Logistic Regression

Overview

Binomial (or binary) logistic regression is a form of regression which is used when the dependent is a dichotomy and the independents are of any type. *Multinomial logistic regression* exists to handle the case of dependents with more classes than two. When multiple classes of the dependent variable can be ranked, then *ordinal logistic regression* is preferred to multinomial logistic regression. Continuous variables are not used as dependents in logistic regression. Unlike logit regression, there can be only one dependent variable.

Logistic reqression can be used to predict a dependent variable on the basis of continuous and/or categorical independents and to determine the percent of variance in the dependent variable explained by the independents; to rank the relative importance of independents; to assess interaction effects; and to understand the impact of covariate control variables.

Logistic regression applies maximum likelihood estimation after transforming the dependent into a logit variable (the natural log of the odds of the dependent occurring or not). In this way, logistic regression estimates the probability of a certain event occurring. Note that logistic regression calculates changes in the log odds of the dependent, not changes in the dependent itself as OLS regression does.

Logistic regression has many analogies to OLS regression: logit coefficients correspond to b coefficients in the logistic regression equation, the standardized logit coefficients correspond to beta weights, and a pseudo R^2 statistic is available to summarize the strength of the relationship. Unlike OLS regression, however, logistic regression does not assume linearity of relationship between the independent variables and the dependent, does not require normally distributed variables, does not assume homoscedasticity, and in general has less stringent requirements. It does, however, require that observations are independent and that the independent variables be linearly related to the logit of the dependent. The success of the logistic regression can be assessed by looking at the classification table, showing correct and incorrect classifications of the dichotomous, ordinal, or polytomous dependent. Also, goodness-of-fit tests such as model chi-square are available as indicators of model appropriateness as is the Wald statistic to test the significance of individual independent variables. .

In SPSS, binomial logistic regression is under Analyze - Regression - Binary Logistic, and the multinomial version is under Analyze - Regression - Multinomial Logistic. Logit regression, discussed separately, is another related option in SPSS for using loglinear methods to analyze one or more dependents. Where both are applicable, logit regression has numerically equivalent results to logistic regression, but with different output options. For the same class of problems, logistic regression has become more popular among social scientists.

Key Terms and Concepts

- **Design variables** are nominal or ordinal independents (factors) entered as dummy variables. In SPSS binomial logistic regression, categorical independent variables must be declared by clicking on the "Categorical" button in the Logistic Regression dialog box. SPSS multinomial logistic regression will convert categorical variables to dummies automatically by leaving out the last category, which becomes the reference category. Researchers may prefer to create dummy variables manually so as to control which category is omitted and thus becomes the reference category. For more on the selection of dummy variables, click here.

- **Covariates** are interval independents in most programs. However, in SPSS dialog all independents are entered as covariates, then one clicks the Categorical button in the Logistic Regression dialog to declare any of those entered as categorical.

- **Interaction terms**. As in OLS regression, one can add interaction terms to the model (ex., age*income). For continuous covariates, one simply creates a new variable which is the product of two existing ones. For categorical variables, one also multiplies but you have to multiply two category codes as shown in the Categorical Variables Codings table of SPSS output (ex, race(1)*religion(2)). The codes will be 1's and 0's so most of the products in the new variables will be 0's unless you recode some products (ex., setting 0*0 to 1).

- **Maximum likelihood estimation, MLE**, is the method used to calculate the logit coefficients. This contrasts to the use of ordinary least squares (OLS) estimation of coefficients in regression. OLS seeks to minimize the sum of squared distances of the data points to the regression line. MLE seeks to maximize the log likelihood, LL, which reflects how likely it is (the odds) that the observed values of the dependent may be predicted from the observed values of the independents.

  MLE is an iterative algorithm which starts with an initial arbitrary "guesstimate" of what the logit coefficients should be, the MLE algorithm determines the direction and size change in the logit coefficients which will increase LL. After this initial function is estimated, the residuals are tested and a re-estimate is made with an improved function, and the process is repeated (usually about a half-dozen times) until *convergence* is reached (that is, until LL does not change significantly). There are several alternative convergence criteria.

- **Ordinal** and **multinomial logistic regression** are extensions of binary logistic regression that allow the simultaneous comparison of more than one contrast. That is, the log odds of three or more contrasts are estimated simultaneously (ex., the probability of A vs. B, A vs.C, B vs.C., etc.).

- **SPSS and SAS**. In SPSS, select Analyze, Regression, Binary (or Multinomial), Logistic; select the dependent and the covariates; Continue; OK. SAS's PROC CATMOD computes both simple and multinomial logistic regression, whereas PROC LOGIST is for simple (dichotomous) logistic regression. CATMOD uses a conventional model command: ex., model wsat*supsat*qman=_response_ /nogls ml ;. Note that in the model command, nogls suppresses generalized least squares estimation and ml specifies maximum likelihood estimation.

- **Significance Tests**

  - **Log likelihood**: A "likelihood" is a probability, specifically the probability that the observed values of the dependent may be predicted from the observed values of the independents. Like any probability, the likelihood varies from 0 to 1. The log likelihood (LL) is its log and varies from 0 to minus infinity (it is negative because the log of any number less than 1 is negative). LL is calculated through *iteration*, using maximum likelihood estimation (MLE). Log likelihood is the basis for tests of a logistic model.

    - **The likelihood ratio** is a function of log likelihood. Because -2LL has approximately a chi-square distribution, -2LL can be used for assessing the significance of logistic regression, analogous to the use of the sum of squared errors in OLS regression. The -2LL statistic is the likelihood ratio. It is also called goodness of fit, deviance chi-square, scaled deviance, deviation chi-square, DM, or L-square. It reflects the significance of the unexplained variance in the dependent. In SPSS output, this statistic is found in the "-2 Log Likelihood" column of the "Iteration History" table or the "Likelihood Ratio Tests" table. The likelihood ratio is not used directly in significance testing, but it it the basis for the likelihood ratio test, which is the test of the difference between two likelihood ratios (two -2LL's), as discussed below. In general, as the model becomes better, -2LL will decrease in magnitude.

    - **The likelihood ratio test**, also called the log-likelihood test, is based on -2LL (deviance). The likelihood ratio test is a test of the significance of the difference between the likelihood ratio (-2LL) for the researcher's model minus the likelihood ratio for a reduced model. This difference is called "model chi-square." The likelihood ratio test is generally preferred over its alternative, the Wald test, discussed below. There are three main forms of the likelihood ratio test::

      1. *Test of the overall model*. When the reduced model is the baseline model with the constant only, the likelihood ratio test tests the significance of the researcher's model as a whole. A well-fitting model is significant at the .05 level or better, meaning the researcher's model is significantly different from the one with the constant only. The likelihood ratio test appears in the "Model Fitting Information" table in SPSS output.

         Thus the likelihood ratio test of a model tests the difference between -2LL for the full model and -2LL for *initial chi-square* in the null model. This is called the the *model chi-square test*. The null model, also called the initial model, is logit(p) = the constant. That is, initial chi-square is -2LL for the model which accepts the null hypothesis that all the b coefficients are 0. This implies that that none of the independents are linearly related to the log odds of the dependent. Model chi-square thus tests the null hypothesis that all population logistic regression coefficients except the constant are zero. It is an overall model test which does not assure that every independent is significant. Warning: If the log-likelihood test statistic shows a small p value (<=.05) for a model

with a large effect size, ignore contrary findings based on the Wald statistic discussed below as it is biased toward Type II errors in such instances - instead assume good model fit overall.

Degrees of freedom in this test equal the number of terms in the model minus 1 (for the constant). This is the same as the difference in the number of terms between the two models, since the null model has only one term. Model chi-square measures the improvement in fit that the explanatory variables make compared to the null model. Model chi-square is a likelihood ratio test which reflects the difference between error not knowing the independents (initial chi-square) and error when the independents are included in the model (deviance). When probability (model chi-square) <= .05, we reject the null hypothesis that knowing the independents makes no difference in predicting the dependent in logistic regression.

2. *Testing for interaction effects.* A common use of the likelihood ratio test is to test the difference between a full model and a reduced model dropping an interaction effect. If model chi-square (which is -2LL for the full model minus -2LL for the reduced model) is significant, then the interaction effect is contributing significantly to the full model and should be retained.

3. *Test of individual model parameters.* The likelihood ratio test assesses the overall logistic model but does not tell us if particular independents are more important than others. This can be done, however, by comparing the difference in -2LL for the overall model with a nested model which drops one of the independents. We can use the likelihood ratio test to drop one variable from the model to create a nested reduced model. In this situation, the likelihood ratio test tests if the logistic regression coefficient for the dropped variable can be treated as 0, thereby justifying dropping the variable from the model. A nonsignificant likelihood ratio test indicates no difference between the full and the reduced models, hence justifying dropping the given variable so as to have a more parsimonious model that works just as well. Note that the likelihood ratio test of individual parameters is a better criterion than the alternative Wald statistic when considering which variables to drop from the logistic regression model. In SPSS output, the "Likelihood Ratio Tests" table contains the likelihood ratio tests of individual model parameters.

4. *Tests for model refinement.* In general, the likelihood ratio test can be used to test the difference between a given model and any nested model which is a subset of the given model. It cannot be used to compare two non-nested models. Chi-square is the difference in likelihood ratios (-2LL) for the two models, and degrees of freedom is the difference in degrees of freedom for the two models. If the computed chi-square is equal or greater than the critical value of chi-square (in a chi-square table) for the given df, then the models are significantly different. If the difference is significant, then the researcher concludes that the variables

dropped in the nested model do matter significantly in predicting the dependent. If the difference is below the critical value, there is a finding of non-significance and the researcher concludes that dropping the variables makes no difference in prediction and for reasons of parsimony the variables are dropped from the model. That is, chi-square difference can be used to help decide which variables to drop from or add to the model. This is discussed further in the next section.

