Accepted Version of Manuscript published in *Computations Statistics*. Cite as:

Bergtold, J.S., Pokharel, K.P., Featherstone, A.M. *et al.* On the examination of the reliability of statistical software for estimating regression models with discrete dependent variables. *Comput Stat* **33,** 757–786 (2018). https://doi.org/10.1007/s00180-017-0776-5

On the Examination of the Reliability of Statistical Software for Estimating Regression Models with Discrete Dependent Variables

Jason S. Bergtold[a], Krishna P. Pokharel[b], Allen M. Featherstone[c], and Lijia Mo[d]

[a] Associate Professor, Department of Agricultural Economics, Kansas State University, 307 Waters Hall, Manhattan, KS 66506-4011 (E-mail: bergtold@ksu.edu, Phone: 785.532.0984) (Corresponding Author).

[b] Graduate Research Assistant, Department of Agricultural Economics, Kansas State University, 342 Waters Hall, Manhattan, KS 66506-4011 (E-mail: kpokharel@ksu.edu, Phone: 785.532.6702).

[c] Professor, Department of Agricultural Economics, Kansas State University, 342 Waters Hall, Manhattan, KS 66506-4011 (E-mail: afeather@ksu.edu, Phone: 785.532.4441).

[d] Graduate Research Assistant, Department of Agricultural Economics, Kansas State University, 342 Waters Hall, Manhattan, KS 66506-4011 (E-mai: lmo2@ksu.edu, Phone: 785.532.6702).

On the Examination of the Reliability of Statistical Software for Estimating Regression Models with Discrete Dependent Variables

Abstract

The numerical reliability of statistical software packages was examined for logistic regression models, including SAS 9.4, MATLAB R2015b, R 3.3.1., Stata/IC 14, and LIMDEP 10. Thirty unique benchmark datasets were created by simulating alternative conditional binary choice processes examining rare events, near-multicollinearity, quasi-separation and nonlinear transformation of variables. Certified benchmark estimates for parameters and standard errors of associated datasets were obtained following standards set-out by the National Institute of Standards and Technology. The logarithm of relative error was used as a measure of accuracy for numerical reliability. The paper finds that choice of software package and procedure for estimating logistic regressions will impact accuracy and use of default settings in these packages may significantly reduce reliability of results in different situations.

KEY WORDS: Accuracy; Benchmark Datasets; Logistic Regression; Maximum Likelihood Estimation; Econometric Software

MSC Codes: 62-04, 62J12, 62P99

On the Examination of the Reliability of Statistical Software for Estimating Regression Models with Discrete Dependent Variables

1. INTRODUCTION

The primary objectives of statistical software are to provide descriptive statistical analysis of data and estimate statistical models. Researchers use the results from statistical software packages for policy analysis, prediction, inference, etc. Researchers often assume the estimated results are reliable, meaning the results for statistical models are numerically accurate. Numerical accuracy of results, regardless of the use of the statistical software package, is one of the crucial factors for credible research. However, research has shown that different statistical packages may provide different estimated results for the same problem (McCullough and Vinod 1999). A scenario may exist where two researchers are solving the same problem with the same data and methods, but using different statistical packages and find different estimates. In such a case, either the statistical packages (or at least one of them) may be inaccurate or one of the authors (or both) did not properly estimate the statistical model or procedures (McCullough 1998; Odeh et al. 2010). In the former case, the problem can be avoided by utilizing statistical software packages that have undergone rigorous and passable assessment of the numerical accuracy of the estimation routines being used. However, in most cases, the results are considered inaccurate either due to data problems or statistical procedures rather than considering the statistical software as a possible source of error. In general, researchers assume that the built-in estimation procedures of software packages are reliable and interpret the results assuming they were correctly estimated (Odeh et al. 2010).

Researchers often focus on user-friendliness and speed of software packages, often ignoring the numerical accuracy of software, as they trust the software developer has ensured

this (McCullough 2000b). If the estimated results are not reliable, then it has strong negative implications for policy analysis, statistical inference, prediction etc., which weakens the work of applied researchers (Tomek 1993). Given the nonlinear nature and the widespread use of discrete choice models, these models could be prone to numerical issues as evidenced by McCullough and Vinod (2003) and Stokes (2004). McCullough and Vinod (2003) state "from the fact that a computer software package produces a solution to an estimation problem, it does not necessarily follow that the solution is accurate, or even that a solution exists (p. 873)." Estimates can be sensitive to the nonlinear algorithms or solvers used to estimate the results and the associated parameters, including starting points, line search procedures, the way in which derivatives are computed, and the termination/convergence criteria (McCullough and Vinod 2003; Odeh et al. 2010).[1] The two main problems for many nonlinear models are: (i) that an algorithm provides solutions while another may fail, and (ii) even if both algorithms provide solutions, one may be more numerically accurate (McCullough and Renfro 2000). Thus, reliability of estimation results and subsequent prediction and inference depends on the numerical accuracy of the estimation procedure and associated nonlinear algorithm used.

The purpose of this paper is to (i) develop a suite of benchmark datasets for testing logistic regression procedures using maximum likelihood estimation, and (ii) examine the numerical reliability of statistical software packages used for the estimation of logistic regression models. The procedures outlined in the paper were used to develop thirty unique and varying benchmark datasets to provide a robust method for testing the reliability of

[1] In any empirical work, researchers do not know true values, thus cross validation of research results becomes critical for verifying the numerical reliability of estimates from nonlinear models.

logistic regression estimation procedures. These benchmark datasets were then used to assess the procedures in five commonly used statistical software packages: SAS 9.4, MATLAB R2015b, R 3.3.1, Stata/IC 14, and LIMDEP 10. Assessment of statistical software examined the numerical accuracy of the results (parameter estimates and standard errors) estimated by these packages using the benchmark datasets produced with available algorithms, multiple starting points, multiple convergence criteria, different estimation commands/procedures and whether errors are correctly identified and reported during estimation. To the authors' knowledge no systematic attempt to assess the actual statistical reliability of econometric and statistical software procedures for estimating discrete choice models based on certified results has been undertaken. This paper starts with the basic logistic regression model for this assessment and the reliability results here will likely apply and extend to other discrete choice techniques in the literature, including other binary choice models, multinomial regression models, and other logistic regression techniques.

2. BACKGROUND

Many past studies have examined the reliability of software packages including SAS, MATLAB, STATA, and LIMDEP. (e.g. Musa et al. 1987; McCullough 1998, 1999b; Kolenikov 2001; Keeling and Pavur 2007; Odeh et al. 2010). These studies have mainly focused on linear and nonlinear regression models using National Institute of Science and Technology (NIST) benchmark datasets (NIST 2014). These datasets do not consider models with discrete dependent variables. Thus, there has not been any systematic examination of the numerical reliability of software packages for discrete choice models, including logistic regression estimation routines.

Huber and Train (2001) examined similarities and differences between classical and Bayesian methods for mixed logit models and found that Bayesian approaches had benefits for numerical accuracy in small samples. However, this study did not examine the reliability of software. Oster (2002, 2003) used exact methods to compare StatXact, LogXact, Stata, Testimate, and SAS based on hardware requirements, documentation, data entry, estimation results etc. Chang and Lusk (2011) compared the maximum likelihood estimator for SAS 9.2, LIMDEP 9 (contains NLOGIT 4), and Hole's model (a user written add-in module) for STATA 11 to examine the accuracy of the mixed logit model estimation using Monte Carlo simulation methods. They used the default algorithm and tolerance level for each software package. Results showed that the solution procedures reached convergence, except for Hole's model when sample size was 200 observations. In addition, all packages provided accurate estimates of willingness-to-pay measures, but bias was observed when the sample size was less than 200 observations. Chang and Lusk (2011) essentially compared the performance of different software packages against each other, but did not use certified values from a certified benchmark dataset to assess the reliability of the procedures examined. McKenzie and Takaoka (2003) found that LIMDEP 8.0 was unable to indicate a problem with an unidentified probit/logit model, given that the parameters of the probit model being examined were not able to be identified (i.e. estimated).

