PROCEDURES FOR LARGE REGRESSION PROBLEMS REQUIRING MAXIMUM LIKELIHOOD ESTIMATION

Frank E. Harrell, Jr., Kerry L. Lee and Ray A. McKinnis
Duke University

ABSTRACT

This paper deals with methods of handling large regression problems that require iterative maximum likelihood calculations. Ways of fitting a single model or of selecting variables by using forward stepwise or backward elimination algorithms are surveyed. Two supplemental SAS procedures using these methods (PHGLM for Cox Proportional Hazard General Linear Model and LOGIST for logistic binary regression) are discussed.

INTRODUCTION

In many statistical regression problems the normal theory does not apply. This is true for example when the dependent variable is not normally distributed or when its value is truncated or censored as with failure-time data. Instead of calculating the closed-form (least squares) maximum likelihood estimates (MLEs) for the normal case, MLEs must be calculated by an iterative trial and error procedure.

In general, the maximum likelihood (ML) method requires scanning the dataset at each iteration for the trial and error solution. This presents computational problems which are magnified when trying to develop a model by some stepwise strategy.

This paper will discuss the necessary components of a procedure for ML estimation of parameters of a single regression model when there are many variables and/or observations. Then the implementation of time-saving stepwise strategies will be presented. First, two examples of commonly used regression models requiring ML estimation will be given.

LOGISTIC MULTIPLE REGRESSION

The logistic model is used when the dependent variable is binary (0-1). Let Yi denote the dependent variable value for the ith observation, i=1,2,...,n and let the corresponding vector of independent variables be Xi=(Xi1,Xi2,...,Xip) and the regression parameters be B=(B1,...,Bp)'. Here Xi1=1 so that B1 is the intercept parameter. This statistical model assumes that the probability Yi=1 is $1/(1+\exp(-XiB))$. Nothing is assumed about the distribution of the Xij, which may be continuous or categorical indicator variables. This model is a competitor of the classical discriminant model which assumes multivariate normality of the Xi.

SAS procedure LOGIST fits the logistic model. Aside from the features listed below common to both new procedures, LOGIST will also output a SAS dataset containing predicted probabilities that Yi=1 along with confidence intervals on these predicted probabilities. The procedure also prints a classification table for each fitted model showing the number of correct and incorrect predictions of Yi along with several statistics measuring the predictive accuracy. Hence, the procedure can do the work of a stepwise discriminant analysis program (when there are two groups) while also providing a significance test for each independent variable with few assumptions.

COX PROPORTIONAL HAZARD GENERAL LINEAR MODEL

The Cox (1972) proportional hazard linear model is a distribution-free regression model for a continuous dependent variable Yi. The model depends only on the rank ordering of the dependent variable vector and the results are invariant with respect to monotonic transformations on the dependent variable. The model is a competitor for the normal general linear model. Its greatest use has been in analyzing failure-time data as it easily handles censored failure times (i.e., for items not yet failed, the failure time is known only to exceed some value).

Let the probability density function and cumulative distribution function of Yi be fi(y) and Fi(y) respectively. Let Si(y) denote the survival function or 1-Fi(y), and Hi(y) denote the hazard function, fi(y)/Si(y). Let the vector of independent variables be Xi=(Xi1,...,Xip) for the ith observation and B=(B1,...,Bp)' denote the vector of regression parameters. The Cox model does not have an intercept term since it is invariant with respect to shifts in Yi. The basic assmption of the model is that Hi(y)=H(y)exp(XiB). No specific function is assumed for H(y).

The assumption of the proportional hazard model can equivalently be stated as follows (see Kalbfleisch 1978): for some monotonic function g(y), g(Yi)-XiB has the extreme value type I distribution with density exp(-t-exp(-t)). The model assumes nothing about the distribution of a Yi; it only assumes how the distribution of Yi relates to that of Yj, i≠j.

The PHGLM procedure fits the proportional hazard model. The dependent variable may represent complete observations or some of the Yi may be censored. To calculate the index of model fit described below PHGLM first calculates an ad hoc measure of the effective sample size if there are any censored observations. It is the value n* such that a sample of size n* with no censoring has the same log likelihood evaluated at B=0 as the given sample.