- **Chi-square (Hosmer-Lemeshow) test of goodness of fit**. If chi-square goodness of fit is not significant, then the model has adequate fit. By the same token, if the test is significant, the model does not adequately fit the data. This test appears in SPSS multinomial logistic regression output in the "Goodness of Fit" table, with a Pearson chi-square and a deviance (likelihood ratio) chi-square version (both usually close). One must check "Goodness of fit" under the Statistics button. This test is preferred over classification tables when assessing model fit.

  - *Hosmer and Lemeshow's goodness of fit test*, not to be confused with a similarly named, obsolete goodness of fit test discussed below, is another name for a chi-square goodness of fit test. It is available under the Options button in the SPSS binary logistic regression dialog. The test divides subjects into deciles based on predicted probabilities, then computes a chi-square from observed and expected frequencies. Then a probability (p) value is computed from the chi-square distribution with 8 degrees of freedom to test the fit of the logistic model. If the H-L goodness-of-fit test statistic is greater than .05, as we want for well-fitting models, we fail to reject the null hypothesis that there is no difference between observed and model-predicted values, implying that the model's estimates fit the data at an acceptable level. That is, well-fitting models show nonsignificance on the H-L goodness-of-fit test, indicating model prediction is not significantly different from observed values. This does not mean that the model necessarily explains much of the variance in the dependent, only that however much or little it does explain is significant. As the sample size gets large, the H-L statistic can find smaller and smaller differences between observed and model-predicted values to be significant. On the other hand, the H-L statistic assumes sampling adequacy, with a rule of thumb being enough cases so that no group has an expected value < 1 and 95% of cells (typically, 10 decile groups times 2 outcome categories = 20 cells) have an expected frequency > 5. Collapsing groups may not solve a sampling adequacy problem since when the number of groups is small, the H-L test will be biased toward nonsignificance (will overestimate model fit).

- **More about likelihood ratio tests of chi-square difference between nested models**. *Block chi-square* is a synonym for the likelihood ratio (chi-square difference) test, referring to the change in -2LL due to entering a block of variables. The main Logistic Regression dialog allows the researcher to enter independent variables in blocks. Blocks may contain one or more variables. There are three major uses of the likelihood ratio test with nested models:

- *Stepwise logistic regression*: The forward or backward stepwise logistic regression method utilizes the likelihood ratio test (chi-square difference) to determine automatically which variables to add or drop from the model. This brute-force method runs the risk of modeling noise in the data and is considered useful only for exploratory purposes. Selecting model variables on a theoretic basis and using the "Enter" method is preferred. However, problems of overfitting the model to noise in the current data may be mitigated by cross-validation, fitting the model to one a test subset of the data and validating the model using a hold-out validation subset. Note that *step chi-square* is the likelihood ratio test, which tests the change in -2LL between steps. Earlier versions of SPSS referred to this as "improvement chi-square." Stepwise procedures are selected in the Method drop-down list of the logistic regression dialog, in turn giving the following choices:

  - *Forward selection vs. backward elimination*: Forward selection is the usual option, starting with the constant-only model and adding variables one at a time in the order they are best by some criterion (see below) until some cutoff level is reached (ex., until the step at which all variables not in the model have a significance higher than .05). Backward selection starts with all variables and deletes one at a time, in the order they are worst by some criterion.

  - *Variable entry criterion for forward selection.* Rao's efficient score statistic (see below) is used as the forward selection criterion for adding variables to the model. It is similar but not identical to a likelihood ratio test of the coefficient for an individual explanatory variable. It appears in the "score" column of the "Variables not in the equation" table of SPSS output.

    - **Score statistic**: Rao's efficient score, labeled simply "score" in SPSS output, is test for whether the logistic regression coefficient for a given explanatory variable is zero. It is mainly used as the criterion for variable inclusion in forward stepwise logistic regression (discussed above), because of its advantage of being a non-iterative and therefore computationally fast method of testing individual parameters compared to the likelihood ratio test. In essence, the score statistic is similar to the first iteration of the likelihood ratio method, where LR typically goes on to three or four more iterations to refine its estimate. In addition to testing the significance of each variable, the score procedure generates an "Overall statistics" significance test for the model as a whole. A finding of nonsignificance (ex., p>.05) on the score statistic leads to acceptance of the null hypothesis that coefficients are zero and the variable may be dropped. SPSS continues by this method until no remaining predictor variables have a score statistic significance of .05 or better.

- *Variable removal criteria for backward elimination.* For eliminating variables in backward elimination, the researcher may choose from among the likelihood ratio test (the preferred method), the Wald statistic, or the conditional statistic (a computationally faster approximation to the likelihood ratio test and preferred when use of the likelihood ratio criterion proves computationally too time consuming). The likelihood ratio test computes -2LL for the current model, then reestimates -2LL with the target variable removed. The conditional statistic is similar except that the -2LL for the target variable removed model is a one-pass estimate rather than an interative reestimation as in the likelihood ratio test. The conditional statistic is considered not as accurate as the likelihood ratio test but more so than the third possible criterion, the Wald test.

- *Which step is the best model?* Stepwise methods do not necessarily identify "best models" at all as they work by fitting an automated model to the current dataset, raising the danger of overfitting to noise in the particular dataset at hand. However, there are three possible methods of selecting the "final model" that emerges from the stepwise procedure:

  1. *Last step.* The final model is the last step model, where adding another variable would not improve the model significantly.

  2. *Lowest AIC.* The "Step Summary" table will print the Akaike Information Criterion (AIC) for each step. AIC is commonly used to compare models, where the lower the AIC, the better. The step with the lowest AIC thus becomes the "final model."

  3. *Lowest BIC.* The "Step Summary" table will print the Bayesian Information Criterion (BIC) for each step. BIC is also used to compare models, again where the lower the BIC, the better. The step with the lowest BIC thus becomes the "final model." Often BIC will point to a more parsimonious model than will AIC as its formula factors in degrees of freedom, which is related to number of variables.

- Click here for further discussion of stepwise methods.

- *Sequential logistic regression* is analysis of nested models where the researcher is testing the control effects of a set of covariates. The logistic regression model is run against the dependent for the full model with independents and covariates, then is run again with the block of independents dropped. If chi-square difference is not significant, then the researcher concludes that the independent variables are controlled by the covariates (that is, they have no effect once the effect of the covariates is

taken into account). Alternatively, the nested model may be just the independents, with the covariates dropped. In that case a finding of non-significance implies that the covariates have no control effect.

- *Assessing dummy variables*: Running a full model and then a model with all the variables in a dummy set dropped (ex., East, West, North for the variable Region) allows assessment of dummy variables, using chi-square difference. Note that even though SPSS computes log-likelihood ratio tests of individual parameters for each level of a dummy variable, the log-likelihood ratio tests of individual parameters (discussed below) should not be used, but rather use the likelihood ratio test (chi-square difference method) for the set of dummy variables pertaining to a given variable. Because all dummy variables associated with the categorical variable are entered as a block this is sometimes called the "block chi-square" test and its value is considered more reliable than the Wald test, which can be misleading for large effects in finite samples.

- **Wald statistic (test):** The Wald statistic is an alternative test which is commonly used to test the significance of individual logistic regression coefficients for each independent variable (that is, to test the null hypothesis in logistic regression that a particular logit (effect) coefficient is zero). For dichotomous independents, the Wald statistic is the squared ratio of the unstandardized logit coefficient to its standard error. The Wald statistic and its corresponding p probability level is part of SPSS output in the "Variables in the Equation" table. This corresponds to significance testing of b coefficients in OLS regression. The researcher may well want to drop independents from the model when their effect is not significant by the Wald statistic. The Wald test appears in the "Parameter Estimates" table in SPSS logistic regression output.

  Warning: Menard (p. 39) warns that for large logit coefficients, standard error is inflated, lowering the Wald statistic and leading to Type II errors (false negatives: thinking the effect is not significant when it is). That is, there is a flaw in the Wald statistic such that very large effects may lead to large standard errors and small Wald chi-square values. For models with large logit coefficients or when dummy variables are involved, it is better to test the difference using the likelihood ratio test of the difference of models with and without the parameter. Also note that the Wald statistic is sensitive to violations of the large-sample assumption of logistic regression. Put another way, the likelihood ratio test is considered more reliable for small samples (Agresti, 1996). For these reasons, the likelihood ratio test of individual model parameters is generally preferred.

- **Logistic coefficients and correlation.** Note that a logistic coefficient may be found to be significant when the corresponding correlation is found to be not significant, and vice versa. To make certain global statements about the significance of an independent variable, both the correlation and the parameter estimate (b) should be significant. Among the reasons why correlations and logistic coefficients may differ in significance are these: (1) logistic coefficients are partial coefficients, controlling for other variables in the model, whereas

correlation coefficients are uncontrolled; (2) logistic coefficients reflect linear and nonlinear relationships, whereas correlation reflects only linear relationships; and (3) a significant parameter estimate (b) means there is a relation of the independent variable to the dependent variable for selected control groups, but not necessarily overall.

- **Confidence interval for the logistic regression coefficient**. The confidence interval around the logistic regression coefficient is plus or minus 1.96*ASE, where ASE is the asymptotic standard error of logistic b. "Asymptotic" in ASE means the smallest possible value for the standard error when the data fit the model. It is also the highest possible precision. The real (enlarged) standard error is typically slightly larger than ASE. One typically uses real SE if one hypothesizes that noise in the data are systematic and one uses ASE if one hypothesizes that noise in the data are random. As the latter is typical, ASE is used here.

- **Goodness of Fit (obsolete)**, also known as *Hosmer and Lemeshow's Goodness of Fit Index* or *C-hat,* is an alternative to model chi-square for assessing the significance of a logistic regression model. Menard (p. 21) notes it may be better when the number of combinations of values of the independents is approximately equal to the number of cases under analysis. This measure was included in SPSS output as "Goodness of Fit" prior to Release 10. However, it was removed from the reformatted output for SPSS Release 10 because, as noted by David Nichols, senior statistician for SPSS, it "is done on individual cases and does not follow a known distribution under the null hypothesis that the data were generated by the fitted model, so it's not of any real use" (SPSSX-L listserv message, 3 Dec. 1999).