Logistic regression is one of the most widely used discrete choice models in a large number of disciplines. The nonlinear benchmark tests from NIST examining the accuracy of nonlinear least squares problems could potentially provide an assessment of maximum likelihood estimators, however, they are not designed for them (Altman et al. 2004). Cameron and Trivedi (2009) find that the nonlinear least squares (NLS) estimator for models arising

from the simple exponential family (i.e. of which the logistic regression is a member of) when the conditional mean is nonlinear and heteroskedastic errors are present does not provide accurate estimates. This arises from the fact that the NLS estimator can be significantly less efficient than MLE. In addition, the standard error estimates may not be valid due to the presence of heteroskedastic errors. Thus, logistic (and other discrete choice) regression model estimation procedures (using maximum likelihood estimation) in software packages may require the use of additional benchmark testing beyond those traditionally used to test nonlinear least squares procedures, because the logistic regression model is a specialized model of the maximum likelihood estimator (McCullough 1999a; NIST 2014).

The two main contributions of this paper are the development of a unique set of benchmark datasets that can be used to test additional and newer versions of statistical software and a comparative examination of the numerical reliability of alternative software packages. The results from this study show the strengths and weaknesses of software packages that provide logistic regression estimation procedures. The software vendors may address inadequacies if they exist. Due to past reliability studies, software vendors have fixed problems in newer versions of their software. For example, results obtained for nonlinear regression estimates from LIMDEP 8.0 are better than LIMDEP 7.0 (Odeh et al. 2010). In addition, researchers can use the information to choose a software package based on the properties of their data. The benchmark datasets developed highlight some of main concerns encountered in estimation of logistic regression models (e.g. low cut-off or probability of occurrence, small samples, multicollinearity, and quasi-separation).

3. LOGISTIC REGRESSION MODEL

Let Y_i be a Bernoulli random variable with mean $E(Y_i) = p = \mathbf{P}(Y_i = 1)$ and variance equal to $p\ (1 - p)$. Let $\mathbf{X}_i$ be a $(K \times 1)$ vector of explanatory variables. Then the conditional probability (or mean) is given by $\mathbf{P}(Y_i = 1|\mathbf{X}_i) = [1 + exp\{-\eta(\mathbf{X_i}; \boldsymbol{\beta})\}]^{-1}$, where $\eta(\mathbf{X_i}; \boldsymbol{\beta})$ is referred to as the predictor or index function. This can be represented as a statistical model:

$$Y_i = [1 + exp\{-\eta(\mathbf{X_i}; \boldsymbol{\beta})\}]^{-1} + u_i, \qquad (1)$$

where u_i is a zero mean IID random error term. The functional form of $\eta(\mathbf{X_i}; \boldsymbol{\beta})$ is usually chosen to be linear in the explanatory variables, parameters or both, but is dependent upon the distributional properties of the explanatory variables (Bergtold et al. 2010; Kay and Little 1987).

While a number of alternative estimators for the logistic regression model are available, the most commonly used is the maximum likelihood estimator (MLE), given the distributional properties of the dependent variable are known. The log-likelihood function used to estimate the logistic regression model given by (1) is:

$$L(\boldsymbol{\beta}; Y_i, \mathbf{X_i}) = \sum_i Y_i \ln\big(F(\mathbf{X_i}; \boldsymbol{\beta})\big) + (1 - Y_i)\ln\big(1 - F(\mathbf{X_i}; \boldsymbol{\beta})\big), \qquad (2)$$

where $F(\mathbf{X_i}; \boldsymbol{\beta}) = [1 + exp\{-\eta(\mathbf{X_i}; \boldsymbol{\beta})\}]^{-1}$. To find the estimates of $\boldsymbol{\beta}$, the log-likelihood function given by equation (2) is maximized given the data. Because the log-likelihood function is non-linear in the parameters and a closed-form solution for the MLE estimator of $\boldsymbol{\beta}$ may not be available, iterative numerical methods are used to maximize the log-likelihood function and obtain parameter estimates. These methods require the use of a nonlinear algorithm that has various components: starting point, choice of search procedure, gradient calculation procedure, and termination criteria (Train 2003). Each of these chosen components may have an effect on

the numerical accuracy and reliability of model estimation. An algorithm either uses the gradient of the log-likelihood function:

$$\nabla_{\beta} L = \sum_i \big(Y_i - F(\mathbf{X_i}; \boldsymbol{\beta})\big)\mathbf{X_i}, \qquad (3)$$

or a numerical approximation of it during optimization. Once an optimal solution or estimate of $\boldsymbol{\beta}$ is obtained, the asymptotic covariance matrix of the MLE estimator for $\boldsymbol{\beta}$ may be calculated using the inverse of the Hessian of the log-likelihood function:

$$\mathrm{Cov}\big(\widehat{\boldsymbol{\beta}}\big) = \big(\sum_{\boldsymbol{i}} F\big(\mathbf{X_i}; \widehat{\boldsymbol{\beta}}\big)\,(\mathbf{1} - F\big(\mathbf{X_i}; \widehat{\boldsymbol{\beta}}\big)\mathbf{X_i}\mathbf{X_i'}\big)^{-\mathbf{1}}. \qquad (4)$$

This estimator can be calculated using the analytic solution given by equation (4) or using a numerical approximation of it (Cameron and Trivedi 2009).

3. BENCHMARK DATA SETS

This section provides the methods used to define a set of benchmark datasets that can be used to test the reliability of logistic regression estimation procedures in statistical software packages. McCullough and Renfro (1998) state that benchmark datasets should be chosen to answer two questions: (1) "What models *can* software packages estimate?" and (2) "What models *should* they be able to estimate?" (p. 60). Following McCullough and Renfro (1998), a suite of benchmark datasets are developed to estimate logistic regression models with predictor (index) functions that are linear in the parameters ($\boldsymbol{\beta}$), under the assumption that standard statistical software packages should be able to estimate such models. Furthermore, to examine the extent to which alternative statistical software packages can reliably estimate such models, a suite of benchmark datasets are developed using different models that take into account a number of issues that arise in estimation of logistic regression models including functional form, low probability of occurrence (cut-off value), near-multicollinearity between

regressors, and quasi-separation.[2] The development of thirty benchmark datasets that examine different estimation issues involving estimation of logistic regression models provides a starting point to examine the reliability of software packages that provide discrete choice model estimation routines and to help establish a set of standards for statistical software implementing such procedures.

3.1 Data Generation

Arnold et al. (1999) show that the existence of the logistic regression model depends on the compatibility between the conditional distribution $f(Y_i|\mathbf{X_i};\boldsymbol{\beta})$ and the inverse conditional distribution$f(\mathbf{X_i}|Y_i;\boldsymbol{\theta})$. That is: $f(Y_i|\mathbf{X_i};\boldsymbol{\beta})f(\mathbf{X_i};\boldsymbol{\vartheta}) = f(\mathbf{X_i}|Y_i;\boldsymbol{\theta})f(Y_i;p) = f(Y_i,\mathbf{X_i};\boldsymbol{\varphi})$, where $f(\mathbf{X_i};\boldsymbol{\vartheta})$ is the multivariate marginal distribution of $\mathbf{X_i}$, $f(Y_i;p)$ is the marginal distribution of Y_i; $f(Y_i,\mathbf{X_i};\boldsymbol{\varphi})$ is the multivariate distribution of Y_i and $\mathbf{X_i}$; and $\boldsymbol{\vartheta}$ and $\boldsymbol{\varphi}$ are appropriate sets of parameters. Thus, data can be generated using the conditional distribution $f(Y_i|\mathbf{X_i};\boldsymbol{\beta})$ and marginal distribution $f(\mathbf{X_i};\boldsymbol{\vartheta})$ or the inverse conditional distribution $f(\mathbf{X_i}|Y_i;\boldsymbol{\theta})$ and marginal distribution $f(Y_i;p)$. Bergtold et al. (2010) and Scrucca and Weisberg (2004) generate data for the logistic regression model using the inverse conditional distribution and marginal distribution of Y_i. This allows for a more parsimonious method of generating data as it provides a systematic way to specify the predictor (index) function, capturing potential nonlinear terms that are based on the distributional assumptions concerning $\mathbf{X_i}$. This arises from the result that $\eta(\mathbf{X_i};\boldsymbol{\beta}) = \ln\left(\frac{f(\mathbf{X_i}|Y_i=1;\boldsymbol{\theta})}{f(\mathbf{X_i}|Y_i=0;\boldsymbol{\theta})}\right) + \ln\left(\frac{p}{1-p}\right)$ (Bergtold et al.