FITTING A REGRESSION MODEL BY MAXIMUM LIKELIHOOD

The following is a list of requirements we perceive a procedure that handles large ML problems should meet, along with the ways in which the two new procedures meet these requirements.

1. The procedure should allow for any number of observations (n) and a very large number of independent variables (p<n).

This was accomplished by allocating storage for calculations dynamically and holding as many observations in core as would fit; the remaining observations are stored on a SAS utility file. This is the same way SAS PROC NLIN works.

2. Because of the cost of scanning the data, the estimation process should converge in a few iterations even when there are many parameters to estimate.

We chose the Newton-Raphson method for maximizing the likelihood. Runs estimating as many as 63 parameters have converged with as few as 8 total passes of the data. When the Newton-Raphson method will not converge, it usually diverges quickly so that aborted runs are not very expensive.

3. The procedure should use default values for initial parameter estimates or allow the user to specify initial estimates for any of the parameters.

The default starting value is zero. The procedures optionally use starting values from a secondary SAS input dataset (usually the variables in this dataset are named the same as those in the model). This feature is useful when there are scores of variables being analyzed on separate runs over a period of time; the user merely specifies the name of a permanent dataset to specify starting values when iterations will not converge using default starting values.

4. The procedure should be able to output into a SAS dataset the final vector of parameter estimates and its covariance matrix for testing general linear hypotheses (e.g. using PROC MATRIX).

The dataset created contains the (transposed) vector of parameter estimates as the first observation and the square covariance matrix in the remaining observations. The variable names for this dataset are the names of the independent variables in the model. This format is suitable for PROC MATRIX:

```
PROC MATRIX;
FETCH B DATA=STATS(OBS=1);B=B';
FETCH V DATA=STATS(FIRSTOBS=2);
```

The format is also suitable for PROC SCORE to calculate predicted values and for future runs of PHGLM or LOGIST using the old estimates as new starting values.

STEPWISE VARIABLE SELECTION

Quite often the statistician desires to find a subset of the independent variables that adequately describes the dependent variable. When using ML estimation, the optimal stepwise method would be at each step to add the variable to the model that increases the likelihood the most or to delete the variable that decreases the likelihood the least. For each variable examined this involves re-iterating to find the MLEs for the new set of variables. If there are 50 candidate variables and 5 iterations are required for convergence for each set of parameters, 250 scans of the data would be required at each step. This is not affordable in most cases.

For forward stepwise regression, what is needed is an algorithm for finding the next variable to enter the model with only one scan of the data. To accomplish this, the PHGLM and LOGIST procedures use a modification of the strategy described in Bartolucci and Fraser (1977) who proposed the use of Rao's (1973) efficient score statistic in variable selection. The method is as follows: let B denote the vector of MLEs for variables in the model and let Ui(B,0) be the

derivative of the log likelihood with respect to the ith parameter (for the ith variable not in the model) evaluated at B for variables in the model and at zero for variable i. Similarly let Ii(B,0) denote the negative of the second derivative. The statistic for entering a variable is Qi=Ui(B,0)**2/Ii(B,0). The Qi statistic may be thought of as the locally most powerful test statistic treating the parameters already in the model as known nuisance parameters which are really estimated by their MLEs. Qi has roughly a chi-square distribution with one degree of freedom.

The variable possessing the largest Qi is added to the model. At any step, MLE chi-square statistics (MLEX) ((parameter estimate/standard error)**2 using asymptotic normality of estimate) are computed for all variables in the model and any not significant at a chosen level are dropped. The stepwise process continues until no other variable has a Qi that is significant at the .10 level (by default). The default significance level for a MLEX statistic for a variable to stay in the model is .05. Using the .10 level for Qi allows for some disagreement between the two types of statistics. In practice, the two generally agree well (they can be compared when a variable is actually added to the model). This method has been found to be inexpensive even for large datasets. Harris, et al(1979) used this procedure to model survival of 1214 patients using 81 candidate variables. The resulting 13 variable model cost $4 and 49 CPU seconds to develop.

