THE PRACTICAL VALUE OF LOGISTIC REGRESSION

Frank E. Harrell, Jr.
Kerry L. Lee
Duke University Medical Center, Durham NC

Abstract

Logistic multiple regression using the method of maximum likelihood is now the method of choice for many regression-type problems involving binary, ordinal, or nominal dependent variables. Logistic regression does not require grouping of observations to obtain valid estimates of effects and of outcome probabilities, and it has been shown in the binary case to provide more accurate probability estimates than linear discriminant analysis when the assumptions of the latter (i.e., multivariate normality of predictor variables with common covariance matrix) are violated. Even when multivariate normality holds, logistic regression has been shown to yield probability estimates virtually as accurate as those obtained using discriminant analysis.

The assumptions of the logistic regression model are for the most part straightforward and easy to verify. A general purpose SAS macro language program to verify the assumptions of the binary or ordinal model graphically will be discussed. Examples demonstrating the advantages of logistic regression for binary and ordinal dependent variables over other methods will also be presented.

Background

Walker and Duncan [1] formulated the general logistic multiple regression models. These models allow a mixture of continuous and nominal independent or predictor variables to be related to a binary, ordinal, or nominal dependent or response variable. No grouping of continuous predictor variables is required and no two observations need have the same values of the predictors, as the method of maximum likelihood is used to obtain estimates of the regression coefficients.

For a binary response variable Y_i=0 or 1 for the i^{th} observation, the binary logistic model assumes that

$$\text{Prob}\,(Y_i=1 \mid X_{i1},X_{i2},\ldots,X_{ip}) = \frac{1}{1+e^{-(\alpha+\beta_1 X_{i1}+\beta_2 X_{i2}+\ldots+\beta_p X_{ip})}}, \qquad (1)$$

where "Prob" denotes "probability" and "|" denotes "given" or "conditioned on", and the predictors are $X_{i1},\ldots, X_{ip}$ for the i^{th} observation. Here α is the intercept and the βs are the regression coefficients.

The nominal or polychotomous logistic model is a generalization of (1). The polychotomous model has the disadvantage of requiring a large number of parameters to be estimated (specifically (p+1) x (c-1) where c is the number of categories of Y), resulting in efficiency problems.

The ordinal logistic or proportional odds model [1, section 6], for an ordinal dependent variable having values 0,1,...,K assumes that for $1 \le j \le K$,

$$\text{Prob}(Y_i \ge j \mid X_{i1},\ldots,X_{ip}) = \frac{1}{1+e^{-(\alpha_j+\beta_1 X_{i1}+\ldots+\beta_p X_{ip})}}. \qquad (2)$$

A separate intercept parameter α_j is required for each level of Y_i=1,2,...,K. and $\text{Prob}(Y_i=0 \mid X_{i1},\ldots,X_{ip})$ is obtained from 1-$\text{Prob}(Y_i \ge 1 \mid X_{i1},\ldots,X_{ip})$. This model utilizes the ordering of Ys; the estimates of the αs will be in order. The model assumes that the odds that $Y_i \ge a$ is a constant (not depending on the Xs) multiple of the odds that $Y_i \ge b$ for fixed values of the Xs. Loosely speaking, this implies that what causes Y_i to increase from, say, 1 to 2 is an extension of what causes it to increase from 0 to 1. No assumptions are made regarding the spacing of scale intervals. The values 0-K are used for convenience.

The LOGIST procedure [2] in the SUGI Supplemental Library can fit binary and ordinal dependent variables, perform tests of association between one or more Xs and Y, test for lack of fit of the model, compute various indexes of model prediction ability, and compute predicted probabilities. For a binary response, LOGIST is run with the statement

```
PROC LOGIST options;
MODEL Y=X1...XP/model-building options;
```

An ordinal logistic analysis is obtained by specifying

```
PROC LOGIST K=maximum Y value  options;
  MODEL Y=X1...XP/model-building options;
```

The FUNCAT procedure [3] can fit the polychotomous (and binary) logistic regression models.

Logistic models are being used with increasing frequency to model binary, ordinal, and polychotomous responses. Some of the areas of application are listed below:

1. Predicting the probability that a particular person has a specific disease
2. Predicting the probability of an event by a fixed time period
3. Predicting whether or not a consumer buys a certain product
4. Predicting an ordered response, e.g., "good", "better", "best"
5. Predicting the severity of a disease or other outcome
6. Testing whether a variable is an "independent risk factor"
7. Generalizing the Wilcoxon-Mann-Whitney, Kruskal-Wallis and Spearman rank tests
8. Testing for differences among several variables between two or more groups (with fewer assumptions than Hotelling's T^2)
9. Generalizing Cochran-Mantel-Haenszel tests [4].