[2] Quasi-separation (also known as quasi complete separation) occurs when a collection of covariates can almost completely separate the outcome groups in the discrete choice model. That is, only a few observations are left that make the outcome groups overlap or in terms of discriminant analysis, the discriminant can almost perfectly delineate the outcome groups (Hosmer et al. 2013).

2010; Kay and Little 1987). Following the inverse conditional approach, data are generated in two steps. First p is specified and a random sequence of data is generated for Y_i based on $f(Y_i; p)$, which is a Bernoulli distribution. Second, using the randomly generated Y_i, data is generated for $\mathbf{X}_i$ using the assumed distribution for $f(\mathbf{X}_i|Y_i; \boldsymbol{\theta})$ (Bergtold et al. 2010).

Thirty benchmark datasets are generated using the inverse conditional approach for reliability testing purposes. Twenty-nine of these datasets are randomly generated using MATLAB as described above. The final dataset is empirical survey data from a survey examining conservation practice adoption on farms in Alabama. This final dataset was used to provide an example with a significant number of covariates. Table 1 summarizes the thirty benchmark data sets and associated parameters used for data generation. The table provides the dataset name, cutoff point, functional form of the predictor, level of multicollinearity between covariates, number of observations, and amount of variation within covariates (if relevant). The datasets vary in difficulty by changing the conditions under which they were generated. These conditions include: changing the $P(Y = 1)$ (i.e. the cutoff value); varying the amount of noise or variation in the data (through the variance parameters of the dataset); varying the degree of near multicollinearity between covariates; introducing different nonlinear transformations with respect to the explanatory variables into the predictor/index function; and introducing quasi-separation into a dataset. For example, a number of datasets were generated with explanatory variables that were nearly collinear by changing the degree of correlation between covariates from 0.75 to 0.995. Collinearity is a significant challenge in environmental, economic and ecological research and different software handles it differently. Similarly, for cut-off points, we generated datasets with cutoffs (i.e. $P(Y_i = 1)$) from 0.015% to 19% with different size datasets. The cutoff value for the cutoff4 benchmark dataset was set at

0.015%, which may arise in modeling of extreme events, such as credit default or credit card fraud. Cases may exist where data has a significant amount of variation. The multivariate3 dataset is generated with covariates that have significantly high variability, modeled using high variances and moderate correlations between covariates. Other benchmark datasets can help to evaluate models with index functions that are nonlinear in the covariates, such as multivariate1 and multivariate2, as well as multivariate5 and multivariate6.

3.2 Estimation of Certified Parameter Values

For each benchmark dataset and associated logistic regression model, certified values for the parameters ($\boldsymbol{\beta}$), associated asymptotic standard errors, and log-likelihood function are estimated using procedures designed following those used for the Statistical Reference Datasets (StRD) from the National Institute of Standards and Technology (NIST) (NIST 2014). All benchmark estimation was completed using Mathematic 7.0 (Wolfram 2015b). Mathematica has the ability to do "extreme precise computations via arbitrary precision calculation (McCullough 2000a, p. 200)." That is, this software package can perform the optimization of the log-likelihood function with high precision and offers a number of algorithmic options. McCullough (2000a) found that earlier versions were able to perfectly replicate the certified values (reported to 11 significant digits) for the StRD nonlinear least squares benchmark datasets, by setting Mathematic 4.0 to do all calculations with 30 significant digits. The authors were able to replicate these results with Mathematica 7.0. Given this performance, Mathematica 7.0 was used to obtain the certified values for this study.

Following NIST standards, certified values for the parameters ($\boldsymbol{\beta}$), associated asymptotic standard errors, and log-likelihood function were generated manually using code

generated by the authors. To simulate and maintain high precision during estimation, all calculations were set to be calculated with 50 significant digits. For each benchmark dataset generated, the associated log-likelihood function given in equation (2) was optimized using three alternative algorithms in Mathematica 7.0 at two alternative starting points: Broyden-Fletcher-Goldfarb-Shanno Quasi-Newton algorithm; Conjugate Gradient algorithm using the Fletcher and Reeves update; and a derivative free approach called the principal axis method (Wolfram 2015a). The two starting points for parameter estimates are (i) the null vector with the value for the intercept replaced by the unconditional log odds, and (ii) the estimated parameters from the associated linear probability model. For derivative based algorithms, the gradient was calculated analytically using equation (3), providing greater precision than numerical derivative approximation methods (McCullough 1998). Parameter estimates and the optimal value of the log-likelihood function were certified when estimates from two of the three algorithms used matched, having 11 or more significant digits in common. Certified asymptotic standard errors were then estimated analytically at the same level of precision in Mathematica 7.0 (i.e. 50 digits of carry-through) using equation (4) once certified parameter estimates were obtained. Full descriptions of the benchmark datasets with certified values and the generated data are available for all 30 benchmark datasets as a supplementary file to this paper.

4. RELIABILITY ASSESSMENT METHODS

We assess five commonly used software packages in statistics and econometrics that are used to estimate logistic regression models: (i) SAS version 9.4, SAS Institute Inc., release July 10, 2013; (ii) MATLAB version 8.6.0.267246 (R2015b), MathWorks, release September 3, 2015;

(iii) R version 3.*3.1*, R Core Team, release June 21, 2016; (iv) STATA/IC version 14, StataCorp LP, release April 7, 2015; and (v) LIMDEP version 10, Econometric Software Inc. (which contains NLOGIT version 5), release June 8, 2012. Our analysis was performed on a personal computer with 64-bit operating system on Microsoft Windows 7 Professional Service Pack 1. Each of these packages and the associated logistic regression estimation routines are summarized in Table 2. For both SAS and STATA, multiple procedures were examined to estimate logistic regression models. The choice of these estimation procedures within the packages was based upon the ability to output actual estimates at the level of precision computed by the program. Thus, the sets of procedures examined are not necessarily exhaustive of the procedures that can estimate logistic regression models. For example, in SAS, other logistic regression procedures exist, including PROC SURVEY LOGISTIC and PROC GENMOD.

To examine the numerical reliability of statistical software, the logarithm of relative error (LRE) is used as a measure of accuracy following McCullough (1998) and other reliability studies (e.g. Odeh et al. 2010). The LRE is calculated as:

$$LRE = -log_{10}\left[\frac{|q-c|}{|c|}\right]$$

where q denotes the estimated value and c stands for the certified (correct) value. If the certified value is zero, the LRE measure is undefined. In such a case, the log absolute error ($LAE = -\log_{10}|q|$) can be used in its place. The first nonzero digit and the digits succeeding it are considered. The LRE measures the number of significant digits of the estimated results in comparison to the certified value. For example, an LRE value of 6.5 indicates that the estimated result is accurate to six significant digits. The higher the LRE score, the more accurate the estimate. If a program reports a negative LRE value, meaning the estimated result

is far from the certified value, the LRE is reported as zero (McCullough and Wilson 1999). For non-linear models, McCullough (1998) uses a minimum LRE score of 4 as a threshold to indicate if a software program is accurate. If more accuracy is needed, the results can be reinterpreted with a higher minimum LRE threshold (e.g. 6). For each benchmark dataset, a number of parameter estimates and standard errors are estimated. The LREs for each parameter and standard error are calculated and the minimum LRE value obtained for the estimated parameters and associated standard errors are reported. The minimum LRE, which represents the least accurate estimated parameter, is reported to indicate how far the estimated solution was from obtaining the certified solution. Reporting an average or a range of values may understate how inaccurate an estimated solution may be. Full reliability results are available from the authors upon request.

