

Presenting Logistic Regression Models to Non-Statisticians: Adjusted Probabilities and Adjusted Risk Ratios

David J. Pasta, Ovation Research Group, Palo Alto, CA
Miriam G. Cisternas, MGC Data Services, Carlsbad, CA

ABSTRACT

Analysts are often required to present results from logistic regressions to non-statisticians. The standard practice of presenting logistic regression results using odds ratios can be a challenge for individuals with little statistical training, who tend to find their interpretation difficult. One alternative to using logistic regression and presenting the results as odds ratios is to change the statistical model to one that estimates probabilities or risk ratios directly. That approach can be less than satisfactory. We prefer to continue to use logistic regression, and we offer two alternatives to the presentation of odds ratios: adjusted probabilities and adjusted risk ratios, both of which are based on the logistic model and the dataset from which it is estimated. An adjusted probability represents the estimated probability of the event for the independent variable or subgroup in question after having adjusted for all other factors in the model. It is similar in spirit to the adjusted means provided by the LSMEANS statement in linear models procedures. The adjusted risk ratio represents the ratio of one adjusted probability to another. We discuss ways of obtaining confidence intervals around adjusted probabilities and adjusted risk ratios.

INTRODUCTION

The presentation and interpretation of the results of a linear model is generally fairly straightforward. The parameter estimates are easily interpreted as the increase in the predicted (outcome variable) for each unit increase in the predictor (explanatory variable). In SAS[®] procedures such as GLM or MIXED, one can use the LSMEANS statements to obtain the estimated (adjusted) means for each category of a CLASS variable in a convenient table. Those adjusted means are calculated as though the other explanatory variables were fixed at their means (for continuous variables) or equally distributed across categories (by default) or distributed across categories according to the marginal distribution observed in the data (with the OBSMARGIN or OM option specified).

There have been many advances in the statistical software for estimating linear models in recent years, most notably the development of PROC MIXED. Much more noteworthy, however, has been the extension of statistical modeling to more and more nonlinear modeling situations. Especially important is the generalized linear model, as popularized by Nelder

and Wedderburn (1972). In the generalized linear model, a nonlinear function of the response variable is assumed to be linearly related to a set of predictor variables. The choice of the so-called link function that specifies the relationship between the response variable and the predictors determines the name given to the analysis. When the logit function is used, defined as $\text{logit}(p)=\log(p/(1-p))$, the analysis is called logit regression or, more commonly, logistic regression. As another example, when the link function is the inverse of the cumulative normal distribution, the analysis is called a normit regression model or a probit regression model.

The natural way of presenting results from logistic regressions is with odds ratios. The odds of a result that happens with probability p is $p/(1-p)$. For an explanatory variable with two values, odds ratios arise in logistic regression as the ratio of the odds of having an event when the explanatory variable is "yes" to the odds of having the event when the explanatory variable is "no". When the variable has no explanatory power, the odds ratio is 1. There is no theoretical upper limit to the odds ratio; the lower limit is zero.

One can also calculate the odds ratio for other explanatory variables, continuous or ordinal, as the variable increases by one unit. Thus the odds ratio gives the (multiplicative) increase in the odds of the outcome for each unit increase in the predictor (explanatory) variable. The increase in odds from several predictors can be calculated by multiplying together the respective odd ratios. Equivalently, and sometimes more conveniently, one can work with the logarithm of the odds ratio. Under the logistic model, the log odds are linearly related to the predictors, so the coefficient in the logistic regression gives the increase in log-odds of the outcome for each unit increase in the predictor.

The presentation of logistic regression results as odds ratios or logarithms of odds ratios represents a substantial difficulty, as individuals with little or no statistical training tend to find odds ratios hard to understand. One "solution" to the difficulty of understanding odds ratios is to avoid them by changing the underlying statistical model. This popular approach is not without its difficulties, however. We prefer to keep the logistic regression model, but to present the results not only as odds

ratios but also as adjusted probabilities and adjusted risk ratios.