Logistic models are so widely applicable because they do not assume anything about the distribution of the Xs, because they allow usage of non-continuous Xs, and because they utilize information in ordered response variable categories. Weighted least squares methods such as that of Grizzle, Starmer, and Koch [5], suffer when there are few ties in the Xs. Discriminant analysis suffers when the Xs are not normally distributed [6], especially when one or more of the predictors is discrete. Even when all assumptions of discriminant analysis hold, logistic regression is virtually as efficient [7].

An example demonstrating the advantages of the ordinal logistic model is shown in [2]. Suppose that a patient received one of three treatments, A, B, and C, and an investigator is interested in testing whether there are any differences among the treatments in the severity of symptoms (including death). Here Y is coded 0 for no symptoms, 1 for presence of pain, and 2 for death. The hypothetical frequency table follows:

Treatment	Y=0	Y=1	Y=2
A	7	2	0
B	3	4	5
C	2	5	2

If one applied the standard χ^2 test for association to this contingency table, the (likelihood ratio) test statistic would be $\chi^2=7.40$ with 4 d.f., and the p-value is .12. This is identical to the likelihood ratio test statistic arising from the polychotomous logistic model. The ordinal logistic model yields a likelihood ratio χ^2 statistic of 6.99 with 2 d.f., with p=.03. The ordinal model takes the ordering of Y into account in addition to the high degree of ties. The corresponding Kruskal-Wallis test would not give accurate significance levels with a large number of ties in Y.

Another very important property of logistic models is that their assumptions are verifiable. For example, instead of examining whether or not the Xs have a multivariate normal distribution, we examine the shape of the relationship between X and the probability that Y is in a certain category.

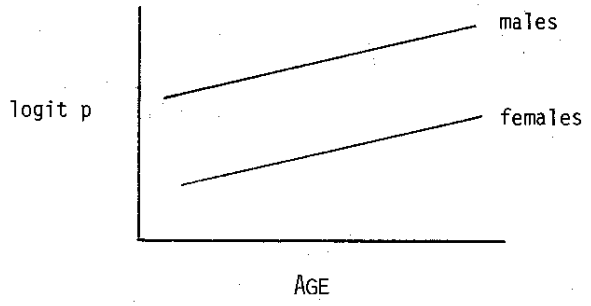
Examining Model Assumptions

Another method of stating the binary logistic model leads to simple methods of validating its assumptions graphically. Equation (1) can be re-written

$$\text{logit Prob}(Y_i=i \mid X_{i1},\ldots,X_{ip}) = \alpha+\beta_1X_{i1}+\beta_2X_{i2}+\ldots+\beta_pX_{ip}, \quad (3)$$

where logit p= log [p/(1-p)] is the log-odds. Hence the model is a linear regression model in the log odds that $Y_i=1$. Since there is no "error term" and hence no "distribution of residuals", the only ways the model can be invalid are non-linearity in one or more Xs, simple or complex interactions among the Xs, or non-independence of the observations. Of course, if the relationship between X and logit Prob ($Y_i=1$) were very complex (e.g., logit Prob ($Y_i=1$) = log ($\alpha+\beta_1X_{i1}+\ldots+\beta_pX_{ip}$)), a simple polynomial in the Xs would not fit the data; the analyst would probably not learn this merely from testing quadratic and cross-product terms.

Suppose that the only predictor variable was the sex of a subject, coded 0 for male, and 1 for female. There is no way that the logistic model can't fit the data -- the model in that case is just fitting two cell proportions. Now suppose that age was the only predictor. The lack of fit could be tested by including a square and perhaps a cubic term in the model. If both age and sex are independent variables, the model (without interaction) assumes the relationships shown in the following figure.



A formal test of the model assumptions, having reasonable power against many alternatives, can be obtained by testing simultaneously for interaction and a quadratic age relationship. This can be done with PROC LOGIST by specifying

```
PROC LOGIST; MODEL Y=age sex agesex age2/
  SW INCLUDE=2 PRINTI PRINTQ SLE=0;
```

LOGIST will print a residual χ^2 with 2 d.f. for testing jointly the added effect of age x sex and age^2, adjusting for the main effects of age and sex.