Users estimating logistic regression models may consider changing software settings/options, such as choice of algorithm, choice of different starting points, lowering tolerance levels for termination criteria, and use of analytic derivatives to improve accuracy. A program that fails to give the minimum level of accuracy may give less accurate results for more difficult problems. Thus, for each software package and logistic regression estimation command examined, the benchmark dataset models are estimated using the (i) default settings (i.e. the naïve approach) and (ii) user determined settings. Optional user settings were determined by adjusting the choice of algorithm and level of convergence required for termination to achieve the highest obtainable minimum LRE using the following procedures (when these could be changed by the user). For each statistical procedure and software package, each benchmark dataset was estimated for each available algorithm for each starting point. For example, each benchmark data set was estimated using PROC LOGISTIC in SAS

9.4 using both the Fisher's Scoring and Newton-Raphson algorithms for each starting point. To determine the optimal user setting for the tolerance level, for each benchmark dataset, algorithm and starting point combination, the associated logistic regression model was estimated starting at the default tolerance level. The model was then re-estimated, decreasing the tolerance level by a magnitude or 1e-3 each time. This was repeated until the minimum LRE did not change for two consecutive reductions in the tolerance level. For example, using PROC LOGISTIC in SAS 9.4, the default tolerance level is 1e-8 for the gradient. This tolerance level was decreased from 1e-8 to 1e-11, then to 1e-14, and so on, until the minimum LRE for the parameter estimates remained constant. Table 2 summarizes the software and settings available that were examined in this study for each software package, including estimation commands, estimation algorithms available, ability to change starting points, and convergence or termination criteria for the estimation algorithm. If a package's default settings give lower LRE values than optional user setting, then it indicates that default settings provided less accurate estimation results compared to optional user settings, which may likely be the case for estimation of logistic regression models (McCullough and Vinod 2003).

A set of starting points can be changed by users in most of the logistic regression packages examined (Table 2). The one exception here is the GLMFIT command in MATLAB that determines its own starting point. Since convergence of algorithms can be sensitive to choice of starting points for nonlinear problems, alternative starting points may provide more accurate results (McCullough and Vinod 2003). When the starting points for estimation routines can be changed, accuracy for each benchmark dataset and model was examined for both starting points provided in the benchmark datasets. The two starting points are (i) zeros

with the intercept equal to the unconditional odds ratio of the dependent variable and (ii) estimates from the corresponding linear probability model (in the parameters).

There are many numerical algorithms that can be used for maximizing the log-likelihood function during maximum likelihood estimation of the logistic regression model. The most widely used optimization method (algorithm) is the Newton Raphson (NR) (Train 2003). Other algorithms available include the Broyden-Fletcher-Goldfarb-Shanno (BFGS) Quasi-Newton algorithm, Berndt-Hall-Hall-Hausman (BHHH) algorithm, Davidon-Fletcher-Powell (DFP) Quasi-Newton algorithm, conjugate gradient methods, and Fisher Scoring (FS) algorithms, among others (see Bazaraa et al. 2006; Greene 2002). Each algorithm requires stopping or termination criteria to indicate that a local optimal solution has been obtained. Researchers many times have the option of choosing among a number of nonlinear algorithms and convergence criteria for the chosen algorithm. For example, in the *PROC LOGISTIC* and *PROC QLIM* statements (commands) in SAS (SAS Manual 2009), the default gradient convergence is set equal to $1E-8$, but researchers can change it, but in some packages, including MATLAB and R, the convergence criteria cannot be changed. A number of alternative convergence criteria may be considered. These include: (i) $|L(\beta_{n+1}) - L(\beta_n)| < \varepsilon$; (ii) $\max(|\beta_{n+1} - \beta_n|) < \varepsilon$; (iii) $g^{\mathrm{T}}(-H^{-1})g < \varepsilon$ where g is the gradient and H is the Hessian of the log-likelihood function; and (iv) $||g(\beta_n)|| < \varepsilon$, where ε is a very small number, usually less that $1E$-4. For all algorithms assessed, derivatives were calculated using numerical methods (differencing).

The results (comparison of minimum LRE scores for different setting) that will be generated for each software package provides information about the strengths and weaknesses of the different statistical software that gives flexibility to researchers to choose appropriate

software packages and commands; adjust settings based on their problem situation; and allow software vendors to address potential reliability issues in newer versions of their software.

5. NUMERICAL RELIABILITY ASSESSMENT RESULTS

Thirty logistic regression models were estimated using the benchmark datasets in each software package. Estimation was conducted for two alternative starting points using default and optional user settings (as determined by the authors) in each statistical software package. If a software package has more than one procedure to estimate logistic regression models, then results are reported for each procedure examined. Results for the minimum LRE for both the default and optional user settings for both sets of starting points are reported in Tables 3 to 6. Tables 3 (parameter estimates) and 5 (standard error estimates) present the results for the default settings for each procedure and statistical package. Tables 4 (parameter estimates) and 6 (standard error estimates) present the results using user optional settings. For MATLAB and R, only default settings are examined as users have limited options for controlling the estimation algorithms. A value of NS indicates that an algorithm did not converge or no optimal solution was obtained for the particular setting (i.e. the model did not estimate). Following the literature, an LRE score greater than or equal to 4 is required to meet the minimum standard of reliability of nonlinear regression estimation following McCullough (1998). In some cases, a user may want higher reliability. In that case, they may want to consider higher LREs, such as 6, to assess software reliability.

For parameter estimates obtained using default settings, the LOGIT command in STATA (for both starting points, D1 and D2), GLM in R (for starting point one), and GLMFIT in MATLAB met a minimum LRE criteria of 4 for 26 of the thirty benchmark

datasets (Table 3). Similarly, PROC LOGISTIC in SAS estimated parameters reliably on 19 and 17 of the 30 benchmark models for each set of starting points. The LOGIT command in LIMDEP met the minimum LRE criteria for 27 and 26 of the benchmark models using the default setting for each set of starting pointss. Similar results were found for the minimum LRE scores of the estimated standard errors (Table 5).

In general, the optional user settings were able to improve algorithmic performance and reliability. In some cases, it allowed a procedure for a given statistical package to meet the minimum LRE score. For example, consider the Cutoff5 benchmark dataset and model estimated using PROC LOGISTIC in SAS. For both starting points, the minimum LRE for the parameter and standard error estimates were 2.8 (Table 3) and 4.3 (Table 5) respectively. When optional user settings were used by lowering the tolerance criteria, the minimum LRE for both starting points for the parameters and standard errors increased to 6.9 (Table 4) and 8.5 (Table 6), respectively.

The results for each statistical package are examined and discussed in more detail below. If any warning or error messages were obtained, they are also reported in this section. A nonlinear procedure may fail in at least one of two ways. The procedure could report an error message after a failed attempt to solve the problem or the procedure could incorrectly solve the problem and report the result without providing an error message. The former issue can be considered a miserable failure of the procedure, whereas the second one is more likely a disastrous failure of the procedure (Murray 1972).

5.1 Stata 14

All thirty datasets were estimated with the LOGIT, BINREG, and GLM procedures in STATA 14. Users can change tolerance (convergence) criteria, estimation algorithm, and starting points for each procedure (Table 2).

Using default settings for model estimation, STATA 14 met the minimum LRE criteria for 26 of the thirty benchmark datasets for the LOGIT and GLM procedures for both starting points. However, the BINREG command only met the minimum LRE criteria of 4 for 22 of the datasets for both starting points. Similarly, for the standard error estimates, the minimum LRE criteria was met for 27 of the datasets for the LOGIT and GLM procedures for both starting points. The BINREG procedure only met the minimum LRE criteria for 23 of the datasets for both starting points.

After examining all potential combinations of user optional settings, the logistic regression commands met the minimum LRE criteria for 26 of the thirty benchmark datasets for both starting points. For standard error estimates, LOGIT, BINREG and GLM performed reliably on 27, 26 and 25 of the benchmark datasets for starting point one, and 25, 22 and 21 of the benchmark datasets for starting point two, respectively. On average, the user determined setting provided a small increase in reliability over the default settings. For the BINREG procedure though, using the user determined settings allowed for reliable estimation of the Cutoff1, Cutoff 3, Cutoff 4 and Cutoff 7 benchmark models. Under default settings, the algorithm used to estimate each of these benchmark models did not converge to a solution.

