

**Effect Size and Confidence Intervals in General Linear Models
for Categorical Data Analysis**

Randall E. Schumacker
University of North Texas
(940) 565-3962
rschumacker@unt.edu

Robert Mount
Executive Director
Dallas ISD
(972) 749-5982
mount@dallasisd.org

Paper presented at the American Educational Research Association annual meeting
Montreal, Canada
April 15, 2005
Marriott Montreal Chateau Champlain, Salle de Bal Ballroom & Foyer
3:05pm – 3:45pm

ABSTRACTEffect Size and Confidence Intervals in General Linear Models
for Categorical Data Analysis

In general linear models for categorical data analysis, goodness-of-fit statistics only provide a broad significance test of whether the model fits the sample data. Hypothesis testing has traditionally reported the chi-square or G^2 likelihood ratio (deviance) statistic and associated p-value when testing the significance of a model or comparing alternative models. The effect size (log odds ratio) and confidence interval (ASE) need to receive more attention when interpreting categorical response data using the logistic regression model. This trend is supported by recent efforts in general linear models for continuous data (t-test, analysis of variance, least squares regression) that have criticized the sole use of statistical significance testing and the $p < .05$ criteria for a Type I error rate.

Effect Size and Confidence Intervals in General Linear Models for Categorical Data Analysis

The American Psychological Association has recently advocated that hypothesis testing go beyond statistical significance testing at $p < .05$ for Type I error rate (Wilkinson, L., & APA Task Force on Statistical Inference, 1999). The traditional statistical significance testing has placed an emphasis upon the probability of a statistical value occurring beyond a chance level given the sampling distribution of the statistic (Harlow, Mulaik, and Steiger, 1997). Recently, more emphasis has been placed on the practical interpretation of results that include effect size, confidence interval, and confidence intervals around effect sizes, however the discussion centered on statistical applications that use continuous data (Kirk, 1996). The present author(s) will highlight applications that use the general linear model for categorical data analysis (DeMaris, 1992; Fox, 1997). The logistic regression goodness-of-fit criteria for categorical data analysis will be presented (Klienbaum, 1994). Our presentation will go beyond the statistical test of significance and highlight the important role that effect size (odds ratio, log odds ratio, relative risk or probability ratio) and confidence interval (asymptotic standard error; ASE) have in the general linear model for categorical data analysis.

Categorical data analysis techniques are used when subject responses are binary and mutually exclusive. The typical method of analyzing relationships amongst categorical variables is to use the chi-square statistic or phi correlation coefficient (Upton, 1978). The general linear model for categorical response variables however has become more widely used in the behavioral sciences because many research questions

involve a categorical dependent variable and one or more categorical independent variables.

Logistic regression is a special case of log-linear regression where both the dependent and independent variables are categorical in nature (Hosmer, & Lemeshow, 1989; Klienbaum, 1994). It offers distinct advantages over the chi-square method for analysis of categorical variables. In logit models, natural log odds of the frequencies are computed that allow different models and different model parameters to be compared given the additive nature of the G^2 component for each model. If a non-significant likelihood-ratio chi-square (G^2) value is computed, then a given model fits the observed data.

Goodness-of-fit Criteria

A theoretical logit regression model is generally postulated (null model or base model). A common practice is then to create alternative models where each new model contains parameters of the previous model, plus a hypothesized new parameter. The theoretical model can be tested beginning with a null model and adding parameters, or with a saturated model deleting parameters. Several logit regression models may “fit” equally well based on various goodness-of-fit criteria that are used to determine whether the model fits the data in the logit regression model. The goodness-of-fit criteria typically reported are:

1. Pearson chi-square
2. Likelihood-ratio chi-square (G^2)
3. Predictive efficacy (R-squared type measure)
4. Deviance ($-2 [L_M - L_S]$)

Pearson chi-square is calculated as: $\chi^2 = \sum (O - E)^2 / E$. The chi-square distribution is defined by degrees of freedom, df. The mean of the chi-square distribution is equal to df with the standard deviation equal to $\sqrt{2df}$. As the degrees of freedom, df, increases the chi-square sampling distribution goes from a right skewed distribution to a normal distribution.

The likelihood-ratio chi-square (G^2) is based on the ratio of maximum likelihood values, Λ , and expressed in logarithm form as $-2 \log(\Lambda)$. The G^2 statistic can also be expressed as: $G^2 = 2 \sum O_{ij} \log(O_{ij} / E_{ij})$ where O is the observed cell frequency, E is the expected cell frequency, and the I and J subscripts represent the individual cells in the cross-tabulated table. The log transformation yields an approximate chi-squared sampling distribution with a minimum value of zero and larger values suggesting rejection of the null hypothesis. The p-value simply indicates the strength of evidence against the null hypothesis.

