficulty by the following computations.

$$\text{Exp}(B \pm 1.96 \times \text{S.E.})$$

where S.E. is the standard error. For example, the 95% confidence interval for the odds ratio of 3.873 (E2) is:

$$\text{Exp}(1.3541 \pm 1.96 \times 1.2619) = (0.327, 45.94).$$

This confidence limit suggests that if we were to repeat the same study, 95 out of 100 times the odds ratio we would obtain would fall into the limit of 0.327- 45.94. If the 95% confidence interval does not include a value of 1, the odds ratio is significant. Otherwise, the odds ratio is not significant. Our obtained confidence interval (0.327, 45.94) includes 1; there-fore, the odds ratio of 3.873 is not significant.

Other Applications

The odds ratio can also be applied to a logistic regression with multi-category independent variables. For example, if a drinking variable has four categories—abstainer, light drinker, moderate drinker, and heavy drinker—we can create a design variable (also called the dummy variable) using the abstainer as the reference group. The specification of the design variables is presented in the following:

		Design Variables		
Drinking (code)		D1	D2	D3
(1)	Abstainer	0	0	0
(2)	Light	1	0	0
(3)	Moderate	0	1	0
(4)	Heavy	0	0	1

By entering all design variables into, the logistic regression model as the following:

PROC LOGISTIC; MODEL D = D1 D2 D3;

where the letter D is the dependent variables and D1, D2, and D3 are three design variables. The yielded coefficient and the odds ratio for D1 are parameters for light drinkers relative to abstainers; the coefficient and the odds ratio for D2 are parameters for moderate drinkers relative to abstainers; and the coefficient and the odds ratio for D3 are for heavy drinkers relative to abstainers. Researchers may select any drinking level as the reference group based on their interest of comparison. Using the design variables, the odds ratio and the logistic regression will have even wider applications to data analysis, whether the data are nominal or ordinal, dichotomous or polytomous.

CONCLUSIONS

For a dichotomous variable, the odds ratio is usually the parameter of interest in 2 x 2 crosstabulation and in a logistic regression due to its ease of interpretation. In logistic regression, the odds ratio can be obtained from the estimated logistic regression coefficient, regardless of how the variable is measured. We have not covered the topics concerning assessing the fit of a logistic model which interested readers may read Hosmer et al.'s helpful article.[4]

REFERENCES

1. SPSS base system User's Guide, Chicago: SPSS Inc. 1990.
2. SAS/STAT User's Guide. Version 6. Cary: SAS Institute Inc. 1990.
3. SPSS Advanced Statistics User's Guide, Chicago: SPSS Inc. 1990.
4. Hosmer DW, Taber S, Lemeshow S: The importance of assessing the fit of logistic regression models: A case study. Am J Public Health 1991; 81(12):1630-1635.