AVOIDING CALCULATION OF ODDS RATIOS

Some authors have chosen to change the underlying statistical model to avoid working with odds ratios. Instead of assuming that the log-odds of an event is linearly related to the explanatory variables, as with logistic regression, instead they assume (for example) that the log of the probability is linearly related to the explanatory values. This leads to the log-binomial model, which can be estimated by using PROC GENMOD and specifying the binomial distribution and the logarithmic link function. Among the authors advocating this approach are Skov et al. (1998) and Deddens et al. (2003). This literature tends to use the term “prevalence ratio” or “risk ratio” to refer to the ratio of predicted probabilities. We do not distinguish those two terms here, although there are certainly occasions where a careful distinction between prevalence and incidence is critical. Another estimation approach that has been used is to use the Cox proportional hazard model to estimate the prevalence ratio (Lee and Chia, 1993; Lee, 1994).

There are both practical and theoretical difficulties with these approaches. One practical difficulty with the log-binomial modeling using PROC GENMOD approach is that GENMOD frequently fails to find a maximum likelihood solution. In some cases this is due to the solution lying on a boundary of the parameter space, where the derivative is not necessarily zero at the solution. In other cases it appears to be related to numerical instabilities in the algorithm. One way to avoid the boundary-space problem would be to use software designed to find maximum likelihood estimates under linear constraints, but that is not as conveniently available as GENMOD. Another solution to that specific problem is given in Deddens et al. (2003), where a near-optimal solution is obtained in the interior of the parameter space that is near the correct boundary solution. This simple technique can mitigate the practical problem when the issue is a boundary solution, but it does not address other convergence problems that seem to arise in very large data sets nor does it address the theoretical problems. The Cox proportional hazard approach appears to have trouble when the solution is on the boundary and may have inflated standard errors (Deddens et al., 2003).

Regardless of the practical difficulties, there are serious theoretical difficulties with these approaches that estimate prevalence ratios directly. They make a different assumption about the relationship between the probabilities and the explanatory variables. At first

blush, this might appear to be purely a matter of taste. There are good reasons to prefer the log-odds model, however. One reason is that when modeling the prevalence directly with a log-linear function, the relationship between the explanatory variable and the outcome is necessarily concave up (because of the shape of the exponential function). With log-odds being treated as linear, the relationship can be concave up or concave down. But perhaps the most compelling reason to prefer log-odds is that it treats the event and the nonevent symmetrically, whereas there is an absence of symmetry in the outcome being modeled using the other approaches. When the outcome is a rare disease, or maybe any disease, it seems reasonable to model the probability of contracting the disease. But when the outcome is “Treatment A” versus “Treatment B” it may be less clear which probability to model – that of getting treatment A or that of getting treatment B. It is a distressing fact that for the approaches that model risk or prevalence directly it matters whether you predict the probability of getting treatment A or the probability of getting treatment B. For the logistic regression (log-odds) approach, interchanging the two gives an equivalent model — the same answer but with all the signs changed.

In our opinion, the lack of symmetry in the approaches that estimate the probabilities directly make them unsuitable for most applications of logistic regression, although there are circumstances where they may be reasonable.

WHAT ABOUT CONVERTING ODDS RATIOS TO RISK RATIOS?

A different approach to the problem of interpreting odds ratios is to act as though they are intended to estimate risk ratios and try to improve their accuracy as estimates. This is the approach of Zhang and Yu (1999). Throughout this short article, beginning with the title (“What’s relative risk? A method of correcting the odds ratio in cohort studies of common outcomes”) and evidenced in nearly every paragraph, the authors take the position that the odds ratio is supposed to be an estimate of the relative risk as measured by the risk ratio but not a very good one – one that needs “correcting”. In fact, the odds ratio from a logistic regression is designed to be an estimate of the population odds ratio, not the population risk ratio. It makes no more sense to speak of “correcting” the odds ratio to get a better estimate of risk ratio than to “correct” the risk ratio to get a better estimate of the odds ratio.