The user-defined algorithm and tolerance settings that provided the lowest LRE scores varied by benchmark dataset, starting point and procedure used. There was no consistency in the setting that provided the most reliable results. For example, for the Multicollinearity3

benchmark dataset, the optimal user settings for the LOGIT procedure were found using the DFP algorithm for starting point 1 and the BFGS algorithm for starting point 2. For the BINREG and GLM procedures, the lowest LRE was obtained using the BFGS algorithm for starting point 1 and the NR algorithm for starting point 2. The default tolerance setting provided the same reliability as lower tolerance levels for all procedures and starting points. Thus, it is recommended that researchers examine different algorithms and associated parameter settings in STATA 14 to ensure that they are obtaining the most reliable estimation results.

At times, estimation algorithms in the different procedures were not able to reliably estimate models. For the Multicollinearity1, Multivariate4, Cutoff6 and Quasisep2 benchmark datasets, use of the default and user determined settings for each procedure reported low LRE values (the number of significant digits was less than 4 for parameter estimates) for the procedures assessed. The estimation routines converged to a solution that were inaccurate. No error messages were reported during estimation of these models. In contrast, when a model failed to estimate (e.g. a NS result in Tables 3 and 4) an error message was provided indicating the algorithm did not converge to a solution.

5.2 Matlab 2015b

All the benchmark models were estimated with GLMFIT in the statistics toolbox. Users are not able to change algorithm, convergence criteria, or starting points in the procedure. Thus, a limitation of the GLMFIT procedure in MATLAB is that it only provides the use of one algorithm (Newton-Raphson) to estimate logistic regression models. This software package reliably estimated 26 and 28 models out of the thirty benchmark datasets for the parameter

estimates and associated standard errors, respectively. MATLAB failed to reliably estimate the Multicollinearity1, Multivariate4, Cutoff6 and Quasisep2 datasets and did not provide any error messages during estimation.

5.3 R 3.3.1

This software package uses the GLM command for the estimation of logistic regression models. For starting point one with default and user optional settings, 26 models met the minimum LRE criteria of four. For starting point two for both default and user optional settings, 25 models met the minimum criteria. However, for standard errors only 23 datasets met the LRE criteria for both starting points. R provided a warning message for Multivariate5 for the second set of starting points: "fitted probabilities numerically 0 or 1 occurred," indicating a problem with estimation. R provides a limited number of user options to control estimation. For the GLM command, users only can change starting points. The only estimation procedure tested here was iterated reweighted least squares (R Core Team 2013).

5.4 Limdep 10

LIMDEP uses the LOGIT (BLOGIT) command to estimate logistic regression models (Econometric Software, Inc. 2012). This software package reliably estimated 27 and 26 datasets for starting points one and two using default settings. LIMDEP estimated 27 models reliably with user optional setting for both starting points. Results for standard error estimation were similar (Tables 5 and 6). At times, the user determined settings provided significantly better results than the default settings. For example, consider the Cutoff1 and Cutoff2 benchmark datasets. The package provided a minimum LRE of 5.8 and 5.9 for starting point 1,

respectively. Using the user option settings, the minimum LRE was able to be increased to 10.2 and 10.8, respectively. As with STATA 14, the best algorithm and tolerance settings were dataset and starting point specific. Thus, users should try different combinations of algorithms and associated parameter settings to ensure reliable model estimation. It should be mentioned that LIMDEP was the only package to be able to reliably estimate the models associated with the Multicolinearity1 and Quasisep2 benchmark datasets.

LIMDEP did provide error messages when algorithms failed to converge or problems were encountered. For example, for Multivariate7, estimating the associated logistic regression using the BHHH algorithm failed. LIMDEP provided a warning: "the likelihood is flat, try refitting and examining the derivatives."

5.5 Sas 9.4

Logistic regression models were estimated with PROC LOGISTIC and PROC QLIM in SAS. For parameter estimates using PROC LOGISTIC with default settings, 19 and 17 of the benchmark models estimated met the minimum LRE for starting points one and two, respectively. In contrast, 26 models were reliably estimated using user optional settings for both starting points. Of particular interest is that the marginal increase in reliability of parameter estimates using user optional settings was significant for this procedure. Both algorithmic options available in this procedure performed similarly. Significant gains in reliability were obtained by significantly lowering the tolerance level for the estimation algorithms (e.g. to 1e-15). Results for reliable estimation of standard errors was similar

between the default and user optional settings. PROC LOGISTIC reliably estimated standard errors for over 25 of the benchmark datasets for default and user optional settings.

Using PROC QLIM, 24 benchmark models were reliably estimated using the default settings for both starting points. Likewise, 27 and 26 models were estimated reliably using user optional determined settings for both starting points, respectively. PROC QLIM was able to reliably estimate standard errors for 22 and 21 of the benchmark datasets for starting points one and two in both settings, respectively. Using default settings, PROC QLIM performed more reliably than PROC LOGISTIC. With user optional settings, performance was more equivalent. In addition, PROC QLIM was the only procedure in all the software packages examined to be able to reliably estimate the Multivariate4 benchmark model.

The PROC LOGISTIC statement showed some warning or error messages. For example, for both algorithms and tolerance levels for both sets of starting points, the procedure showed the following error messages when estimating Multivariate4: "in calculating the expected values, predicted probabilities less than 1e-6 and greater than 0.999999 were changed to 1e-6 and 0.999999, respectively." In PROC QLIM, when the conjugate gradient algorithm did not converge, it showed an error message: "optimization cannot be completed." The PROC QLIM statement, in general, reported an error message when an algorithm did not converge or optimal solution (i.e. estimates) was not found.

5.6 Discussion

The results show that many of the packages were able to reliably estimate quite a few of the benchmark models, assuming a minimum LRE of 4 was reliable enough. Even if the same number of models were estimated reliably using default and user optional settings, reliability

was often improved (i.e. obtaining higher LREs) with user optional settings. In some cases, the marginal gain in performance was significant. For example, the Base benchmark dataset, the minimum LRE score for parameter estimates for starting point two using the default setting in LIMDEP was 6.6 (Table 3), but with user optional setting it increased to 10.6 (Table 5). Users should consider varying user optional settings in software packages where they are available to ensure they are obtaining the most reliable results.

If a modeler requires greater reliability, the modeler may want to examine the results using a minimum LRE of 6. Using this criteria and user optional settings, MATLAB reliably estimated parameters for 26 and standard errors for 27 of the benchmark datasets. LIMDEP reliably estimated parameters for 27 (26) and standard errors for 28 (27) of the benchmark datasets for starting point 1 (2). This comparison can be carried out on all of the statistical software packages examined. Results indicate that some procedures and statistical packages provide higher reliability when the minimum LRE is increased, but this does vary significantly across packages. If higher reliability is needed, both MATLAB and LIMDEP would be good choices. For other packages (including LIMDEP), users would likely want to change user optional settings to obtain the most reliable results.

This study indicates that the results are sensitive in nonlinear models to the choice of statistical software, algorithm, and tolerance level used during estimation. For example, the only package to reliably estimate the multicollinearity1 benchmark dataset was LIMDEP. Thus, LIMDEP may want to consider the use of this package if similar forms of multicollinearity exist in their data. Finally, it should be emphasized, that replication of the results of published articles might not be possible if researchers do not document the name of software package, command or routines (and options) used for data analysis. Since replication

is the basis of science, if results cannot be replicated, then they are likely harder to trust. This makes it difficult to assess the relevance of the research within a profession's accumulated body of knowledge or as a basis for policy analysis (McCullough and Vinod 2003).

6. CONCLUSION

The numerical reliability of estimating logistic regression models for five statistical and econometric software packages widely used by applied researchers in multiple disciplines was examined. The packages were SAS 9.4, MATLAB R2015b, R 3.3.1, STATA/IC 14, and LIMDEP 10. To test the reliability, thirty unique benchmark datasets were created following the procedures established by the National Institute of Science and Technology for their nonlinear regression benchmark datasets. Logistic regression models for the thirty benchmark datasets were estimated for different procedures in each of the statistical software packages. The reliability of the software packages and associated procedures was assessed using the minimum LRE of the parameters and asymptotic standard errors obtained, computed using the benchmark values for the parameter and standard error estimates. We followed previous literature and tested the default settings for each package and then adjusted the options in each software package to obtain an optimal user optional setting to try and obtain closer estimates to the certified values for each benchmark dataset. In reality, the certified benchmark values will be unknown, thus modelers and researchers should follow the suggestions of McCullough and Vinod (2003) to verify their results. Furthermore, the authors believe that the results should extend to other binary choice models (e.g. probit model) and multinomial models, given the similarity to logistic regression models.

