Examining the Reliability of Logistic Regression Estimation Software

Lijia Mo[1], Jason S. Bergtold[2], and Allen M. Featherstone[3]
[1]Ph.D Candidate, [2]Assistant Professor, [3]Professor
Department of Agricultural Economics
Kansas State University
342 Waters Hall
Manhattan, KS 66506-4011

Poster prepared for presentation at the Agricultural & Applied Economics Association 2010 AAEA, CAES, & WAEA Joint Annual Meeting, Denver, Colorado, July 25-27, 2010

Examining the Reliability of Logistic Regression Estimation Software

Lijia Mo[1], *Jason S. Bergtold*[2], *and Allen M. Featherstone* [3]

[1]Ph.D Candidate, [2]Assistant Professor, [3]Professor, Department of Agricultural Economics, Kansas State University

Contact Information:
Lijia Mo
Ph.D Candidate
Department of Agricultural Economics
Kansas State University
342 Waters Hall
Manhattan, KS 66502-4011
Email: lmo2@ksu.edu

1. Introduction

Software reliability tests help to improve the quality of statistical software. Previous work has predominately examined linear and nonlinear least squares estimation procedures and found that default nonlinear algorithmic options should not be relied upon (McCullough 1998, 1999). Systematic testing of discrete choice models for econometric software has yet been undertaken.

2. Purpose and Objectives

The purpose of this research is to examine the reliability of logistic regression estimation options in econometric software packages.

Specific objectives include:

- Test the reliability of logistic regression packages, including SAS, STATA, MATLAB, R, SHAZAM, EVIEWS, MINITAB, SPSS, and LIMDEP.
- Develop and utilize benchmark datasets and certified estimated values to evaluate the accuracy and reliability of each software package.
- Evaluate software reliability under alternative nonlinear algorithmic options, including starting value, choice of algorithm/estimator and termination criteria.

3. Logistic Regression Model

$$Y_i = \left[1+\exp\left(-\beta'\phi(\mathbf{X}_i)\right)\right]^{-1} + u_i$$

where Y_i is a binary dependent variable,
$E(Y_i \mid \mathbf{X}_i = \mathbf{x}_i) = \mathbf{P}(Y_i = 1 \mid \mathbf{X}_i = \mathbf{x}_i) = \left[1+\exp\left(-\beta'\phi(\mathbf{X})\right)\right]^{-1}$
is the conditional mean linear in the parameters, $\phi(\mathbf{X}_i)$ is a vector of functions of the elements of **X** (e.g. linear terms, squares, interaction terms, etc.) and u_i is a mean zero error term.

The logistic regression model is usually estimated using the method of maximum likelihood via the log likelihood function. Given the nonlinear nature of the estimation process iterative numerical methods must be used for estimation. These methods include:Newton-Raphson (NR), Fisher (or Method of) Scoring, Berndt, Hall, Hall, Hausman (BHHH), BFGSQuasi-Newton, and Conjugate Gradient Methods.

4. Data and Methods

Creation of benchmark datasets and certified parameter values follows that set forth by the National Institute for Standards and Technology (NIST) (2003). Assessment of reliability follows procedures set forth by McCullough (1998, 1999). The approach is outline below.

Benchmark Datasets

Datasets where randomly generated following simulation procedures in Bergtold et al. (2010) using MATLAB (version 7.5). Two datasets are presented here:

Dataset 1: Logistic regression model with four normally distributed explanatory variables with 1000 observations. The index function or predictor is linear in the variables and all explanatory variables exhibit a high degree of multicollinearity.

Dataset 2: Logistic regression model with four normally distributed explanatory variables with 5000 observations. The index function or predictor is linear in the variables, multicollinearity is present, and the $\mathbf{P}(Y_i=1) = 0.0005$.

Certified Value Estimation

Certified values for parameters and standard errors of the logistic regression models associated with datasets 1 and 2 where obtained following procedures used by the National Institute of Standards and Technology (2003). Mathematica 7.0 was used for certified value estimation using the method of maximum likelihood. Certified values were verified by estimating using 3 separate nonlinear algorithms, analytical derivatives, and 40 significant digits of precision.

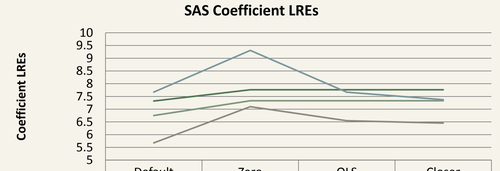
Software Reliability and Assessment

Results for two software packages are presented here:

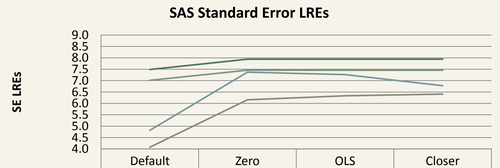
- **SAS**: PROC LOGISTIC and PROC QLIM
- **LIMDEP**: LOGIT/BLOGIT

Assessment of reliability of parameter and standard error estimation is based on the log relative error (LRE), which measures the number of significant digits relative to the certified value. The higher the value the closer the estimate (McCullough, 1998, 1999). Four starting values were used in estimation: package default, zero, OLS, and "close" starting point.