- **Interpreting Parameter Estimates**

  - **Logit coefficients (logits)**, also called unstandardized *logistic regression coefficients* or *effect coefficients* or simply "parameter estimates" in SPSS output, correspond to b coefficients in OLS regression. Logit coefficients, which are on the right-hand side of the logistic equation, are not to be confused with logits, which is the term on the left-hand side. Both logit and regression coefficients can be used to construct prediction equations and generate predicted values, which in logistic regression are called logistic scores. The SPSS table which lists the b coefficiences also lists the standard error of b, the Wald statistic and its significance (discussed below), and the odds ratio (labeled Exp(b)) as well as confidence limits on the odds ratio.

    Probabilities, odds, and odds ratios are all important basic terms in logistic regression. See the more extensive coverage in the separate section on log-linear analysis.

  - **Parameter estimates and logits.** In SPSS and most statistical output for logistic regression, the "parameter estimate" is the b coefficient used to predict the log odds (logit) of the dependent variable. Let z be the logit for a dependent variable, then the logistic prediction equation is:

$z = \ln(\text{odds(event)}) = \ln(\text{prob(event)}/\text{prob(nonevent)}) = \ln(\text{prob(event)}/[1 - \text{prob(event)}])$
$= b_0 + b_1X_1 + b_2X_2 + ... + b_kX_k$
where b_0 is the constant and there are k independent (X) variables. Some of the X variables may in fact be interaction terms.

For a one-independent model, z would equal the constant, plus the b coefficient times the value of X_1, when predicting odds(event) for persons with a particular value of X_1. If X_1 is a binary (0,1) variable, then $z = X_0$ (that is, the constant) for the "0" group on X_1 and equals the constant plus the b coefficient for the "1" group. To convert the log odds (which is z, which is the logit) back into an odds ratio, the natural logarithmic base e is raised to the zth power: odds(event) = exp(z) = odds the binary dependent is 1 rather than 0. If X_1 is a continuous variable, then z equals the constant plus the b coefficient times the value of X_1. For models with additional constants, z is the constant plus the crossproducts of the b coefficients times the values of the X (independent) variables. Exp(z) is the log odds of the dependent, or the estimate of odds(event).

To summarize, logits are the log odds of the event occurring (usually, that the dependent = 1 rather than 0). The "z" in the logistic formula above is the logit. Odds(event) = Exp(z). Where OLS regression has an identity link function, logistic regression has a logit link function (that is, logistic regression calculates changes in the log odds of the dependent, not changes in the dependent itself as OLS regression does). Parameter estimates (b coefficients) associated with explanatory variables are estimators of the change in the logit caused by a unit change in the independent. In SPSS output, the parameter estimates appear in the "B" column of the "Variables in the Equation" table. Logits do not appear but must be estimated using the logistic regression equation above, inserting appropriate values for the constant and X variable(s). The b coefficients vary between plus and minus infinity, with 0 indicating the given explanatory variable does not affect the logit (that is, makes no difference in the probability of the dependent value equaling the value of the event, usually 1); positive or negative b coefficients indicate the explanatory variable increases or decreases the logit of the dependent. Exp(b) is the odds ratio for the explanatory variable, discussed below. Note that when b=0, Exp(b)=1, so therefore an odds ratio of 1 corresponds to an explanatory variable which does not affect the dependent variable.

- **Odds ratios**. In contrast to exp(z), exp(b) is the odds ratio. The odds ratio is the natural log base, e, to the exponent, b, where b = the parameter estimate. "Exp(B)" in SPSS output refers to odds ratios. Exp(b), which is the odds ratio for a given independent variable, represents the factor by which the odds(event) change for a one-unit change in the independent variable. Put another way, Exp(b) is the ratio of odds for two groups where each group has a values of X_j which are one unit apart from the values of X_j in the other group. A positive Exp(b) means the independent variable increases the logit and therefore increases odds(event). If Exp(b) = 1.0, the independent variable has no effect. If Exp(b) is less than 1.0, then the independent variable decreases the logit and decreases odds(event). For instance, if $b_1 = 2.303$, then the corresponding odds ratio (the exponential function, e^b) is 10, then we may

say that when the independent variable increases one unit, the odds that the dependent = 1 increase by a factor of 10, when other variables are controlled. In SPSS, odds ratios appear as "Exp(B)" in the "Variables in the Equation" table.

odds ratio = exp(b)
b = ln(odds ratio)

A second simple example: Some 20 people take a performance test, where 0=fail and 1=success. For males, 3 fail and 7 succeed. For females, 7 fail and 3 succeed. Then p(success) for males = 7/10 = .70; and q(failure) for males = 3/10 = .30. Likewise p(success) for females = 3/10 = .30, and q(failure) for females = 7/20 = .70. Therefore the *odds* of success for males is the ratio of the probabilities = .7/.3 = 2.3333. The odds of success for females = .3/.7 = .4286, rounded off. Then the *odds ratio* for success (for performance = 1) for males:females is 2.3333/.4286 = 5.4444. Since the parameter estimate is the natural log of the odds ratio, therefore b(gender) = ln(5.4444) = 1.6946. Conversely, if the b for gender was 1.6946 we could convert it to an odds ratio using the function exp(1.6946) = 5.4444. And we would say that the odds of success (the odds that the dependent variable performance = 1) are 5.4444 times as large for males as for females. (If you try this on your calculator results will not be exact due to rounding error: there are actually more than the four decimal places shown above, and also delta must be set to 0).

- *Comparing the change in odds for different values of X.* The odds ratio, which is Exp(b), is the factor by which odds(event) changes for a 1 unit change in X. But what if years_education was the X variable and one wanted to know the change factor for X=12 years vs. X=16 years? Here, the X difference is 4 units. The change factor is not Exp(b)*4. Rather, odds(event) changes by a factor of $Exp(b)^4$. That is, odds(event) changes by a factor of Exp(b) raised to the power of the number of units change in X.

- *Comparing the change in odds when interaction terms are in the model.* In general, Exp(b) is the odds ratio and represents the factor by which odds(event) is multiplied for a unit increase in the X variable. However, the effect of the X variable is not properly gauged in this manner if X is also involved in interaction effects which are also in the model. Before exponentiating, the b coefficient must be adjusted to include the interaction b terms. Let X_1 be years_education and let X_2 be a dichotomous variable called "school_type" coded 0=private school, 1=public school, and let the interaction term $X_1 * X_2$ also be in the model. Let the b coefficients be .864 for X_1, .280 for X_2, and .010 for $X_1 * X_2$. The adjusted b, which we shall label b*, for years education = .864 + .010*school_type. For private schools, b* =.864. For public schools b* = .874. Exp(b*) is then the estimate of the odds ratio, which will be different for different values of the variable (here, school_type) with which the X variable (here, years_education) is interacting.

- *Logit coefficients vs. logits.* Although b coefficients are called logit or logistic coefficients, they are not logits. That is, b is the parameter estimate and z is the logit. Parameter estimates (b) have to do with changes in the logit, z, while the logit has to do with estimates of odds(event). To avoid confusion, many

researchers refer to b as the "parameter estimate."

- *Logit coefficients, and why they are preferred over odds ratios in modeling.* Note that for the case of decrease the odds ratio can vary only from 0 to .999, while for the case of increase it can vary from 1.001 to infinity. This asymmetry is a drawback to using the odds ratio as a measure of strength of relationship. Odds ratios are preferred for interpretation, but logit coefficients are preferred in the actual mathematics of logistic models. *Warning:* The odds ratio is a different way of presenting the same information as the unstandardized logit (effect) coefficient discussed in the section on logistic regression, and like it, is not recommended when comparing the relative strengths of the independents. The standardized logit (effect) coefficient is used for this purpose.

- *Parameter estimates in multinomial logistic regression.* In multinomial logistic analysis, where the dependent may have more than the usual 0-or-1 values, the comparison is always with the last value rather than with the value of 1. The parameter estimates table for multinomial logistic regression will contain factor or covariate parameters for each category of the categorical dependent variable except the last category (by default - however SPSS multinomial lets the researcher set the reference category as the first or other custom category). If the predictor is a covariate, there will be a single set of parameters for each value of the categorical dependent except the reference category. If the predictor is a factor, there will be one parameter row for each of that predictor's categories except the reference category. If a parameter estimate (b coefficient) is significant and positive, then that parameter increases the odds of the given response (category) of the dependent (response) variable compared to the reference category response. If negative, that parameter decreases to odds of that response compared to the reference category response. Additional light can be thrown on the predictive power of the logits by requesting cell probabilities under the Statistics button, giving actual and predicted counts and percentages for each combination of categories of the dependent and predictor variables.

  For example, let "candidate" be a categorical dependent variable with three levels: the first parameter estimate will be the log of the odds (probability candidate=1: probability candidate=3), and the second parameter estimate will the the log odds of (p(candidate=2):p(candidate=3)). Let the explanatory variable be gender, with 0=female and 1=male, such that the reference category is 1=male. Let the reference category of the dependent equal 3, the default. There will be two parameter estimates for gender: one for candidate 1 and one for candidate 2. Let the parameter estimate for gender=0 for candidate 1 be .500. Then the odds ratio is exp(.500) = 1.649. We can then say the odds of a female selecting candidate 1 compared to candidate 3 is 1.649 times (about 65% greater than) the odds a male would. Warning: This is a statement about odds - do not directly transform it into a statement about probabilities/likelihood/chances.

- *Effect size.*. The odds ratio is a measure of effect size. The ratio of odds ratios of the independents is the ratio of relative importance of the independent variables in terms of effect on the dependent variable's odds. (Note standardized

logit coefficients may also be used, as discussed below, but then one is discussing relative importance of the independent variables in terms of effect on the dependent variable's log odds, which is less intuitive.).