Software reliability testing results suggest that logistic regression estimation procedures in the software packages were able to meet the minimum LRE requirement of 4 for

many of the benchmark datasets. It was not expected that a package is able to reliably estimate logistic regression models for all thirty datasets to be considered reliable. A minimum LRE criteria of 4 may not be reliable enough in some situations and a modeler should instead examine these results using a higher minimum LRE (e.g. of 6). It did become apparent though, that users should be careful when only using default settings. The BINREG procedure in STATA and PROC LOGISTIC in SAS both performed comparatively worse using the default settings. When user optional settings were determined by changing tolerance criteria and algorithm choice, reliability results significantly improved for both packages. Overall, user optional settings resulted in better and more accurate performance than default settings. In some cases, no default settings were available to change, limiting the flexibility of the package as in the case of MATLAB and R.

This study expands on the reliability testing of software packages for statistical estimation by considering discrete choice models using maximum likelihood estimation. Furthermore, the study provides thirty unique benchmark datasets with certified parameter and standard error estimates for reliability testing that can be used to test other statistical software packages and future versions of software

ACKNOWLEDGEMENTS: Partial support for this research was obtained from the National Science Foundation Grant: From Crops to Commuting: Integrating the Social, Technological, and Agricultural Aspects of Renewable and Sustainable Biorefining (I-STAR); NSF Award No.: DGE-0903701. The analysis and conclusions set forth are those of the authors based on the independent assessments of statistical software.

REFERENCES

Altman, M., J. Gill, and M. P. McDonald (2004), *Numerical issues in statistical computing for the social scientist*. NJ: John Wiley & Sons.

Arnold, B.C., E. Castillo. J.M. Sarabia (1999), *Conditional specification of statistical models*. New York, NY: Springer Verlag.

Bazaraa, M.S., H.D. Sherali and C.M. Shetty (2006), *Nonlinear programming: theory and algorithms*. Hoboken, NJ, John Wiley & Sons, Inc.

Bergtold, J. S., A. Spanos, and E. Onukwugha (2010), Bernoulli regression models: Revisiting the specification of statistical models with binary dependent variables. *Journal of Choice Modelling 3*(2), 1–28.

Cameron, A.C. and P.K. Trivedi (2009), *Microeconometrics: Methods and Applications*. New York, NY, Cambridge University Press.

Chang, J. B. and J. L. Lusk (2011), Mixed logit models: accuracy and software choice. *Journal of Applied Econometrics 26*(1), 167–172.

Econometric Software, Inc. (2012), Limdep 10 and Nlogit 5. http://www.limdep.com/features/documentation.php. (Accessed on August 15, 2015).

Greene, W. H. (2002). *Econometric Analysis*. Englewood Cliffs, H.J Prentice.

Hosmer, D.W., S. Lemeshow and R.X. Sturdivant (2013), *Applied Logistic Regression*, 3rd ed. Hoboken, NJ, John Wiley and Sons, Inc.

Huber, J. and K. Train (2001), On the similarity of classical and Bayesian estimates of individual mean partworths. *Marketing Letters 12*(3), 259–269.

Kay, R. and S. Little (1987), Transformations of the explanatory variables in the logistic regression models for binary data. *Biometrika* 74: 495 – 501.

Keeling, K. B. and R. J. Pavur (2007), A comparative study of the reliability of nine statistical software packages. *Computational Statistics & Data Analysis 51*(8), 3811–3831.

Kolenikov, S. (2001), Review of Stata 7. *Journal of Applied Econometrics 16*(5), 637–646.

Koro¨si, G., L. Matyas, and I. Sz´ekely (1993), Comparative review of some econometric software packages. *Journal of Economic Surveys 7*(1), 105–118.

MATLAB (2012), *(R2012a),* Natick, Massachusetts: The MathWorks Inc.

McCullough, B. (2000a), The accuracy of Mathematica 4 as a statistical package. *Computational statistics 15*(2), 279–300.

— (2000b), Is it safe to assume that software is accurate? *International Journal of Forecasting 16*(3), 349–357.

— (2003), *Some details of nonlinear estimation*. in M. Altman, J. Gill, and M. McDonald, eds., Numerical Methods in Statistical Computing for the Social Sciences. New York: Wiley.

McCullough, B.D. and C.G. Renfro (1998), Benchmarks and software standards: A case study of GARCH procedures. *Journal of Economic and Social Measurement* 25: 59 -71.

McCullough, B. and C. G. Renfro (2000), Some numerical aspects of nonlinear estimation. *Journal of Economic and Social Measurement 26*(1), 63–77.

McCullough, B. D. (1998), Assessing the reliability of statistical software: Part I. *The American Statistician 52*(4), 358–366.

— (1999a), Assessing the reliability of statistical software: Part II. *The American Statistician 53*(2), 149–159.

— (1999b), Econometric software reliability: Eviews, Limdep, Shazam and Tsp. *Journal of Applied Econometrics 14*(2), 191–202.

McCullough, B. D. and H. D. Vinod (1999), The numerical reliability of econometric software. *Journal of Economic Literature 37*, 633–665.

— (2003), Verifying the solution from a nonlinear solver: A case study. *American Economic Review 93*(3), 873–892.

McCullough, B. D. and B. Wilson (1999), On the accuracy of statistical procedures in Microsoft excel 97. *Computational Statistics & Data Analysis 31*(1), 27–37.

McKenzie, C. R. and S. Takaoka (2003), 2002: a LIMDEP odyssey. *Journal of Applied Econometrics 18*(2), 241–247.

Murray,W. (1972), Failure, the Causes and Cures. In Numerical Methods for Unconstrained Optimization, ed. W. Murray, pp. 107–122. New York: Academic Press.

Musa, J. D., A. Iannino, and K. Okumoto (1987), *Software reliability: measurement, prediction, application*. McGraw-Hill, Inc.

National Institute of Standards and Technology (2014), Statistical reference datasets. http://www.itl.nist.gov/div898/strd. (Accessed on April 15, 2014).

Odeh, O. O., A. M. Featherstone, and J. S. Bergtold (2010), Reliability of statistical software. *American Journal of Agricultural Economics 92*(5), 1472–1479.

Oster, R. A. (2002), An examination of statistical software packages for categorical data analysis using exact methods. *The American Statistician 56*(3), 235–246.

— (2003), An examination of statistical software packages for categorical data analysis using exact methods—part ii. *The American Statistician 57*(3), 201–213.

R Core Team (2013), *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

SAS Manual (2009), Sas/stat 13.2 user's guide. http://support.sas.com/documentation/cdl/en/statug/67523/PDF/default/statug.pdf. (Accessed on August 25, 2014).

Scrucca, L. and S. Weisberg. 2004. "A simulation study to investigate the behavior of the log-density ratio under normality." *Communications in Statistics: Simulation and Computation* 33: 159 – 178.

Simon, S. D. and J. P. LeSage (1988), Benchmarking numerical accuracy of statistical algorithms. *Computational Statistics & Data Analysis 7*(2), 197–209.

StataCorp (2015). Stata 14 base reference manual. http://www.stata.com/ manuals14/r.pdf. (Accessed on July 15, 2015).

Stokes, H. H. (2004), On the advantage of using two or more econometric software systems to solve the same problem. *Journal of Economic and Social Measurement 29*(1), 307–320.

Tomek, W. G. (1993), Confirmation and replication in empirical econometrics: A step toward improved scholarship. *American Journal of Agricultural Economics 75*(Special Issue), 6–14.

Train, K. E. (2003), *Discrete choice methods with simulation*. Cambridge University Press.