On the other hand, many non-statisticians are taught the shortcut, “you can think of the odds ratio as pretty much like a risk ratio as long as the (base) prevalence of the event being studied is low.” The rule of thumb is

that as long as the prevalence of the “disease” (event of interest) is below about 10% in the “nonexposed” (base) population, the estimated odds ratio is pretty close to the risk ratio.

The Zhang and Yu suggestion is to solve the “formula” for the risk ratio in terms of the odds ratio. This gives them what they insist on calling the “corrected” odds ratio as $OR / [(1 - P_0) + (P_0 \times OR)]$. This formula assumes that P_0 , the incidence of the outcome of interest in the nonexposed group, is known or can be estimated. They propose the creation of confidence bounds by applying the same formula to the confidence bounds of the odds ratio, which neglects the uncertainty in the estimate of P_0 (McNutt et al., 1999) but is perhaps justifiable as a simplification.

This approach at least allows the underlying model to remain in terms of log-odds being linearly related to the parameters. Is there a better way to make the results of logistic regression easier for non-statisticians to understand? Let us return to the linear model and adjusted means and work from that approach.

ADJUSTED MEANS

Analysts who work frequently with linear models using SAS often use the LSMEANS statement in PROC GLM or PROC MIXED to obtain “adjusted means.” Those adjusted means are convenient for reporting the results of general linear models. In earlier work (Pasta, Cisternas, and Williamson 1998; Pasta and Cisternas, 2003), we have discussed appropriate analogues for nonlinear and generalized linear models and how to estimate their standard errors. It is appropriate to review some of that material here.

The mathematics and statistics of linear regression, analysis of variance, and analysis of covariance are essentially identical; the primary differences have more to do with terminology than substance. In multiple linear regression you model a quantitative response variable as a linear combination of quantitative predictor variables. The model is linear in the sense that the relationship between the response variable and any predictor variable in the model is a straight line. It is permitted for the predictor variables to be nonlinearly related to other variables, which may or may not be included in the model. For example, the total cost of a hospital stay (the response variable) might be predicted by a linear model containing an intercept and three predictors: the length of stay, the square of the length of stay, and the cube of the length of stay. Although such a regression equation is often referred to as curvilinear, it is linear in the predictors and therefore can be considered a linear model.

In PROC GLM and PROC MIXED, the LSMEANS statement provides a convenient way to obtain least-

squares (adjusted) means for even complicated general linear models. When the OBSMARGINS option is not specified on LSMEANS, these adjusted means can be thought of as the mean value that would have been obtained if all the continuous variables were set at their means and all the qualitative (CLASS statement) variables were spread equally over each of the possible levels. This can lead to unreasonable adjusted means, as shown in Potter and Pasta (1997). Better is to specify OBSMARGINS on the LSMEANS statement, so that the adjusted means spread the qualitative variables across the levels in proportion to the observed distribution

For simple linear models without any interactions between a binary predictor variable and other predictors, the least-squares means for the two levels of the binary predictor variable with the OM option can be calculated from a model in either of two equivalent ways. One way is to set every variable except the predictor variable of interest at its mean value and calculate the predicted value under the model as though this observation had a value of 1 for the predictor variable, and then calculate the predicted value again as though this observation had a value of 0 for the predictor variable. This method will be referred to as the “predicted value of the mean” approach. The precise definition of “setting a variable at its mean value” for qualitative variables has to do with the details of the OM option; it may be easiest to think of creating a set of dummy variables and taking the mean of those variables.

The other way to calculate the least-squares means is called the “mean of the predicted values” approach. In this approach, two predicted values are computed for all of the observations in the input data set, once with the predictor variable coded with a value of 1 and once with the predictor variable coded with a value of 0. The means of those predicted values are the least-squares means for the “1” group and the “0” group, respectively.

The important thing to understand is that these two methods are equivalent for the linear model, but generally are not equivalent for nonlinear models. Either method extends to nonlinear models and to the case where there are interactions between the predictor of interest and other predictors. The second method, the “mean of the predicted values” method, is perhaps easier to extend. The predictor variable of interest and all other variables (including interactions) derived from that variable can be coded first as though all the observations have a “1” and then again as though all the observations have a “0”.