- *Confidence interval on the odds ratio*. SPSS labels the odds ratio "Exp(B)" and prints "Low" and "High" confidence levels for it. If the low-high range contains the value 1.0, then being in that variable value category makes no difference on the odds of the dependent, compared to being in the reference (usually highest) value for that variable. That is, when the 95% confidence interval around the odds ratio includes the value of 1.0, indicating that a change in value of the independent variable is not associated in change in the odds of the dependent variable assuming a given value, then that variable is not considered a useful predictor in the logistic model.

- *Types of variables*. Parameter estimates (b) and odds ratios (Exp(b)) may be output for dichotomies, categorical variables, or continuous variables. Their interpretation is similar in each case, though often researchers will not interpret odds ratios in terms of statements illustrated below, but instead will simply use odds ratios as effect size measures and comment on their relative sizes when comparing independent variable effects, or will comment on the change in the odds ratio for a particular explanatory variable between a model and some nested alternative model.

    - *Dichotomies*. If b is positive, then as the dichotomous independent variable moves from 0 to 1, the log odds (logit) of the dependent also increases. If the odds ratio is 4.3 for hs_degree (1=having a high school degree, 0 = not), for instance, where the dependent is employed (0=not employed, 1=employed), we say that the odds of a person with a high school degree being employed are 4.3 times the odds of a person without a high school degree. In multinomial logistic regression, the odds are those that the dependent=the highest category rather than dependent=1 as in binary logistic regression.

    - *Categorical variables*: Categorical variables must be interpreted in terms of the left-out reference category, as in OLS regression. If b is positive, then when the dummy = 1 (that category of the categrorical variable is present), the log odds (logit) of the dependent also increase. Thus the parameter estimate for a categorical dummy variable refers to the change in log odds when the dummy=1, compared to the reference category equaling 1 (being present). If the odds ratio is 4.3 for religion=2 (Catholic), where the reference category is "Agnostics," and the dependent is 1=attend religious movies and 0=don't attend, then the odds a Catholic attends religious movies is 4.3 times the odds that an agnostic does. *Warning*: Note that dichotomous variables may be entered as categorical dummies rather than as simple variables (which would be the norm). If if entered as categorical variables, then their odds ratios will be computed differently and must be interpreted comparative to the reference category rather than as simple increase/decrease in odds ratio.

- *Continuous covariates*: When the parameter estimate, b, is transformed into an odds ratio, it may be expressed as a percent increase in odds. For instance, consider the example of number of publications of professors (see Allison, 1999: 188). Let the logit coefficient for "number of articles published" be +.0737, where the dependent variable is "being promoted". The odds ratio which corresponds to +.0737 is approximately 1.08 (e to the .0737 power). Therefore one may say, "each additional article published increases the odds of promotion by about 8%, controlling for other variables in the model.." (Obviously, this is the same as saying the original dependent odds increases by 108%, or noting that one multiplies the original dependent odds by 1.08. By the same token, it is not the same as saying that the *probability* of promotion increases by 8%.) To take another example, let income be a continuous explanatory variable measured in ten thousands of dollars, with a parameter estimate of 1.5 in a model predicting home ownership=1, no home ownership=0. A 1 unit increase in income (one $10,000 unit) is then associated with a 1.5 increase in the log odds of home ownership. However, it is more intuitive to convert to an odds ratio: exp(1.5) = 4.48, allowing one to say that a unit ($10,000) change in income increases the odds of the event ownership=1 about 4.5 times.

- **Probability interpretations**. While logistic coefficients are usually interpreted as odds, not probabilities, it is possible to use probabilities. Exp(z) is odds(event). Therefore the quantity (1 - Exp(z)) is odds(nonevent). Therefore P(event) = Exp(z)/(1 - Exp(z)). Recall z = the constant plus the sum of crossproducts of the b coefficients times the values of their respective X (independent) variables. For dichotomous independents assuming the values (0,1), the crossproduct term is null when X = 0 and is b when X=1. For continuous independents, different probabilities will be computed depending on the value of X. That is, P(event) varies depending on the value of X.

- **Measures of Model Fit**

  - **The Akaike Information Criterion, AIC,** is a common information theory statistic used when comparing alternative models. It is output by SAS's PROC LOGISTIC. Lower is better model fit.

  - **The Schwartz Information Criterion, SIC** is a modified version of AIC and is part of SAS's PROC LOGISTIC output. Compared to AIC, SIC penalizes overparameterization more (rewards model parsimony). Lower is better model fit. It is common to use both AIC and SIC when assessing alternative logistic models.

  - **Model chi-square**. Model chi-square is based on log likelihoods and as discussed above, model chi-square should be significant in a well-fitting model. Significance means that the fit of the researcher's full model differs significantly from that of the constant-only null model.

  - **Pearson Goodness of Fit, GOF**, also called Peason chi-square, is printed by SPSS logistic output below model chi-square. This alternative test should be non-significant

for a well-fitting model. Non-significance corresponds to failing to reject the null hypothesis that the observed likelihood does not differ from 1.

- **Measures of Effect Size**

  - **R-squared**. There is no widely-accepted direct analog to OLS regression's R^2. This is because an R^2 measure seeks to make a statement about the "percent of variance explained," but the variance of a dichotomous or categorical dependent variable depends on the frequency distribution of that variable. For a dichotomous dependent variable, for instance, variance is at a maximum for a 50-50 split and the more lopsided the split, the lower the variance. This means that R-squared measures for logistic regressions with differing marginal distributions of their respective dependent variables cannot be compared directly, and comparison of logistic R-squared measures with R^2 from OLS regression is also problematic. Nonetheless, a number of logistic R-squared measures have been proposed, all of which should be reported as approximations to OLS R^2, not as actual percent of variance explained.

    Note that R^2-like measures below are not goodness-of-fit tests but rather attempt to measure strength of association. For small samples, for instance, an R^2-like measure might be high when goodness of fit was unacceptable by the likelihood ratio test.

    1. **Cox and Snell's R-Square** is an attempt to imitate the interpretation of multiple R-Square based on the likelihood, but its maximum can be (and usually is) less than 1.0, making it difficult to interpret. It is part of SPSS output in the "Model Summary" table.

    2. **Nagelkerke's R-Square** is a further modification of the Cox and Snell coefficient to assure that it can vary from 0 to 1. That is, Nagelkerke's R^2 divides Cox and Snell's R^2 by its maximum in order to achieve a measure that ranges from 0 to 1. Therefore Nagelkerke's R^2 will normally be higher than the Cox and Snell measure but will tend to run lower than the corresponding OLS R^2. Nagelkerke's R^2 is part of SPSS output in the "Model Summary" table and is the most-reported of the R-squared estimates. See Nagelkerke (1991).

    3. **Pseudo-R-Square** is a Aldrich and Nelson's coefficient which serves as an analog to the squared contingency coefficient, with an interpretation like R-square. Its maximum is less than 1. It may be used in either dichotomous or multinomial logistic regression.

    4. **Hagle and Mitchell's Pseudo-R-Square** is an adjustment to Aldrich and Nelson's Pseudo R-Square and generally gives higher values which compensate for the tendency of the latter to underestimate model strength.

    5. **R-square** is OLS R-square, which can be used in binary logistic regression (see Menard, p. 23) but not in multinomial logistic regression. To obtain R-square, save the predicted values from logistic regression and run a bivariate regression on the observed dependent values. Note that logistic regression can yield

deceptively high R^2 values when you have many variables relative to the number of cases, keeping in mind that the number of variables includes k-1 dummy variables for every categorical independent variable having k categories.

- **Classification tables** are the 2 x 2 tables in the logistic regression output for dichotomous dependents, or the 2 x n tables for ordinal and polytomous logistic regression, which tally correct and incorrect estimates. The columns are the two predicted values of the dependent, while the rows are the two observed (actual) values of the dependent. In a perfect model, all cases will be on the diagonal and the overall percent correct will be 100%. If the logistic model has homoscedasticity (not a logistic regression assumption), the percent correct will be approximately the same for both rows. Since this takes the form of a crosstabulation,. measures of association (SPSS uses lambda-p and tau-p) may be used in addition to percent correct as a way of summarizing the strength of the table.

  - *Warning.* Classification tables should not be used as goodness-of-fit measures because they ignore actual predicted probabilities and instead use dichotomized predictions based on a cutoff (ex., .5). For instance, in binary logistic regression, predicting a 0-or-1 dependent, the classification table does not reveal how close to 1.0 the correct predictions were nor how close to 0.0 the errors were. A model in which the predictions, correct or not, were mostly close to the .50 cutoff does not have as good a fit as a model where the predicted scores cluster either near 1.0 or 0.0. Also, because the hit rate can vary markedly by sample for the same logistic model, use of the classification table to compare across samples is not recommended.

  - *Split of the dependent variable.* While no particular split of the dependent variable is assumed, the split makes a difference in the classification table. Suppose the dependent is split 99:1. Then one could guess the value of the dependent correctly 99% of the time just by always selecting the more common value. The classification table will likely show 0 predictions in the predicted column for the 1% value of the dependent. The closer to 50:50, the easier it is for a predictor variable to have an effect. Even at some intermediate but lopsided split, such as 85:15, it can be difficult for a predictor to improve on simple guessing (that is, on 85%). A strong predictor variable could improve on the 85% but a weak one might not. This does not mean the predictor variables are non-significant, just that they do not move the estimates enough to make a difference compared to pure guessing. When the classification table for a dichotomous dependent has a zero "Predicted" column, it is likely that the raw correlations of the predictor variables with the dependent variable are not high enough to make a difference.

  - *Terms* associated with classification tables:

    1. *Hit rate:* Number of correct predictions divided by sample size. The hit rate for the model should be compared to the hit rate for the classification table for the constant-only model (Block 0 in SPSS output). The Block 0 rate will be the percentage in the most numerous category (that is, the null

model predicts the most numerous category for all cases).

2. *Sensitivity:* Percent of correct predictions in the reference category of the dependent (ex., 1 for binary logistic regression).