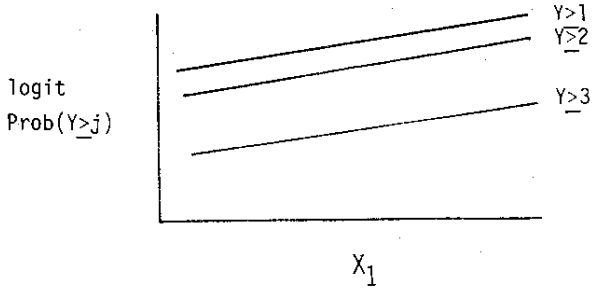
The ordinal logistic model can be stated

$$\text{logit Prob}(Y_i \geq j | X_{i1}, \ldots, X_{ip}) = \alpha_j + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}.$$

For a given category j, the regression assumptions can be verified by plotting the logit of the proportion of $Y \geq j$ versus X. The ordinality (proportional odds) assumption can be checked by noting that

$$\text{logit Prob}(Y_i \geq a | X_{i1}, \ldots, X_{ip}) - \text{logit Prob}(Y_i \geq b | X_{i1}, \ldots, X_{ip}) = \alpha_a - \alpha_b,$$

when $1 \leq a, b \leq K$. Hence the ordinality assumption is equivalent to the logit cumulative probability curves being parallel (equidistant). If there is one X and linearity as well as ordinality hold, the following relationships will obtain:



Linearity can be formally tested just as with the binary model. The ordinality assumption can be formally tested by allowing the regression coefficients to vary with the category of Y and then testing if these coefficients are equal across Y. Such a test is a planned future enhancement in the LOGIST procedure.

A Tool for Checking Model Assumptions Graphically

Since the logistic model assumptions are relatively straightforward, it is surprising that the model is often used with no checking of its assumptions and with no attempt to transform continuous variables to satisfy the linearity assumption. The major reason that data analysts do not always validate the assumptions routinely is that simple scatterplots are not adequate, due to the extreme extent of ties in Y. One must judiciously group on X to compute cell proportions and plot these proportions (or their logits) using suitable X-coordinates.

The simplist method for grouping continuous Xs is to round them. However this results in some intervals having too few points to be able to obtain reliable estimates. Alternative methods that group an X into intervals of varying width that contain a given proportion or number of observations can help solve this problem. For example, it is a common practice to group a variable into deciles and then to compute the proportion of Y=1 in each decile. PROC RANK makes this easy; the only problem remaining is to decide which x-coordinate to use for each decile. The interval midpoint is commonly used for graphing, but the distribution of Xs in each interval may be asymmetric, especially in the lowest and highest decile. The mean X in each interval is a more appropriate summary of what that decile represents.

A SAS* statement-form macro language procedure, called EMPTREND for empirical trend plot, was developed to allow the user to obtain graphs to check binary and ordinal logistic model (as well as other models) assumptions, using only one SAS statement. EMPTREND plots the relationship between one or two predictor variables and a binary, ordinal, or continuous response variable. The first predictor variable, called X, is usually continuous. The optional second predictor is a class variable; it may be discrete or continuous.

EMPTREND first groups the observations by X and optionally by the CLASS variable, using one of three methods selected by the user. It then computes the mean Y, proportion of Y=1, median Y, or all cumulative proportions $Y_i \geq j$ $j=1,2,\ldots,K$ in each X-group, depending on the method chosen. The mean X is also computed for each group.

The X variable may be grouped by rounding, by forming quantile (quintiles, deciles, etc.) groups, or by sorting the dataset on X and separating the dataset into groups having a specified minimum number of observations. For the latter method, the N observations having the lowest X-values form the first group, the next N the second group, and so on. The CLASS variable can either be treated as a discrete variable (without grouping), rounded, or grouped into quantiles.

EMPTREND is invoked by the following statements:

```
%INCLUDE macrolibrary (EMPTREND);
*Must appear once per job;
EMPTREND x y options;
```

Here x are y respectively are the SAS names of the X and Y variables. The options that may appear in the EMPTREND statement follow.