Wolfram (2015a), Unconstrained optimization: methods of local minimization. Wolfram language & system. Online documentation. Available at: http://reference.wolfram.com/language/tutorial/UnconstrainedOptimizationOverview.html. [Last accessed December 3, 2015].

— (2015b), Wolfram *Mathematica* tutorial collection. Available Online: https://www.wolfram.com/learningcenter/tutorialcollection/complete/. [Last Accessed December 3, 2015].

Table 1. Benchmark dataset specifications

Datasets	**P (Y_i =1) Cutoff Point**	**Predictor Functional Form**	**Collinearity (ρ)[a]**	***N***	**Variance[b]**
Base	P = 0.6	$\eta(\mathbf{X};b) = b_0 + b_1X$	-	200	$\sigma_1 = 1$
Multicollinearity1	P = 0.6	$\eta(\mathbf{X};b) = b_0 + b_1X_1 + b_2X_2$	$\rho_{12} = 0.75$	500	$\sigma_1 = 1.5$ $\sigma_2 = 1.5$
Multicollinearity2	P = 0.6	$\eta(\mathbf{X};b) = b_0 + b_1X_1 + b_2X_2$	$\rho_{12} = 0.95$	500	$\sigma_1 = 1.5$ $\sigma_2 = 1.5$
Multicollinearity3	P = 0.6	$\eta(\mathbf{X};b) = b_0 + b_1X_1 + b_2X_2$	$\rho_{12} = 0.995$	500	$\sigma_1 = 1.5$ $\sigma_2 = 1.5$
Multicollinearity4	P = 0.6	$\eta(\mathbf{X};b) = b_0 + b_1X_1 + b_2X_2 + b_3X_1^2 + b_4X_1X_2 + b_5X_2^2$	$\rho_{12} = 0.75$	1000	$\sigma_{10} = 1$ $\sigma_{11} = 1.5$ $\sigma_{20} = 1.5$ $\sigma_{21} = 2$
Multicollinearity5	P = 0.6	$\eta(\mathbf{X};b) = b_0 + b_1X_1 + b_2X_2 + b_3X_1^2 + b_4X_1X_2 + b_5X_2^2$	$\rho_{12} = 0.95$	1000	$\sigma_{10} = 1$ $\sigma_{11} = 1.5$ $\sigma_{20} = 1.5$ $\sigma_{21} = 2$
Multicollinearity6	P = 0.6	$\eta(\mathbf{X};b) = b_0 + b_1X_1 + b_2X_2 + b_3X_1^2 + b_4X_1X_2 + b_5X_2^2$	$\rho_{12} = 0.995$	1000	$\sigma_{10} = 1$ $\sigma_{11} = 1.5$ $\sigma_{20} = 1.5$ $\sigma_{21} = 2$
Multicollinearity7	P = 0.6	$\eta(\mathbf{X};b) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4$	$\rho_{ij} = 0.985$ $\forall$ i, j = 1,.. 4, i ≠ j	1000	$\sigma = 1$
Multicollinearity8	P = 0.4	$\eta(\mathbf{X};b) = b_0 + b_1X_1 + b_2X_2 + b_3\ X_1X_2$	$\rho_0 = 0.3$ $\rho_1 = 0.7$	50	-
Multicollinearity9	P = 0.4	$\eta(\mathbf{X};b) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_{12}X_1X_2 + b_{13}X_1X_3 + b_{23}X_2\ X_3 + b_{123}X_1X_2X_3$	$\rho_0 = 0.3$ $\rho_1 = 0.7$	400	$\sigma = 1$
Multicollinearity10	P = 0.4	$\eta(\mathbf{X};b) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_{11}X_1^2 + b_{12}X_1X_2 + b_{13}X_1X_3 + b_{23}X_2X_3 + b_{112}X_1^2X_2 + b_{113}X_1^2X_3 + b_{123}X_1X_2X_3 + b_{1123}X_1^2X_2X_3$	$\rho_0 = 0.3$ $\rho_1 = 0.7$	325	$\sigma = 1 – 3$
Multicollinearity11	P = 0.4	$\eta(\mathbf{X};b) = b_0 + b_1X_1 + b_2X_2 + b_3X_1X_2$	$\rho_0 = 0.8$ $\rho_1 = 0.4$	89	-

Table 1 Continued.

Datasets	**P (Y_i =1) Cutoff Point**	**Predictor Functional Form**	**Collinearity (ρ)[a]**	***N***	**Variance[b]**
Multivariate1	P =0.6	$\eta(X;b) = b_0 + b_1X_1 + b_2Ln(X2) + b_3X_3$	-	300	$\sigma = 1$
Multivariate2	P = 0.6	$\eta(\mathbf{X};b) = b_0 + b_1X_1 + b_2X_1^2 + b_3X_2 + b_4Ln(X_2) + b_5X_3$	-	300	$\sigma_1 = 0.9$ $\sigma_2 = 1.8$
Multivariate 3	P = 0.4	$\eta(\mathbf{X};b) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5$	$\rho_{ij} = 0.5$ $\forall$ i, j = 1,..,5, i ≠ j	1000	$\sigma_i = 15$
Multivariate4	P = 0.4	$\eta(\mathbf{X};b) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5$	ρ_{ij} = 0.3- 0.7 $\forall$ i, j = 1,…,5, i ≠ j	1000	$\sigma_i = 1$
Multivariate5	P = 0.5	$\eta(\mathbf{X};b) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + b_{11}X_1^2 + b_{12}X_1X_2 + b_{13}X_1X_3 + b_{14}X_1X_4 + b_{15}X_1X_5 + b_{22}X_2^2 + b_{23}X_2X_3 + b_{24}X_2X_4 + b_{25}X_2X_5 + b_{33}X_3^2 + b_{34}X_3X_4 + b_{35}X_3X_5 + b_{44}X_4^2 + b_{45}X_4X_5 + b_{55}X_5^2$	ρ_{ij} = 0.15 - 0.35 $\forall$ i, j = 1,..,5, i ≠ j	100	$\sigma = 1$
Multivariate6	P = 0.4	$\eta(\mathbf{X};b) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + b_{11}X_1^2 + b_{12}X_1X_2 + b_{13}X_1X_3 + b_{14}X_1X_4 + b_{15}X_1X_5 + b_{22}X_2^2 + b_{23}X_2X_3 + b_{24}X_2X_4 + b_{25}X_2X_5 + b_{33}X_3^2 + b_{34}X_3X_4 + b_{35}X_3X_5 + b_{44}X_4^2 + b_{45}X_4X_5 + b_{55}X_5^2$	ρ_{ij} =0.1- 0.6, i, j = 1-5, i ≠ j	200	σ = 0.5 -0. 77
Multivariate 7	P = 0.4	$\eta(\mathbf{X};b) = b_0 + b_1X_1 + b_2X_2$	$\rho = 0.75$	9	$\sigma_1 = 0.25$ $\sigma_2 = 0.40$
Cutoff1	P= 0.05	$\eta(\mathbf{X};b) = b_0 + b_1X_1 + b_2X_2$	$\rho_{12} = 0.99$	50	$\sigma = 1$
Cutoff2	P =0.15	$\eta(\mathbf{X};b) = b_0 + b_1X_1 + b_2Ln(X2) + b_3X_3$	-	32	$\sigma = 1$
Cutoff3	P = 0.0005	$\eta(\mathbf{X};b) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4$	$\rho_{12} = 0.96$ $\rho_{34} = 0.96$	5000	σ = 0.2- 2.5
Cutoff4	P = 0.00015	$\eta(\mathbf{X};b) = b_0 + b_1X_1 + b_2X_2 + b_3X_3$	ρ = 0.7 – (- 0.85)	20000	$\sigma = 1$

Table 1 Continued.