3. *Specificity:* Percent of correct predictions in the given category of the dependent (ex., 0 for binary logistic regression).

4. *False positive rate:* In binary logistic regression, the number of errors where the dependent is predicted to be 1, but is in fact 0, as a percent of total cases which are observed 0's. In multinomial logistic regression, the number of errors where the predicted value of the dependent is higher than the observed value, as a percent of all cases on or above the diagonal.

5. *False negative rate:* In binary logistic regression, the number of errors where the dependent is predicted to be 0, but is in fact 1, as a percent of total cases which are observed 1's. In multinomial logistic regression, the number of errors where the predicted value of the dependent is lower than the observed value, as a percent of all cases on or below the diagonal.

- **The histogram of predicted probabilities**, also called the "classplot" or the "plot of observed groups and predicted probabilities," is part of SPSS output when one chooses "Classification plots" under the Options button in the Logistic Regression dialog. It is an alternative way of assessing correct and incorrect predictions under logistic regression. The X axis is the predicted probability from 0.0 to 1.0 of the dependent being classified "1". The Y axis is frequency: the number of cases classified. Inside the plot are columns of observed 1's and 0's (or equivalent symbols). Thus a column with one "1" and five "0's" set at p = .25 would mean that six cases were predicted to be "1's" with a probability of .25, and thus were classified as "0's." Of these, five actually were "0's" but one (an error) was a "1" on the dependent variable.

  The researcher looks for two things: (1) A U-shaped rather than normal distribution is desirable. A U-shaped distribution indicates the predictions are well-differentiated. A normal distribution indicates many predictions are close to the cut point, which is not as good a model fit.; and (2) There should be few errors. The 1's to the left are false positives. The 0's to the right are false negatives. Examining this plot will also tell such things as how well the model classifies difficult cases (ones near p = .5).

- **Unstandardized logit coefficients** are shown in the "B" column of the "Variables in the Equation" table in SPSS binomial logistic regression output. There will be a B coeffiicient for each independent and for the constant. The logistic regression model is log odds(dependent variable) = (B for var1)*Var1 + (B for var2)*Var2 + ... + (B for varn)*Varn + (B for the constant*Constant.

- **Standardized logit coefficients**, also called *standardized effect coefficients* or *beta weights,* correspond to beta (standardized regression) coefficients and like them may be used to compare the relative strength of the independents. However, odds ratios are

preferred for this purpose, since when using standardized logit coefficients one is discussing relative importance of the independent variables in terms of effect on the dependent variable's logged odds, which is less intuitive than relative to the actual odds of the dependent variable, which is the referent when odds ratios are used. SPSS does not output standardized logit coefficients but note that if one standardizes one's input data first, then the parameter estimates will be standardized logit coefficients. (SPSS does output odds ratios, which are found in the "Exp(B)" column of the "Variables in the Equation" table in binomial regression output.) Alternatively, one may multiply the unstandardized logit coefficients times the standard deviations of the corresponding variables, giving a result which is not the standardized logit coefficient but can be used to rank the relative importance of the independent variables. Note: Menard (p. 48) warned that as of 1995, SAS's "standardized estimate" coefficients were really only partially standardized. Different authors have proposed different algorithms for "standardization," and these result in different values, though generally the same conclusions about the relative importance of the independent variables.

- **Odds ratios**, discussed above, are also used as effect size measures.

- **Partial contribution, R**. Partial R is an alternative method of assessing the relative importance of the independent variables, similar to standardized partial regression coefficients (beta weights) in OLS regression. R is a function of the Wald statistic, Do (discussed below), and the number of degrees of freedom for the variable. SPSS prints R in the "Variables in the Equation" section. Note, however, that there is a flaw in the Wald statistic such that very large effects may lead to large standard errors, small Wald chi-square values, and small or zero partial R's. For this reason it is better to use odds ratios for comparing the importance of independent variables.

- **BIC**, the Bayes Information Criterion, has been proposed by Raftery (1995) as a third way of assessing the independent variables in a logistic regression equation. BIC in the context of logistic regression (and different from its use in SEM) should be greater than 0 to support retaining the variable in the model. As a rule of thumb, BIC of 0-2 is weak, 2 - 6 is moderate, 6 - 10 is strong, and over 10 is very strong.

1. **Lambda-p** is a PRE (proportional reduction in error) measure, which is the ratio of (errors without the model - errors with the model) to errors without the model. If lambda-p is .80, then using the logistic regression model will reduce our errors in classifying the dependent by 80% compared to classifying the dependent by always guessing a case is to be classed the same as the most frequent category of the dichotomous dependent. Lambda-p is an adjustment to classic lambda to assure that the coefficient will be positive when the model helps and negative when, as is possible, the model actually leads to worse predictions than simple guessing based on the most frequent class. Lambda-p varies from 1 to (1 - N), where N is the number of cases. Lambda-p = (f - e)/f, where f is the smallest row frequncy (smallest row marginal in the classification table) and e is the number of errors (the 1,0 and 0,1 cells in the classification table).

2. **Tau-p** is an alternative measure of association. When the classification table has equal marginal distributions, tau-p varies from -1 to +1, but otherwise may

be less than 1. Negative values mean the logistic model does worse than expected by chance. Tau-p can be lower than lambda-p because it penalizes proportional reduction in error for non-random distribution of errors (that is, it wants an equal number of errors in each of the error quadrants in the table.)

3. **Phi-p** is a third alternative discussed by Menard (pp. 29-30) but is not part of SPSS output. Phi-p varies from -1 to +1 for tables with equal marginal distributions.

4. **Binomial d** is a significance test for any of these measures of association, though in each case the number of "errors" is defined differently (see Menard, pp. 30-31).

5. **Separation:** Note that when the independents completely predict the dependent, the error quadrants in the classification table will contain 0's, which is called *complete separation*. When this is nearly the case, as when the error quadrants have only one case, this is called *quasicomplete separation*. When separation occurs, one will get very large logit coefficients with very high standard errors. While separation may indicate powerful and valid prediction, often it is a sign of a problem with the independents, such as definitional overlap between the indicators for the independent and dependent variables.

- The **c statistic** is a measure of the discriminative power of the logistic equation. It varies from .5 (the model's predictions are no better than chance) to 1.0 (the model always assigns higher probabilities to correct cases than to incorrect cases for any pair involving dependent=0 and dependent=1). Thus c is the percent of all possible pairs of cases in which the model assigns a higher probability to a correct case than to an incorrect case. The c statistic is not part of SPSS logistic output but may be calculated using the COMPUTE facility, as described in the SPSS manual's chapter on logistic regression. Alternatively, save the predicted probabilities and then get the area under the ROC curve. In SPSS, select Analyze, Regression, Binary (or Multinomial); select the dependent and covariates; click Save; check to save predicted values (pre_1); Continue; OK. Then select Graphs, ROC Curve; set pre_1 as the test variable; select standard error and confidence interval; OK. In the output, c is labeled as "Area." It will vary from .5 to 1.0.

- **Contrast Analysis**

  - **Repeated contrasts** is an SPSS option (called *profile contrasts* in SAS) which computes the logit coefficient for each category of the independent (except the "reference" category, which is the last one by default). Contrasts are used when one has a categorical independent variable and wants to understand the effects of various levels of that variable. Specifically, a "contrast" is a set of coefficients that sum to 0 over the levels of the independent categorical variable. SPSS automatically creates K-1 internal dummy variables when a covariate is declared to be categorical with K values (by default, SPSS leaves out the last category, making it the reference category). The user can choose various ways of assigning values to these internal variables, including *indicator contrasts*, *deviation contrasts*, or *simple contrasts*. In SPSS, indicator contrasts are now the default (old versions used deviation

contrasts as default).

- *Indicator contrasts* produce estimates comparing each other group to the reference group. David Nichols, senior statistician at SPSS, gives this example of indicator coding output:

```
Parameter codings for indicator contrasts
-------------------------------------------------
                          Parameter
             Value   Freq Coding
                            (1)    (2)
GROUP
                 1    106 1.000   .000
                 2    116  .000  1.000
                 3    107  .000   .000
-------------------------------------------------
```

  This example shows a three-level categorical independent (labeled GROUP), with category values of 1, 2, and 3. The predictor here is called simply GROUP. It takes on the values 1-3, with frequencies listed in the "Freq" column. The two "Coding" columns are the internal values (parameter codings) assigned by SPSS under indicator coding. There are two columns of codings because two dummy variables are created for the three-level variable GROUP. For the first variable, which is Coding (1), cases with a value of 1 for GROUP get a 1, while all other cases get a 0. For the second, cases with a 2 for GROUP get a 1, with all other cases getting a 0.

- *Simple contrasts* compare each group to a reference category (like indicator contrasts). The contrasts estimated for simple contrasts are the same as for indicator contrasts, but the intercept for simple contrasts is an unweighted average of all levels rather than the value for the reference group. That is, with one categorical independent in the model, simple contrast coding means that the intercept is the log odds of a response for an unweighted average over the categories.

- *Deviation contrasts* compare each group other than the excluded group to the unweighted average of all groups. The value for the omitted group is then equal to the negative of the sum of the parameter estimates.

- *Contrasts and ordinality:* For nominal variables, the pattern of contrast coefficients for a given independent should be random and nonsystematic, indicating the nonlinear, nonmonotonic pattern characteristic of a true nominal variable. Contrasts can thus be used as a method of empirically differentiating categorical independents into nominal and ordinal classes.

- **Analysis of residuals** Residuals may be plotted to detect outliers visually. Residual analysis may lead to development of separate models for different types of cases. For logistic regression, it is usual to use the standardized difference between the observed and expected probabilities. SPSS calls this the "standardized residual (ZResid)," while SAS calls this the "chi residual," while Menard (1995) and at other times (including by SPSS in the table of "Observed and Predicted Frequencies" in multinomial logistic output) it is called the "Pearson residual." In a model which fits in every cell formed by the independents, no absolute standardized residual will be > 1.96. Cells which do not meet this criterion signal combinations of independent variables for which the model is not working well.