DATA =	name of dataset to analyze defaults to last one created
ROUND=r	group x by rounding to the nearest r units
GROUPS=g	group x into g quantile groups
N=n	group x into intervals each having at least n observations
NMIN=nmin	minimum number of observations to accept in a group; groups having fewer are discarded. Defaults to 10
CLASS=	optional name of CLASS variable
CROUND=r	round CLASS variable to the nearest r
CGROUPS=g	group CLASS variable into g quantiles
K=	Y is ordinal with maximum value K
MEDIAN	if Y is continuous, plot the median Y instead of the mean (which is the default)
PRINT	print estimates as well as plot them
LOGIT	print and plot logits of proportions if Y is binary or ordinal
NOPLOT	don't make graphs (useful only if PRINT is specified)
SASGRAPH	make graphs with SAS/GRAPH* procedure GPLOT as well as with PROC PLOT
FONT=f	if SASGRAPH is given, use font f in the titles default fonts are triplex, italic, complex, duplex for major -> minor titles
OUT=d	store estimates in SAS dataset d

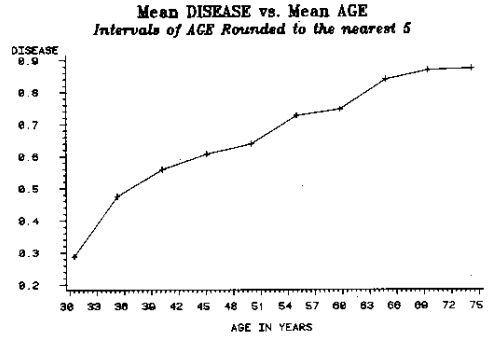
ROUND, GROUPS, and N are mutually exclusive as are CROUND and CGROUPS, K and MEDIAN, MEDIAN and LOGIT. One of ROUND, GROUPS, and N must be specified. The PRINT option is useful for seeing exactly how EMPTREND, grouped the Xs --it causes the minimum, maximum, and mean X in each group to be printed as well as the mean, median, or proportion of Y. If CROUND or CGROUPS is given, PRINT will also print how the CLASS variable was grouped. The positions for x and y in the EMPTREND statement can contain lists of SAS variable names. For example EMPTREND "age bp" "sick dead" ... will result in graphs of age vs. sick, age vs. dead, bp vs sick, and bp vs. dead.

Usage of EMPTREND for examining relationships with binary or ordinal Ys is given by the following series of examples.

Example 1: Round age to the nearest 5 years, print and plot the proportion of patients with DISEASE=1 in each age group versus the mean X in the group. Plot no points represented by fewer than 50 observations.

```
EMPTREND age disease ROUND=5    NMIN=50    PRINT
   SASGRAPH;
```

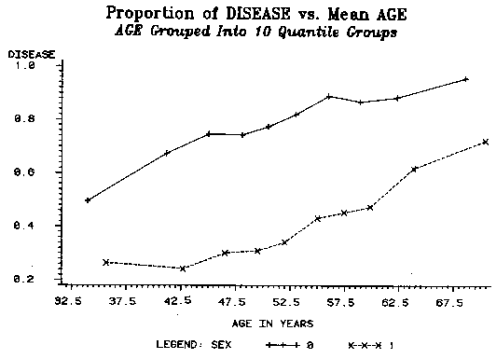
The resulting output is given below.

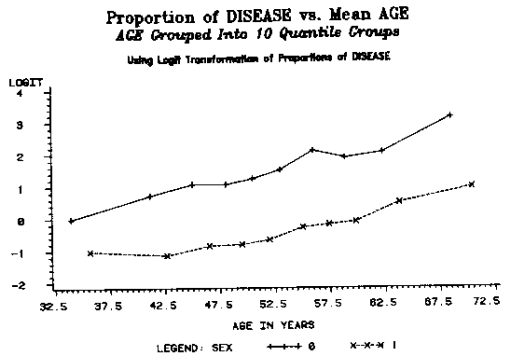


A smooth, consistent relationship between age and the prevalence of disease is obvious from the graph.

Example 2: Group age into deciles, group also by the discrete variable sex, and plot the logit of the proportions.

```
EMPTREND age disease CLASS=sex  GROUPS=10
   LOGIT SASGRAPH;
```

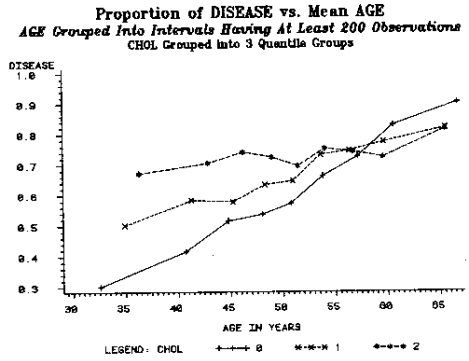


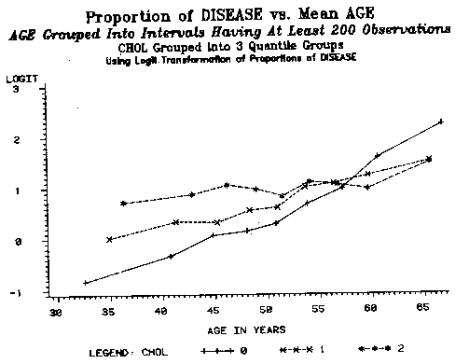


There are independent relationships between age and sex with disease. The relationship for males (sex=0) is apparently linear. Some nonlineality is present for females, which also results in a kind of age x sex interaction.