Dataset 1

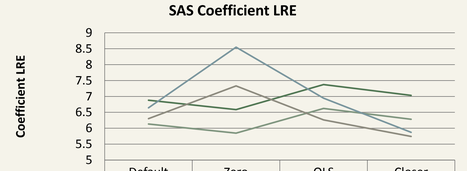


	Default	Zero	OLS	Closer
SASlgtMax	7.3204667	7.7607893	7.7607962	7.7618321
SASlgtMin	6.7488345	7.3302582	7.3302641	7.3313024
SASQlimMax	7.6718713	9.3072959	7.665547	7.3705179
SASQlimMin	5.6762522	7.0975475	6.5461953	6.4507549

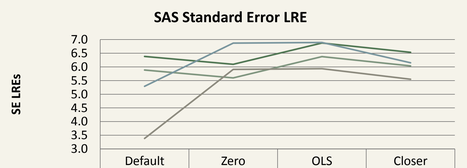


	Default	Zero	OLS	Closer
SASlgtMax	7.4886162	7.9376525	7.9376576	7.9386936
SASlgtMin	7.0121657	7.4570813	7.4570871	7.4581253
SASQlimMax	4.8163593	7.3695372	7.2648133	6.7789257
SASQlimMin	4.0769192	6.1548141	6.3342889	6.4093923

Dataset 2



	Default	Zero	OLS	Closer
SASlgtMax	6.8814373	6.584198	7.37377	7.0323819
SASlgtMin	6.1334718	5.8471002	6.6232998	6.2831373
SASQlimMax	6.6414813	8.5464934	6.9455261	5.8718384
SASQlimMin	6.3014306	7.3311223	6.2619097	5.7436787



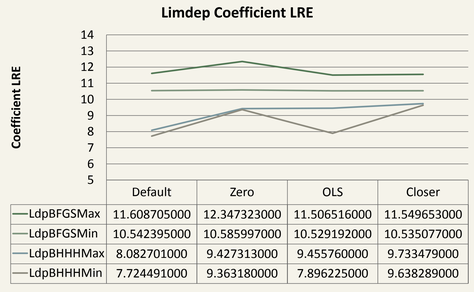
	Default	Zero	OLS	Closer
SASlgtMax	6.3813797	6.0944941	6.8709837	6.5310129
SASlgtMin	5.8845451	5.5981931	6.3743352	6.0341999
SASQlimMax	5.2883499	6.8686259	6.8920382	6.1511284
SASQlimMin	3.382289	5.9077592	5.9328852	5.5482874

Figure 1: Minimum LREs for Parameter Estimates for SAS using PROC Logistic and PROC QLIM for Datasets 1 and 2

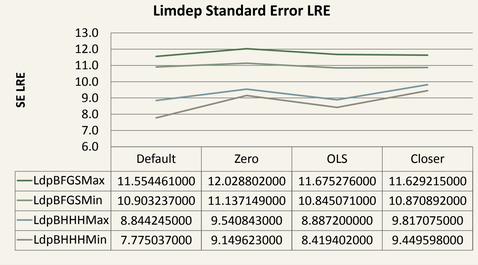
5. Results

A reliable estimate is when the minimum parameter estimate for the coefficient or standard error is greater than or equal to 4 (McCullough 1998, 1999).

- Presented findings suggest that SAS and LIMDEP can provide reliable estimates for logistic regression analysis, but users may need to aware of starting values and how they can affect estimation results. Furthermore, choice of algorithm and procedure for performing estimation my affect results, as well. Users should never rely on default settings to ensure reliability of estimates.
- Performance on both datasets was adequate. Additional research in this study is examining smaller datasets , other cutoff values and nonlinear index/predictor functions. Results suggest that traditional logistic regression software is relatively reliable, but more specialized procedures (e.g. PROC GLIMMIX and PROC SURVEYLOGISTIC in SAS) may have some difficulty in handling estimates for basic logistic regression models.



	Default	Zero	OLS	Closer
LdpBFGSMax	11.608705000	12.347323000	11.506516000	11.549653000
LdpBFGSMin	10.542395000	10.585997000	10.529192000	10.535077000
LdpBHHHMax	8.082701000	9.427313000	9.455760000	9.733479000
LdpBHHHMin	7.724491000	9.363180000	7.896225000	9.638289000



	Default	Zero	OLS	Closer
LdpBFGSMax	11.554461000	12.028802000	11.675276000	11.629215000
LdpBFGSMin	10.903237000	11.137149000	10.845071000	10.870892000
LdpBHHHMax	8.844245000	9.540843000	8.887200000	9.817075000
LdpBHHHMin	7.775037000	9.149623000	8.419402000	9.449598000

Figure 2: Minimum LREs for Parameter Estimates for LIMDEP using Quasi-Newton and BHHH Algorithms for Dataset 2

References:

Bergtold, J., A. Spanos, and E. Onukwugha (2010) Bernoulli Regression Models: Revisiting the Specification of Statistical Models with Binary Dependent Variables.. Journal of Choice Modeling, Vol. 3, No. 2, in press.

McCullough, B. D. (1998). Assessing the Reliability of Statistical Software: Part I. The American Statistician, Vol. 52, No. 4 Nov.: pp. 358-366.

McCullough, B. D. (1999). Assessing the Reliability of Statistical Software: Part II. The American Statistician, Vol. 53, No. 2 May: pp. 149-159.

National Institute of Standard and Technology (NIST). (2003) Statistical Reference Datasets. Available at: http://www.itl.nist.gov/div898/strd .