Note there are other less-used types of residuals in logistic regression: logit residuals, deviance residuals, Studentized residuals, and of course unstandardized (raw) residuals: see Menard, p. 72.

The Save button in the SPSS logistic dialog will save the standardized residual as ZRE_1. One can also save predictions as PRE_1. The DfBeta statistic can be saved as DFB0_1 for the constant, DFB1_1 for the first independent, DFB1_2 for the second independent, etc.

- The **dbeta statistic, DBeta,** is available to indicate cases which are poorly fitted by the model. Called **DfBeta** in SPSS (whose algorithm approximates dbeta), it measures the change in the logit coefficients for a given variable when a case is dropped. There is a DfBeta statistic for each case for each explanatory variable and for the constant. An arbitrary cutoff criterion for cases to be considered outliers is those with dbeta > 1.0 on critical variables in the model.

- The **leverage statistic, h,** is available to identify cases which influence the logistic regression model more than others. The leverage statistic varies from 0 (no influence on the model) to 1 (completely determines the model). The leverage of any given case may be compared to the average leverage, which equals p/n, where p = (k+1)/n, where k= the number of independents and n=the sample size. Note that influential cases may nonetheless have small leverage values when predicted probabilities are <.1 or >.9. Leverage is an option in SPSS, in which a plot of leverage by case id will quickly identify cases with unusual impact.

- **Cook's distance**, D, is a third measure of the influence of a case. Its value is a function of the case's leverage and of the magnitude of its standardized residual. It is a measure of how much deleting a given case affects residuals for all cases. An approximation to Cook's distance is an option in SPSS logistic regression.

Assumptions

- Logistic regression is popular in part because it enables the researcher to overcome many of the restrictive assumptions of OLS regression:

1. Logistic regression does not assume a linear relationship between the dependents and the independents. It may handle nonlinear effects even when exponential and polynomial terms are not explicitly added as additional independents because the logit link function on the left-hand side of the logistic regression equation is non-linear. However, it is also possible and permitted to add explicit interaction and power terms as variables on the right-hand side of the logistic equation, as in OLS regression.
2. The dependent variable need not be normally distributed (but does assume its distribution is within the range of the exponential family of distributions, such as normal, Poisson, binomial, gamma).
3. The dependent variable need not be homoscedastic for each level of the independents; that is, there is no homogeneity of variance assumption: variances need not be the same within categories.

4. Normally distributed error terms are not assumed.
5. Logistic regression does not require that the independents be interval.
6. Logistic regression does not require that the independents be unbounded.

- However, other assumptions still apply:

1. **Meaningful coding**. Logistic coefficients will be difficult to interpret if not coded meaningfully. The convention for binomial logistic regression is to code the dependent class of greatest interest as 1 and the other class as 0, and to code its expected correlates also as +1 to assure positive correlation. For multinomial logistic regression, the class of greatest interest should be the last class. Logistic regression is predicting the log odds of being in the class of greatest interest.

2. **Inclusion of all relevant variables in the regression model**: If relevant variables are omitted, the common variance they share with included variables may be wrongly attributed to those variables, or the error term may be inflated.

3. **Exclusion of all irrelevant variables**: If causally irrelevant variables are included in the model, the common variance they share with included variables may be wrongly attributed to the irrelevant variables. The more the correlation of the irrelevant variable(s) with other independents, the greater the standard errors of the regression coefficients for these independents.

4. **Error terms are assumed to be independent (independent sampling)**. Violations of this assumption can have serious effects. Violations will occur, for instance, in correlated samples and repeated measures designs, such as before-after or matched-pairs studies, cluster sampling, or time-series data. That is, subjects cannot provide multiple observations at different time points.Conditional logit models in Cox regression and logistic models for matched pairs in multinomial logistic regression are available to adapt logistic models to handle non-independent data.

5. **Low error in the explanatory variables**. Ideally assumes low measurement error and no missing cases. See here for further discussion of measurement error in GLM models.

6. **Linearity**. Logistic regression does not require linear relationships between the independent factor or covariates and the dependent, as does OLS regression, but it does assume a linear relationship between the independents and the log odds (logit) of the dependent. When the assumption of linearity in the logits is violated, then logistic regression will underestimate the degree of relationship of the independents to the dependent and will lack power (generating Type II errors, thinking there is no relationship when there actually is). One strategy for mitigating lack of linearity in the logit of a continuous covariate is to divide it into categories and use it as a factor, thereby getting separate parameter estimates for various levels of the variable.

    - *Box-Tidwell Transformation (Test)*: Add to the logistic model interaction terms which are the crossproduct of each independent times its natural logarithm [(X)ln(X)]. If these terms are significant, then there is nonlinearity in the logit. This method is not sensitive to small nonlinearities.

    - *Orthogonal polynomial contrasts*: This option treats a categorical independent as a

categorical variable with categories assumed to be equally spaced. The logit (effect) coefficients for each category of the categorical explanatory variable should not change over the contrasts. This method is not appropriate when the independent has a large number of values, inflating the standard errors of the contrasts. Select polynomial from the contrast list after clicking the Categorical button in the SPSS logistic regression dialog.

- *Logit step tests*. Another simple method of checking for linearity between an ordinal or interval independent variable and the logit of the dependent variable is to (1) create a new variable which divides the existing independent variable into categories of equal intervals, then (2) run a logistic regression with the same dependent but using the newly categorized version of the independent as a categorical variable with the default indicator coding. If there is linearity with the logit, the b coefficients for each class of the newly categorized explanatory variable should increase (or decrease) in roughly linear steps.

    - *The SPSS Visual Bander* can be used to create a categorical variable based on an existing interval or ordinal one. It is invoked from the SPSS menu by selecting Transform, Visual Bander. In the Visual Bander dialog, enter the variable to categorize and click Continue. In the next dialog, select the variable just entered and also give a name for the new variable to be created; click the Make Cutpoints button and select Equal Intervals (the default), then enter the starting point (ex., 0) and the number of categories (ex., 5) and tab to the Width box, which will be filled in with a default value. Click Apply to return to the Visual Bander dialog, then click OK. A new categorized variable of the name provided will be created at the end of the Data Editor spreadsheet.

    - *A logit graph* can be constructed to visually display linearity in the logit, or lack thereof, for a banded (categorical) version of an underlying continuous variable. After a banded version of the variable is created and substituted into the logistic model, separate b coefficients will be output in the "Parameter Estimates" table. Once can go to this table in SPSS output, double-click it, highlight the b coefficients for each band, right-click, and select Create Graph, Line.

7. **Additivity**. Like OLS regression, logistic regression does not account for interaction effects except when interaction terms (usually products of standardized independents) are created as additional variables in the analysis. This is done by using the categorical covariates option in SPSS's logistic procedure.

8. **No multicollinearity**: To the extent that one independent is a linear function of another independent, the problem of multicollinearity will occur in logistic regression, as it does in OLS regression. As the independents increase in correlation with each other, the standard errors of the logit (effect) coefficients will become inflated. Multicollinearity does not change the estimates of the coefficients, only their reliability. High standard errors flag possible multicollinearity. Multicollinearity and its handling is discussed more extensively in the StatNotes section on multiple regression.

9. **No outliers**. As in OLS regression, outliers can affect results significantly. The researcher should analyze standardized residuals for outliers and consider removing them or modeling

them separately.Standardized residuals >2.58 are outliers at the .01 level, which is the customary level (standardized residuals > 1.96 are outliers at the less-used .05 level). Standardized residuals are requested under the "Save" button in the binomial logistic regression dialog box in SPSS. For multinomial logistic regression, checking "Cell Probabilities" under the "Statistics" button will generate actual, observed, and residual values.

10. **Large samples**. Also, unlike OLS regression, logistic regression uses maximum likelihood estimation (MLE) rather than ordinary least squares (OLS) to derive parameters. MLE relies on large-sample asymptotic normality which means that reliability of estimates decline when there are few cases for each observed combination of independent variables. That is, in small samples one may get high standard errors. In the extreme, if there are too few cases in relation to the number of variables, it may be impossible to converge on a solution. Very high parameter estimates (logistic coefficients) may signal inadequate sample size. As a rule of thumb, Peduzzi et al. (1996) recommend that the smaller of the classes of the dependent variable have at least 10 events per parameter in the model.

11. **Sampling adequacy**. Goodness of fit measures like model chi-square assume that for cells formed by the categorical independents, all cell frequencies are >=1 and no more than 20% of cells are < 5. Researchers should run crosstabs to assure this requirement is met. Sometimes one can compensate for small samples by combining categories of categorical independents or by deleting independents altogether.

12. **Expected dispersion**. In logistic regression the expected variance of the dependent can be compared to the observed variance, and discrepancies may be considered under- or overdispersion. If there is moderate discrepancy, standard errors will be over-optimistic and one should use adjusted standard error. Adjusted standard error will make the confidence intervals wider. However, if there are large discrepancies, this indicates a need to respecify the model, or that the sample was not random, or other serious design problems.The expected variance is ybar*(1 - ybar), where ybar is the mean of the fitted (estimated) y. This can be compared with the actual variance in observed y to assess under- or overdispersion. Adjusted SE equals SE * SQRT(D/df), where D is the scaled deviance, which for logistic regression is -2LL, which is -2Log Likelihood in SPSS logistic regression output.

SPSS Output for Logistic Regression

- **Commented SPSS Output for Logistic Regression**

Frequently Asked Questions

- **Why not just use regression with dichotomous dependents?**
- **When is OLS regression preferred to logistic regression?**

- **Why not just use regression with dichotomous dependents?**

  Use of a dichotomous dependent in OLS regression violates the assumptions of normality and homoscedasticity as a normal distribution is impossible with only two values. Also, when the values can only be 0 or 1, residuals (error) will be low for the portions of the regression line near Y=0 and Y=1, but high in the middle -- hence the error term will violate the assumption of homoscedasticity (equal variances) when a dichotomy is used as a dependent. Even with large samples, standard errors and significance tests will be in error because of lack of homoscedasticity. Also, for a dependent which assumes values of 0 and 1, the regression model will allow estimates below 0 and above 1. Also, multiple linear regression does not handle non-linear relationships, whereas log-linear methods do. These objections to the use of regression with dichotomous dependents apply to polytomous dependents also.