Example 3: Group age into intervals having at least 200 observations within tertiles of serum cholesterol (chol).

```
EMPTREND age disease   CLASS=chol  N=200
   CGROUPS=3   LOGIT   SASGRAPH;
```

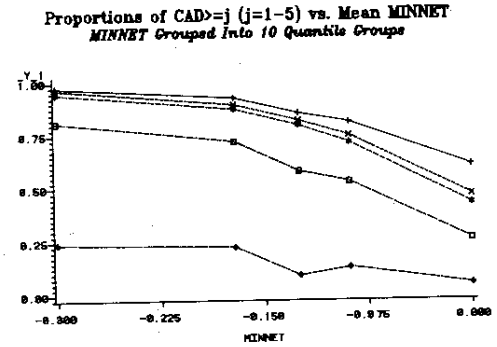


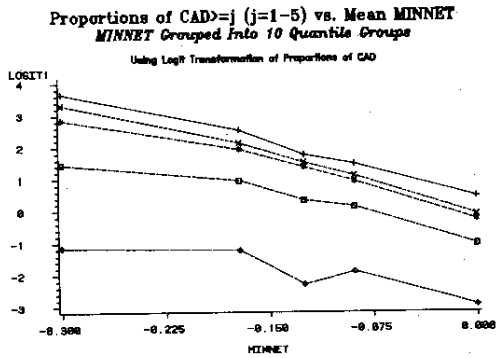


A strong interaction between age and cholesterol is obvious.

Example 4: Group variable minnet into deciles. Within each decile, compute all cumulative probabilities for the ordinal variable cad, which ranges from 0 to 5.

```
EMPTREND minnet cad K=5  GROUPS=10   LOGIT
  SASGRAPH;
```





The logit plot demonstrates a fair degree of linearity. The equal vertical spacings lends support to the ordinal logistic assumption.

Conclusions

The logistic multiple regression models for binary, ordinal, and nominal dependent variables have wide applicability. These models have assumptions that are verifiable and testable. Procedures such as EMPTREND are useful for checking model assumptions graphically and for suggesting data transformations to obtain linearity. There is no excuse for failing to check model assumptions, at least for each predictor variable taken singly.

Acknowledgements

This work was supported by Research Grants HS-03834 and HS-04873 from the National Center for Health Services Research, Research Grant HL-17670 from the National Heart, Lung, and Blood Institute, Training Grant LM-07003 and Grant LM-03373 from the National Library of Medicine, and grants from the Prudential Insurance Company of America, the Kaiser Family Foundation, and the Andrew W. Mellon Foundation.

References

[1] Walker SH, Duncan DB: Estimation of the probability of an event as a function of several independent variables. Biometrika 54:167-79, 1967.

[2] Harrell FE: The LOGIST Procedure. In SUGI Supplemental Library User's Guide, 1983 Edition, ed. S. Joyner. Cary, NC: SAS Institute, Inc.

[3] Ray AA (ed.): SAS User's Guide: Statistics, 1982 Edition. Cary, NC: SAS Institute, Inc.

[4] Day NE, Byar DP: Testing hypotheses in case-control studies - equivalence of Mantel-Haenszel statistics and logit score tests. Biometrics 35:623-30, 1979.

[5] Grizzle JE, Starmer CF, Koch GG: Analysis of categorical data by linear models. Biometrics 25: 489-504, 1969.

[6] Halperin M, Blackwelder WC, Verter JI: Estimation of the multivariate logistic risk function: a comparison of the discriminant function and maximum likelihood approaches. Journal of Chronic Diseases 24:125-58, 1971.

[7] Harrell FE, Lee KL: A comparison of the discrimination of discriminant analysis and logistic regression under multivariate normality. In Biostatistics: Statistics in Biomedical, Public Health and Environmental Sciences, PK Sen, ed. Amsterdam: Elsevier, 1985.

SAS and SAS/GRAPH are registered trademarks of SAS Institute, Inc., Cary, N.C., U.S.A.