When there are not too many (say <100) variables to have in one model, a backward elimination method may be used. The variable with the smallest MLEX can be deleted from the model at each step and then iterations can be performed to estimate parameters for the new list. A much faster method which we will call "fast backward elimination" is the one proposed by Lawless and Singhal (1978). This algorithm is an approximate way to eliminate non-significant variables from the first model fitted. First the variable with the smallest MLEX is deleted. No more iterations are made to re-estimate the parameters; instead the estimated covariance matrix is swept (using SAS subroutine $SWEEP) to compute approximate MLEs and a new covariance matrix under the restriction that the parameter for the variable deleted is zero. Next, approximate MLEXs are computed using these approximate estimates. At each step, the residual chi square, which measures the cumulative significance of all variables deleted, is printed. The backward process continues until one of two user-selected criteria is met: either the next variable to delete would have a significant single degree of freedom chi square, or the residual chi square with the next variable deleted would be significant.

The fast backward elimination option is extremely efficient once the MLEs are computed for the first model, as no matrix inversions or scans of the data are required to eliminate variables. In most cases, the final model is the same as the model derived by deleting variables using the true MLEXs.

For each model in which MLEs are computed, the procedures optionally print the estimates, standard errors, and test statistics (MLEXs) for each variable. A statistic called D which measures the fit of the model independent of the sample size is also printed. D is between 0 and 1 and is analogous to the squared multiple correlation coefficient in the normal case; it is the value D such that D(n-p)/(1-D) = model chi-square (-2 log likelihood ratio for all parameters except intercept, if any). Here p is the total number of parameters in the model and n is the the sample size (n* for PHGLM).

One additional feature supported by PHGLM and LOGIST that we have found to be very useful with the forward stepwise technique is that if the user knows in advance an upper bound on the number of variables in the final model he may specify this as the STOP parameter in the MODEL statement. The procedures then allocate storage for covariance matrices of that dimension or less. This allows hundreds of candidate variables to be considered without requiring a large region of core.

STATISTICAL ISSUES IN VARIABLE SELECTION

Variable selection procedures have potential for being abused. For example, a user who claims to have found 10 significant predictor variables out of 500 can be greatly misled. Chances are that as many as 25 significant variables might be found even if there were none- the 25 largest of 500 independent chi-squares are probably each greater than the critical value for an individual chi-square even in the null case. For this reason, we conclude that backward elimination is often superior to the forward stepwise method. With backward elimination, the user gets an optimal (likelihood ratio) test of significance of regression for the entire list of variables, a test that penalizes the user

for the number of variables tested. If the test statistic is not significant, trying to find a significant sub-model can be misleading, for by definition sub-models are found by throwing away variables having small chi-squares.

Of course, the user could protect himself from falsely declaring a model significant with the forward stepwise procedure by judging the model chi-square as if it had p degrees of freedom where p is the total number of candidate variables. However, if an intermediate model were not significant, one would still not know if the overall model would have been judged significant if it were tested.

Mantel (1970) displays other advantages of backward elimination from the standpoint of getting a "best" sub-model.

REFERENCES

Bartolucci AA and Fraser MD (1977). Comparative step-up and composite tests for selecting prognostic indicators associated with survival. Biometrical J. 19, 437-448.

Cox DR (1972). Regression models and life tables (with discussion). J. Roy. Statist. Soc. B. 34, 187-220.

Harrell FE (1979). The LOGIST procedure. SAS Technical Report S-110, SAS Institute, Inc.

Harrell FE (1979). The PHGLM procedure. SAS Technical Report S-109, SAS Institute, Inc.

Harris PJ, Harrell FE, Lee KL, Behar VS and Rosati RA (1979). Survival in medically treated coronary artery disease. Circulation 60, 1259-69.

Kalbfleisch JD (1978). Likelihood methods and nonparametric tests. J. Am. Statist. Assoc. 73, 167-170.

Lawless JF and Singhal K (1978). Efficient screening of nonnormal regression models. Biometrics 34, 318-327.

Mantel N (1970). Why stepdown procedures in variable selection. Technometrics 12, 621-625.

Rao CR (1973). Linear Statistical Inference and its Applications, second edition, p. 418-419. New York: Wiley.

ACKNOWLEDGEMENTS

This work was supported in part by NIH grants LM-07003, LM-03373, and LM-00042 from the National Library of Medicine, HRA-230-76-0300 from the Health Resources Administration, HS-03834 from the National Center for Health Services Research, HL-17670 from NHLBI and grants from the Prudential Insurance Company and the Kaiser Family Foundation.