- **When is OLS regression preferred to logistic regression?**

  With a multinomial dependent when all assumptions of OLS regression are met, OLS

regression usually will have more power than logistic regression. That is, there will be fewer Type II errors (thinking there is no relationship when there actually is). OLS assumptions cannot be met with a binary dependent. Also, the maximum number of independents should be substantially less for logistic as compared to OLS regression as the categorization of logistic dependents means less information content. With a binary dependent, it is impossible to meet the normality assumptions of OLS regression, but if the split is not extreme (not 90:10 or worse), OLS regression will not return dramatically different substantive results. Still, logistic regression is clearly preferred for binary response (dependent) variables.

- **What is the SPSS syntax for logistic regression?**
  With SPSS, logistic regression is found under Analyze - Regression - Binary Logistic or Multinomial Logistic.

```
LOGISTIC REGRESSION /VARIABLES income WITH age SES gender opinion1
     opinion2 region
    /CATEGORICAL=gender, opinion1, opinion2, region
    /CONTRAST(region)=INDICATOR(4)
    /METHOD FSTEP(LR)
    /CLASSPLOT
```

  Above is the SPSS syntax in simplified form. The dependent variable is the variable immediately after the VARIABLES term. The independent variables are those immediately after the WITH term. The CATEGORICAL command specifies any categorical variables; note these must also be listed in the VARIABLES statement. The CONTRAST command tells SPSS which category of a categorical variable is to be dropped when it automatically constructs dummy variables (here it is the 4th value of "region"; this value is the fourth one and is not necessarily coded "4"). The METHOD subcommand sets the method of computation, here specified as FSTEP to indicate forward stepwise logistic regression. Alternatives are BSTEP (backward stepwise logistic regression) and ENTER (enter terms as listed, usually because their order is set by theories which the researcher is testing). ENTER is the default method. The (LR) term following FSTEP specifies that likelihood ratio criteria are to be used in the stepwise addition of variables to the model. The /CLASSPLOT option specifies a histogram of predicted probabilities is to output (see above).

- **Can I create interaction terms in my logistic model, as with OLS regression?**
  Yes. As in OLS regression, interaction terms are constructed as crossproducts of the two interacting variables.

- **Will SPSS's logistic regression procedure handly my categorical variables automatically?**
  No. You must declare your categorical variables categorical if they have more than two values. This is done by clicking on the "Categorical" button in the Logistic Regression dialog box. After this, SPSS will automatically create dummy variables based on the categorical variable.

- **Can I handle missing cases the same in logistic regression as in OLS regression?**
  No. In the linear model assumed by OLS regression, one may choose to estimate missing values based on OLS regression of the variable with missing cases, based on non-missing data. However, the nonlinear model assumed by logistic regression requires a full set of data. Therefore SPSS provides only for LISTWISE deletion of cases with missing data, using the remaining full dataset to calculate logistic parameters.

- **Is it true for logistic regression, as it is for OLS regression, that the beta weight (standardized logit coefficient) for a given independent reflects its explanatory power controlling for other variables in the equation, and that the betas will change if variables are added or dropped from the equation?**

  Yes, the same basic logic applies. This is why it is best in either form of regression to compare two or more models for their relative fit to the data rather than simply to show the data are not inconsistent with a single model. The model, of course, dictates which variables are entered and one uses the ENTER method in SPSS, which is the default method.

- **What is the coefficient in logistic regression which corresponds to R-Square in multiple regression?**

  There is no exactly analogous coefficient. See the discussion of RL-squared, above. *Cox and Snell's R-Square* is an attempt to imitate the interpretation of multiple R-Square, and *Nagelkerke's R-Square* is a further modification of the Cox and Snell coefficient to assure that it can vary from 0 to 1.

- **Is there a logistic regression analogy to adjusted R-square in OLS regression?**

  Yes. **RLA-squared** is adjusted RL-squared, and is similar to adjusted R-square in OLS regression. RLA-squared penalizes RL-squared for the number of independents on the assumption that R-square will become artificially high simply because some independents' chance variations "explain" small parts of the variance of the dependent. RLA-squared = (GM - 2k)/DO, where k = the number of independents.

- **Is multicollinearity a problem for logistic regression the way it is for multiple linear regression?**

  Absolutely. The discussion in "Statnotes" under the "Regression" topic is relevant to logistic regression.

- **What is the logistic equivalent to the VIF test for multicollinearity in OLS regression? Can odds ratios be used?**

  Multicollinearity is a problem when high in either logistic or OLS regression because in either case standard errors of the b coefficients will be high and interpretations of the relative importance of the independent variables will be unreliable. In an OLS regression context, recall that VIF is the reciprocal of tolerance, which is 1 - R-squared. When there is high multicollinearity, R-squared will be high also, so tolerance will be low, and thus VIF will be high. When VIF is high, the b and beta weights are unreliable and subject to misinterpretation. For typical social science research, multicollinearity is considered not a problem if VIF <= 4, a level which corresponds to doubling the standard error of the b coefficient.

  As there is no direct counterpart to R-squared in logistic regression, VIF cannot be computed -- though obviously one could apply the same logic to various psuedo-R-squared measures. Unfortunately, I am not aware of a VIF-type test for logistic regression, and I would think that the same obstacles would exist as for creating a true equivalent to OLS R-squared.

  A high odds ratio would not be evidence of multicollinearity in itself.

  To the extent that one independent is linearly or nonlinearly related to another independent,

multicollinearity could be a problem in logistic regression since, unlike OLS regression, logistic regression does not assume linearity of relationship among independents. Some authors use the VIF test in OLS regression to screen for multicollinearity in logistic regression if nonlinearity is ruled out. In an OLS regression context, nonlinearity exists when eta-square is significantly higher than R-square. In a logistic regression context, the Box-Tidwell transformation and orthogonal polynomial contrasts are ways of testing linearity among the independents.

- **How can one use estimated variance of residuals to test for model misspecification?**
  - The misspecification problem may be assessed by comparing expected variance of residuals with observed variance. Since logistic regression assumes binomial errors, the estimated variance (y) = m(1 - m), where m = estimated mean residual. "Overdispersion" is when the observed variance of the residuals is greater than the expected variance. Overdispersion indicates misspecification of the model, non-random sampling, or an unexpected distribution of the variables. If misspecification is involved, one must respecify the model. If that is not the case, then the computed standard error will be over-optimistic (confidence intervals will be too wide). One suggested remedy is to use adjusted SE = SE*SQRT(s), where s = D/df, where D = dispersion and df=degrees of freedom in the model.

- **How are interaction effects handled in logistic regression?**

  The same as in OLS regression. One must add interaction terms to the model as crossproducts of the standardized independents and/or dummy independents. Some computer programs will allow the researcher to specify the pairs of interacting variables and will do all the computation automatically. In SPSS, use the categorical covariates option: highlight two variables, then click on the button that shows >a*b> to put them in the Covariates box .The significance of an interaction effect is the same as for any other variable, except in the case of a set of dummy variables representing a single ordinal variable.

  When an ordinal variable has been entered as a set of dummy variables, the interaction of another variable with the ordinal variable will involve multiple interaction terms. In this case the significance of the interaction of the two variables is the significance of the change of R-square of the equation with the interaction terms and the equation without the set of terms associated with the ordinal variable. (See the StatNotes section on "Regression" for computing the significance of the difference of two R-squares).

- **Does stepwise logistic regression exist, as it does for OLS regression?**

  Yes, it exists, but it is not supported by all computer packages. It is supported by SPSS. Stepwise regression is used in the exploratory phase of research or for purposes of pure prediction, not theory testing. In the theory testing stage the researcher should base selection of the variables on theory, not on a computer algorithm. Menard (1995: 54) writes, "there appears to be general agreement that the use of computer-controlled stepwise procedures to select variables is inappropriate for theory testing because it capitalizes on random variations in the data and produces results that tend to be idosyncratic and difficult to replicate in any sample other than the sample in which they were originally obtained." Those who use this procedure often focus on *step chi-square* output in SPSS, which represents the change in the likelihood ratio test (model chi-square test) at each step.

- **What if I use the mulitnomial logistic option when my dependent is binary?**

Binary dependents can be fitted in both the binary and multinomial logistic regression options of SPSS, with different options and output. This can be done but the multinomial procedure will aggregate the data, yielding different goodness of fit tests. The SPSS 14 online help manual notes, " An important theoretical distinction is that the Logistic Regression procedure produces all predictions, residuals, influence statistics, and goodness-of-fit tests using data at the individual case level, regardless of how the data are entered and whether or not the number of covariate patterns is smaller than the total number of cases, while the Multinomial Logistic Regression procedure internally aggregates cases to form subpopulations with identical covariate patterns for the predictors, producing predictions, residuals, and goodness-of-fit tests based on these subpopulations. If all predictors are categorical or any continuous predictors take on only a limited number of values—so that there are several cases at each distinct covariate pattern—the subpopulation approach can produce valid goodness-of-fit tests and informative residuals, while the individual case level approach cannot."

- **What is nonparametric logistic regression and how is it more nonlinear?**

  In general, nonparametric regression as discussed in the section on OLS regression can be extended to the case of GLM regression models like logistic regression. See Fox (2000: 58-73).

  GLM nonparametric regression allows the logit of the dependent variable to be a nonlinear function of the parameter estimates of the independent variables. While GLM techniques like logistic regression are nonlinear in that they employ a transform (for logistic regression, the natural log of the odds of a dependent variable) which is nonlinear, in traditional form the result of that transform (the logit of the dependent variable) is a linear function of the terms on the right-hand side of the equation. GLM non-parametric regression relaxes the linearity assumption to allow nonlinear relations over and beyond those of the link function (logit) transformation.