For the linear model, the two methods were equivalent. For the nonlinear and generalized linear models, which method is better? This is to some extent a matter of taste, but we believe that the

second method, the “mean of the predicted values” provides a better analogue to the least-squares means of linear models. It provides, in a very specific sense, a mean value for each group that reflects the observed values of the predictor variables and adjusts for any imbalance between the treatment groups. It also has the advantage of calculating predicted values for plausible values of the predictor variables, rather than for some mythical hybrid observation that is, for example, 42% male and 58% female and also 27% a person with no college, 41% a person with some college, and 32% a college graduate.

Once adjusted predicted values are calculated, it is easy to calculate the ratio of those values for different categories of a CLASS variable. These provide point estimates of adjusted risk ratios based on the logistic regression model.

CALCULATION OF STANDARD ERRORS

We would like to be able to estimate standard errors for adjusted means calculated by the recommended method, the “mean of the predicted values” approach. It turns out that the estimation of the standard error of adjusted means and differences between adjusted means is not difficult using PROC IML. The details are given in Pasta and Cisternas, 2003.

The estimation of the standard error of the ratio of two adjusted means is more problematic. One approach is to use the classical Fieller approach to confidence intervals for ratios of random variables. Another approach is to use the bootstrap. Neither method gives an exact theoretical answer, but both can give approximate answers that are very useful in practice.

REFERENCES

- Deddens, James A., Petersen, Martin R., Lei, Xiudong (2003), “Estimation of prevalence ratios when PROC GENMOD does not converge,” Proceedings of SUGI 28, Paper 270-28.
- Lee J. (1994), “Odds ratio or relative risk for cross-sectional data?”, International Journal of Epidemiology, 23:201-3.
- Lee J, and Chia KS (1993), “Estimation of prevalence rate ratios for cross-sectional data: an example in occupational epidemiology,” British Journal of Industrial Medicine, 50:861-2.
- McNutt L., Hafner J., and Xue X. (1999), “Correcting the odds ratio in cohort studies of common outcomes (Letter)”, Journal of the American Medical Association, 282:529.
- Nelder, J.A. and Wedderburn, R.W.M. (1972), “Generalized Linear Model,” Journal of the Royal Statistical Society, Series A, 135, 761-8.

Pasta, David J. and Cisternas, Miriam G. (2003), “Estimating Standard Errors for CLASS Variables in Generalized Linear Models Using PROC IML,” Proceedings of SUGI 28, Paper 264-28.

Pasta, David J., Cisternas, Miriam G., and Williamson, Cynthia L. (1998), “Estimating Standard Errors of Treatment Effects for Probit Models and for Linear Models of Log-Transformed Variables using PROC IML,” Proceedings of the 6th Annual Western Users of SAS Software Regional Users Group Conference, 211-216.

Potter, Lori and Pasta, David J. (1997), “The Sums of Squares Are All the Same – How Can the LSMEANS Be So Different?”, Proceedings of the 5th Annual Western Users of SAS Software Regional Users Group Conference, 187-92.

Skov T., Deddens J., Petersen M., and Endahl L. (1998), “Prevalence proportion ratios: estimation and hypothesis testing,” International Journal of Epidemiology, 27:91-5.

Zhang J. and Yu K. (1998), “What’s relative risk? A method of correcting the odds ratio in cohort studies of common outcomes,” Journal of the American Medical Association, 280:1690-1.

ACKNOWLEDGMENTS

SAS and SAS/STAT software are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

CONTACT INFORMATION

David J. Pasta
Senior Director, Data Management and Analysis
Ovation Research Group
2970 South Court
Palo Alto, CA 94306
(650) 213-9106 (phone)
(650) 213-9125 (fax)
dpasta@ovation.org

Miriam G. Cisternas, Partner
MGC Data Services
5051 Millay Court
Carlsbad, CA 92008
(760) 804-5746 (phone)
miriam@mgcdata.com
www.mgcdata.com