Predictive efficacy refers to whether a model generates accurate predictions of group membership on the dependent variable. It is possible to have an excellent fit between the logit model and the data without having predictive efficacy. Recall, if $G^2 = 0$, a saturated model exists which perfectly fits the data, yet predictive efficacy can be far from perfect. The R^2 type measure for logistic regression is not meant as a variance accounted for interpretation, as traditionally noted in least squares regression, because it under estimates the proportion of variance explained in the categorical variables. Instead, the R^2 type measure is an approximation for assessing predictive efficacy ranging from zero (0) [independence model] to one (1) [saturated model].

The deviance value provides a way to examine differences in nested logistic regression models. The G^2 from one model is simply subtracted from the G^2 of the second model. This is similar to testing a full versus restricted model in multiple regression. The deviance value is $-2[L_m - L_s]$ where L represents the respective log-likelihood function of each model with the degrees of freedom equal to the difference in the degrees of freedom of the two models. The deviance is the likelihood ratio statistic (G^2) for comparing model M to the saturated model S . Since the saturated model has $G^2 = 0$, this reduces to the G^2 statistic for the hypothesized logistic regression model. If G^2 is non-significant, then additional independent categorical predictors in the model are not needed. This type of test is only appropriate for the likelihood-ratio chi-square and not the Pearson chi-square because adding additional independent categorical predictor variables will never result in a poorer fit of the model to the data.

Effect Size and Confidence Interval Criteria

Effect size measures and the asymptotic standard error (ASE) play a major role in interpreting the practical significance of estimated parameters in general linear models for categorical variables. The parameter estimates in logistic regression are calculated using maximum likelihood estimation and possess asymptotic properties. As sample size increases, the parameter estimates become unbiased and consistent with population parameters. The sampling distribution also approaches normality with variance lower than other unbiased estimation procedures.

The effect size measures typically used in categorical data analysis are:

1. z test
2. odds ratio
3. log odds ratio
4. relative risk or probability ratio

The z test, given larger samples, can be used to test a parameter's significance and compute a confidence interval. The formula for z is: $z = B / ASE$. The confidence interval is computed as: $z \pm 1.96 * \sigma$; where $\sigma = [p(1-p)/n]^{1/2}$. The significance test simply indicates whether an estimated parameter is reasonable whereas the confidence interval yields a range of possible values for the parameter, given sampling error.

Odds ratios are computed as: $Odds = p / 1 - p$. If the probability of success is .8, the probability of failure is .2, and the odds ratio is $.8 / .2 = 4$. This indicates 4 successes for every one failure. Unfortunately, odds ratios in small to moderate samples have skewed sampling distributions and therefore are not widely used.

The log odds ratio or natural logarithm of the odds ratio, $\log(\theta)$, is preferred for interpreting an effect size. Independence of categorical variables is equivalent to $\log(\theta) = 0$, i.e. odds ratio = 1 is equal to \log odds ratio = 0. The sampling distribution of the log odds ratio approximates a normal distribution as sample size increases with a mean of $\log(\theta)$ and standard deviation ASE. Parameter estimates in logit models can be readily interpreted as a log-odds ratio. This is calculated as e^{β} for a single parameter, or $e^{\beta_1 - \beta_2}$ for differences between two parameters. This is useful when examining contrasts between levels of two independent categorical predictor variables.

The relative risk or probability ratio should be interpreted separately from the odds ratio (Cohen, 2000). The relative risk (RR) indicates a probability and is computed

as probability p_1 divided by probability p_2 [$RR = p_1 / p_2$]. In contrast the odds ratio (OR) is $(p_1 / 1 - p_1)$ divided by $(p_2 / 1 - p_2)$. The odds ratio is therefore related, but different from relative risk ($OR = [PR - p_1] / 1 - p_1$) or $RR \times [1 - p_2 / 1 - p_1]$). For logistic model interpretation, a gender coefficient (male = 0 and female = 1) of $e^{1.67}$ would indicate the odds of females over males participating, whereas the statement females were two-thirds more likely than males to participate is a relative risk or probability statement.

The asymptotic standard error (ASE) or standard deviation of the log transform sampling distribution is computed as: $ASE(\log \pi) = [1/n_1 + \dots + 1/n_k]^{1/2}$. A 95% confidence interval around the log odds ratio is then computed as $\log(\pi) \pm 1.96 ASE[\log(\pi)]$. The confidence interval should contain the value 1.0 otherwise the true odds will be different for the two groups being compared. The confidence interval also provides valuable information about the range of minimum and maximum log odd ratios.