  Generalized nonparametric regression is a GLM equivalent to OLS local regression (local polynomial nonparametric regression), which makes the dependent variable a single nonlinear function of the independent variables. The same problems noted for OLS local regression still exist, notably difficulty of interpretation as independent variables increase.

  Generalized additive regression is the GLM equivalent to OLS additive regression, which allow the dependent variable to be the additive sum of nonlinear functions which are different for each of the independent variables. Fox (2000: 74-77) argues that generalized additive regression can reveal nonlinear relationships under certain circumstances where they are obscured using partial residual plots alone, notably when a strong nonlinear relationship among independents exists alongside a strong nonlinear relatinship between an independent and a dependent.

- **How can I use matched pairs data in conditional multinomial logistic regression?**

  Multinomial logistic regression can be used for analysis of matched case-control pairs. In the data setup, every id number has three rows: The case row, the control person row paired with that case, and the difference row obtained by subtraction. Analysis is done on the difference row by using Data, Select cases, and selecting for type = "diff" or similar coding, where "type" is a column with values "case," "control,", and "diff" for the three rows of data for any id. Any categorical variables like religion=1,2,3 must be replaced with sets of

dummies such that religion1 = 0,1, religion2=0,1, and religion3 is omitted as the reference category. That is, by default the highest variable becomes the reference category. All predictors are entered as covariates. There are no factors because all categorical variables have been transformed into 0,1 dummy variables.

The dependent has to be a constant, such as '1.' That is, the researcher must have a column in the data setup, perhaps labeled "control" and with 1's for all rows.The real dependent is whatever the researcher is matching cases on, for instance, people with and without heart attacks but matched otherwise on a set of variables like age, weight, etc.

Under the Model button, the researcher requests no intercept.

Output will be the same as for other multinomial logistic regression models. Note the odds are usually set up as case:control, so that in the example of heart attacks, the cases might be heart-attack people and controls would be non-heart-attack matched pairs. The dependent reference category would become control=non-heart-attack. Let a covariate be the dichotomous variable 0=white, 1=non-white, and its odds ratio be 1.3. Its reference category by default would be 1=non-white. The odds ratio statement would take the form, the odds of a white not getting a heart attack compared to getting one is 1.3 times that of a non-white. Put another way, being white increases the odds of not getting a heart attack by a factor of 1.3.

- **How many independents can I have?**

  There is no precise answer to this question, but the more independents, the more likelihood of multicollinearity. In general, there should be significantly fewer independents than in OLS regression as logistic dependents, being categorized, have lower information content. Also, if you have 20 independents, at the .05 level of significance you would expect one to be found to be significant just by chance. A rule of thumb is that there should be no more than 1 independent for each 10 cases in the sample. In applying this rule of thumb, keep in mind that if there are categorical independents, such as dichotomies, the number of cases should be considered to be the lesser of the groups (ex., in a dichotomy with 480 0's and 20 1's, effective size would be 20), and by the 1:10 rule of thumb, the number of independents should be the smaller group size divided by 10 (in the example, 20/10 = 2 independents maximum).

- **How do I express the logistic regression equation if one or more of my independents is categorical?**

  When a covariate is categorical, SPSS will print out "parameter codings," which are the internal-to-SPSS values which SPSS assigns to the levels of each categorical variable. These parameter codings are the X values which are multiplied by the logit (effect) coefficients to obtain the predicted values.

- **How do I compare logit coefficients across groups formed by a categorical independent variable?**

  There are two strategies. The first strategy is to separate the sample into subgroups, then perform otherwise identical logistic regression for each. One then computes the p value for a Wald chi-square test of the significance of the differences between the corresponding coefficients. The formula for this test, for the case of two subgroup parameter estimates, is Wald chi-square = $[(b_1 - b_2)^2]/\{[se(b_1)]^2 + [se(b_2)]^2$, where the b's are the logit coefficients

for groups 1 and 2 and the se terms are their corresponding standard errors. This chi-square value is read from a table of the chi-square distribution with 1 degree of freedom.

The second strategy is to create an indicator (dummy) variable or set of variables which reflects membership/non-membership in the group, and also to have interaction terms between the indicator dummies and other independent variables, such that the significant interactions are interpreted as indicating significant differences across groups for the corresponding independent variables. When an indicator variable has been entered as a set of dummy variables, its interaction with another variable will involve multiple interaction terms. In this case the significance of the interaction of the indicator variable and another independent variable is the significance of the change of R-square of the equation with the interaction terms and the equation without the set of terms associated with the ordinal variable. (See the StatNotes section on "Regression" for computing the significance of the difference of two R-squares).

Allison (1999: 186) has shown that "Both methods may lead to invalid conclusions if residual variation differs across groups." Unequal residual variation across groups will occur, for instance, whenever an unobserved variable (whose effect is incorporated in the disturbance term) has different impacts on the dependent variable depending on the group. Allison suggests that, as a rule of thumb, if "one group has coefficients that are consistently higher or lower than those in another group, it is a good indication of a potential problem ..." (p, 199). Allison explicated a new test to adjust for unequal residual variation, presenting the code for computation of this test in SAS, LIMDEP, BMDP, and STATA. The test is not implemented directly by SPSS or SAS, at least as of 1999. Note Allison's test is conservative in that it will always yield a chi-square which is smaller than the conventional test, making it harder to prove the existence of cross-group differences.

- **How do I compute the confidence interval for the unstandardized logit (effect) coefficients?**
  To obtain the upper confidence limit at the 95% level, where b is the unstandardized logit coefficient, se is the standard error, and e is the natural logarithm, take e to the power of (b + 1.96*se). Subtract to get the lower CI.

- **SAS's PROC CATMOD for multinomial logistic regression is not user friendly. Where can I get some help?**
  - SAS CATMOD Examples
  - University of Idaho
  - York University, on the CATPLOT module

Bibliography

- Agresti, Alan (1996). *An introduction to categorical data analysis*. NY: John wiley. An excellent, accessible introduction.
- Allison, Paul D. (1999). Comparing logit and probit coefficients across groups. *Sociological Methods and Research*, 28(2): 186-208.
- Cox, D.R. and E. J. Snell (1989). Analysis of binary data (2nd edition). London: Chapman & Hall.
- DeMaris, Alfred (1992). *Logit modeling: Practical applications*. Thousand Oaks, CA: Sage

Publications. Series: Quantitative Applications in the Social Sciences, No. 106.

- Estrella, A. (1998). A new measure of fit for equations with dichotomous dependent variables. *Journal of Business and Economic Statistics* 16(2): 198-205. Discusses proposed measures for an analogy to R^2.
- Fox, John (2000). Multiple and generalized nonparametric regression. Thousand Oaks, CA: Sage Publications. Quantitative Applications in the Social Sciences Series No.131. Covers nonparametric regression models for GLM techniques like logistic regression. Nonparametric regression allows the logit of the dependent to be a nonlinear function of the logits of the independent variables.
- Hosmer, David and Stanley Lemeshow (1989). *Applied Logistic Regression.* NY: Wiley & Sons. A much-cited treatment utilized in SPSS routines.
- Jaccard, James (2001). *Interaction effects in logistic regression.* Thousand Oaks, CA: Sage Publications. Quantitative Applications in the Social Sciences Series, No. 135.
- Kleinbaum, D. G. (1994). *Logistic regression: A self-learning text.* New York: Springer-Verlag. What it says.
- McKelvey, Richard and William Zavoina (1994). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, 4: 103-120. Discusses polytomous and ordinal logits.
- Menard, Scott (2002). *Applied logistic regression analysis, 2nd Edition.* Thousand Oaks, CA: Sage Publications. Series: Quantitative Applications in the Social Sciences, No. 106. First ed., 1995.
- Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, Vol. 78, No. 3: 691-692. Covers the two measures of R-square for logistic regression which are found in SPSS output.
- O'Connell, Ann A. (2005). *Logistic regression models for ordinal response variables.* Thousand Oaks, CA: Sage Publications. Quantittive Applications in the Social Sciences, Volume 146.
- Pampel, Fred C. (2000). *Logistic regression: A primer.* Sage Quantitative Applications in the Social Sciences Series #132. Thousand Oaks, CA: Sage Publications. Pp. 35-38 provide an example with commented SPSS output.
- Peduzzi, P., J. Concato, E. Kemper, T. R. Holford, and A. Feinstein (1996). A simulation of the number of events per variable in logistic regression analysis. *Journal of Climical Epidemiology* 99: 1373-1379.
- Press, S. J. and S. Wilson (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, Vol. 73: 699-705. The authors make the case for the superiority of logistic regression for situations where the assumptions of multivariate normality are not met (ex., when dummy variables are used), though discriminant analysis is held to be better when they are. They conclude that logistic and discriminant analyses will usually yield the same conclusions, except in the case when there are independents which result in predictions very close to 0 and 1 in logistic analysis. This can be revealed by examining a 'plot of observed groups and predicted probabilities' in the SPSS logistic regression output.
- Raftery, A. E. (1995). Bayesian model selection in social research. In P. V. Marsden, ed., *Sociological Methodology 1995*: 111-163. London: Tavistock. Presents BIC criterion for evaluating logits.
- Rice, J. C. (1994). "Logistic regression: An introduction". In B. Thompson, ed., *Advances in social science methodology*, Vol. 3: 191-245. Greenwich, CT: JAI Press. Popular introduction.
- Tabachnick, B.G., and L. S. Fidell (1996). *Using multivariate statistics*, 3rd ed. New York: Harper Collins. Has clear chapter on logistic regression.

- Wright, R.E. (1995). "Logistic regression". In L.G. Grimm & P.R. Yarnold, eds., *Reading and understanding multivariate statistics.* Washington, DC: American Psychological Association. A widely used recent treatment.

Last update 1/6/08.

Back
