Examining the Reliability of Logistic Regression Estimation Software

Lijia Mo¹, Jason S. Bergtold², and Allen M. Featherstone³
¹Ph.D Candidate, ²Assistant Professor, ³Professor
Department of Agricultural Economics
Kansas State University
342 Waters Hall
Manhattan, KS 66506-4011


Copyright 2010 by Lijia Mo, Jason S. Bergtold, and Allen M. Featherstone. All right reserved. Readers may make verbatim copies of this document for non-commercial purpose by any means, provided that this copyright notice appears on all such copies.
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 LMDEP.
- 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

\[ Y_i = [1 + \exp(-\beta \phi(X_i))]^{-1} + \epsilon_i \]

where \( Y_i \) is a binary dependent variable, \( E(Y_i | X_i = x_i) = P(Y_i = 1 | X_i = x_i) = 1 + \exp(-\beta \phi(X_i)) \) is the conditional mean linear in the parameters, \( \phi(X_i) \) is a vector of functions of the elements of \( X_i \) (e.g. linear terms, squares, interaction terms, etc.) and \( \epsilon_i \) 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, Hausman (BHHH), BFGSQuasi-Newton, and Conjugate Gradient Methods.

4. Data and Methods


**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 \( P(\phi_2) = 1 = 0.0005. \)

**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
- **LMDEP**: 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.

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 LMDEP 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 may 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 cutooff 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.

References:

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

**Figure 2**: Minimum LREs for Parameter Estimates for LMDEP using Quasi-Newton and 88HII Algorithms for Dataset 2

**Table 1**: Dataset 1

**Table 2**: Dataset 2