METHODS

The logistic regression model ($\log(\pi) = \alpha + \beta_1 X_1 + \dots + \beta_k X_k$) was applied to a set of categorical data (Stokes, Davis, & Koch, 1995). The goodness-of-fit criteria, effect size, confidence interval, and confidence interval around the effect size are reported. The importance of effect size and confidence interval reporting above and beyond significance testing is then discussed.

Data Analysis

An example data set relating myocardial infarction and aspirin use is provided as follows

(Agresti, 1996):

Group	Yes	No	Total
Placebo	189	10,845	11,034
Aspirin	104	10,933	11,037

The proportion, p_1 , or placebo odds ratio is $189 / 11,034 = .0171$ and indicates that .0171 percent suffered myocardial infarction while taking a placebo. In contrast, proportion, p_2 , or aspirin odds ratio is $104 / 11,037$ and indicates that .0094 suffered myocardial infarction while taking aspirin. The percent difference is .0077 with standard error of .0015. $z = .0077 / .0015 = 5.133$, which is statistically significant. The 95% confidence interval for this true difference is $.0077 + /- 1.96(.0015)$ or $(.005, .011)$, so taking aspirin appears to diminish the risk of myocardial infarction. The relative risk is .0171 divided by .0094 or 1.82. Using relative risk, the proportion of MI cases was 82% higher for the group taking the placebo. The 95% confidence interval is $(1.43, 2.30)$, thus we can be 95% confident that the proportion of MI cases for the group taking the placebo was at least 43% higher than the group taking the aspirin. The relative risk indicates that the difference isn't trivial and may have important health implications.

The natural log odds ratio is $\log(1.83) = .605$. The ASE ($\log \pi$) is computed as $[1/189 + 1/10,845 + 1/104 + 1/10,933]^{1/2} = .123$. The 95% confidence interval for $\log(\pi)$ is $(.365, .846)$. The corresponding confidence interval for π is $(1.43, 2.30)$. Since it does not contain 1.0, the true odds of myocardial infarction appear to be different for the two groups.

RESULTS AND INTERPRETATION

The categorical data example indicates a statistically significant z-test of the difference between the proportion of myocardial infarction cases for the placebo and aspirin usage groups. The effect size (odds ratio, log odds ratio, and relative risk or probability ratio) provides a more practical interpretation of the efficacy of using aspirin to thwart myocardial infarction in patients. Moreover, the confidence interval and especially the confidence interval around the effect size (log odds ratio) provided important additional information to our interpretation of results.

Statistical significance testing has come under attack by scholars in recent years because it is influenced by a researcher's choice of sample size, power, and Type I error rate. The reported research literature however has focused on continuous data analysis techniques and not fully included categorical data analysis methods. The American Psychological Association and Editors of several popular journals are now requiring educational researchers to report effect size and confidence intervals. The use and interpretation of effect size and confidence interval in categorical data analysis is therefore also important to understand and report.

REFERENCES

- Agresti, A. (1996). *An Introduction to Categorical Data Analysis*.
NY: John Wiley & Sons, Inc.
- Cohen, M.P. (2000). Note on the Odds Ratio and the Probability Ratio. *Journal of Educational and Behavioral Statistics*, 25(2), 249-252.
- DeMaris, A. (1992). *Logit modeling: Practical Applications*. Sage University Paper series on Quantitative Applications in the Social Sciences, no. 07-086.
Newbury Park, CA: Sage.
- Fox, J. (1997). *Applied Regression Analysis, Linear Models, and Related Methods*.
Newbury Park, CA: Sage.
- Harlow, L.L., Mulaik, S.A., & Steiger, J.H. (editors.) (1997).
What if there were no significance tests?
NJ: Lawrence Erlbaum Associates, Inc.
- Hosmer, D.W. & S. Lemeshow (1989). *Applied Logistic Regression*.
NY: John Wiley & Sons, Inc.
- Kirk, R. (1996). Practical significance: A concept whose time has come.
Educational and Psychological Measurement, 56, 746-759.
- Kleinbaum, D.G. (1994). *Logistic Regression*.
NY: Springer-Verlag.
- Stokes, M.E., C.S. Davis, & G.G. Koch (1995). *Categorical Data Analysis Using the SAS System*. Cary, NC: SAS Institute, Inc.

Upton, G.J.G. (1978). *The analysis of cross-tabulated data*.

NY: John Wiley & Sons, Inc.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations.

American Psychologist, 54, 594-604.