Datasets	**P (Y_i =1) Cutoff Point**	**Predictor Functional Form**	**Collinearity (ρ)[a]**	***N***	**Variance[b]**
Cutoff5	P = 0.05	$\eta(\mathbf{X};b) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + b_{11}X_1^2 + b_{12}X_1X_2 + b_{13}X_1X_3 + b_{14}X_1X_4 + b_{15}X_1X_5 + b_{22}X_2^2 + b_{23}X_2X_3 + b_{24}X_2X_4 + b_{25}X_2X_5 + b_{33}X_3^2 + b_{34}X_3X_4 + b_{35}X_3X_5 + b_{44}X_4^2 + b_{45}X_4X_5 + b_{55}X_5^2$	ρ_{ij} =0.3 – 0.5 $\forall$ i, j = 1,…,5, i ≠ j	500	σ = 1- 2
Cutoff6	P = 0.10	$\eta(\mathbf{X};b) = b_0 + b_1X_1 + b_2X_2 + b_{12}X_1X_2$	ρ_0 = 0.2, ρ_1 = 0.9	65	-
Cutoff7	P = 0.10	$\eta(\mathbf{X};b) = b_0 + b_1X_1 + b_2X_2 + b_{12}X_1X_2$	ρ_0 = 0.3, ρ_1 = 0.8	17500	-
Cutoff8	P= 0.19	$\eta(\mathbf{X};b) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_{11}X_1^2 + b_{12}X_1X_2 + b_{13}X_1X_3 + b_{23}X_2X_3 + b_{112}X_1^2X_2 + b_{113}X_1^2X_3 + b_{123}X_1X_2X_3 + b_{1123}X_1^2X_2X_3$	ρ_0 = 0.4, ρ_1 = 0.7	200	-
Empirical1[c]	-	$\eta(\mathbf{X};b) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + b_6X_6 + b_7X_7 + b_8X_8 + b_9X_9 + b_{10}X_{10} + b_{11}X_{11} + b_{12}X_{12} + b_{13}X_{13}$	-	1081	-
Quasisep1[d]	-	$\eta(\mathbf{X};b) = b_0 + b_1X$	-	100	-
Quasisep2[d]	-	$\eta(\mathbf{X};b) = b_0 + b_1X + b_2X^2 + b_3X^3 + b_4X^4 + b_5X^5$	-	60	-

[a] Collinearity in the datasets was determined by changing the correlation between the explanatory variables. Given that the inverse conditional distribution is dependent on the value of Y_i = 0,1, the correlation between covariates may change, giving the potential the specification of values for ρ_0 and ρ_1.

[b] Variability was introduced by specifying the value of the standard error of a given covariate.

[c] The empirical dataset is taken from survey data collected by the authors examining conservation practice adoption in Alabama.

[d] The estimation of the datasets examining quasi-separation are generated following the procedures in Ryan (1997).

Table 2. Statistical package assessed and associated details

Software Package	Version	Logistic Regression Command	Estimation Method and Algorithmic Options [a]	Starting Point Option (Yes/No)[b]	Default Tolerance Setting(s)[c]
LIMDEP NLOGIT 5	10	LOGIT	Maximum Likelihood NR (Default) Other options: BFGS, BHHH, and DFP	Yes	Gradient: 1e-6
MATLAB	8.6.0.267246 (R2015b)	GLMFIT	Maximum Likelihood NR (Default)	No	Parameters: 1e-6
R	3.3.1	GLM	IRLS FS	Yes	Objective function :1e-8
SAS	9.4	PROC LOGISTIC, PROC QLIM	Maximum Likelihood LOGISTIC: FS (Default) Other option: NR QLIM: QN (Default) Other options: CONGRA, NR with line search, TRUREG	Yes	Gradient: 1e-8
STATA/IC	14	LOGIT BINREG GLM	Maximum Likelihood NR (Default) Other options: BFGS, BHHH, and DFP	Yes	NR technique: nrtol (1e-5) and other techniques: qtol (1e-5)

[a] Estimation Methods: ML (Maximum Likelihood) and IRLS (Iteratively Reweighted Least Squares)
Algorithms: NR (Newton-Raphson), BHHH (Berndt-Hall-Hall-Hausman), BFGS (Broyden-Fletcher-Goldfarb-Shanno),and DFP (Davidon-Fletcher Powell), FS (Fisher's Scoring), , QN (Quasi Newton), CONGRA (Conjugate Gradient), TRUREG (Trust Region Optimization)

[b] This option indicates if the specified procedure allows the user to specify the starting point.

[c] The convergence criteria used are when the (i) gradient (gradient, relative gradient or scaled gradient) is less than tolerance; (ii) change in the parameter vector is less than tolerance; and (iii) the change in the deviance (for IRLS) or log-likelihood function is less than tolerance. For example, STATA default tolerance setting for the NR technique is nrtol (1e-5) and other technqiues use qtol (1e-5).

Table 3. Minimum LRE for parameter estimates using default settings

	STATA						MT	R		LIMDEP		SAS			
	LOGIT		BINREG		GLM		GF	GLM		LOGIT		LOGISTIC		QLIM	
Dataset	D1	D2	D1	D2	D1	D2	D	D1	D2	D1	D2	D1	D2	D1	D2
Base	10.9	6.7	10.9	6.6	10.9	6.6	10.8	10.9	10.8	10.9	6.6	5.4	6.6	8.2	6.4
Multico1	**0.7**	**0.7**	**0.7**	**0.7**	**0.7**	**0.7**	**0.7**	**0.7**	**0.7**	10	9.1	**0.7**	**0.7**	**0.7**	**0.7**
Multico2	9.7	9.3	9.7	9.3	9.7	9.4	10.5	9.7	10.7	9.7	10.7	4.7	5.2	6.8	10.1
Multico3	7.1	7.9	7.1	9.1	7.1	9.1	10.7	8.4	10.8	8.4	10.8	4.3	5.8	8.9	9.6
Multico4	7	7.1	7	7.1	7	7.1	10.2	10.2	10.2	10.2	10.2	5.5	5.3	8.4	8.4
Multico5	7.1	6.9	7.1	6.9	7.1	6.9	10.5	9.3	9.0	9.3	9.0	4.6	4.4	8.9	8.1
Multico6	7.0	6.6	7.0	6.6	7.0	6.6	10.2	6.9	6.9	6.9	6.9	6.9	6.9	8.1	8.9
Multico7	6.4	6.4	6.3	6	6.3	6.0	10.5	9.7	10.5	9.7	6.6	5.0	6.6	6.6	6.3
Multico8	10.1	9.6	10.1	9.5	10.1	9.5	10.1	10.1	8.2	6.2	8.2	6.2	3.9	7.2	6.3
Multico9	5.9	5.8	5.9	5.8	5.9	5.8	10.3	6.0	9.6	6.0	9.6	6.0	4.3	6.7	6.3
Multico10	4.8	4.8	5.6	5.6	4.8	4.8	10.4	9.1	10.4	9.1	10.4	**3.6**	4.6	**3.9**	4.8
Multico11	5.4	7.5	5.4	7.5	5.4	7.5	10.2	10.2	7.5	5.4	7.5	5.4	**2.9**	4.5	4.7
Multivar1	8.2	8.1	8.2	8.1	8.2	8.1	10.6	9.3	10.6	9.1	7.2	4.4	7.2	7.4	8.0
Multivar2	7.5	7.5	7.5	7.5	7.5	7.5	10.6	9.5	8.6	**1.7**	**1.7**	4.5	4.0	8.0	7.1
Multivar3	10.3	7.7	10.3	7.7	10.3	7.7	10.3	10.3	7.7	6	7.7	6.0	**3.4**	7.5	8.4
Multivar4	**2.4**	**2.4**	**2.4**	**2.4**	**2.4**	**2.4**	**2.4**	**2.4**	**2.4**	**2.4**	**2.4**	**2.4**	**2.4**	7.8	**2.4**
Multivar5	5.1	5.1	5.1	5.1	5.1	5.1	10.3	8.0	**0.0**	8.0	0.0	**3.4**	**3.7**	5.8	5.2
Multivar6	6.7	6	6.6	6.1	6.6	6.1	10.3	7.7	7.5	7.7	7.5	**3.7**	**3.6**	4.5	4.1
Multivar7	7.1	7.1	7.1	7.1	7.1	7.1	10.6	10.6	11	7.2	11	7.2	5.5	5.4	5.7
Cutoff1	8.5	8.1	**NS**	**NS**	8.5	8.1	10.2	5.8	5.7	5.8	10.2	**2.0**	**2.0**	**3.3**	**3.7**
Cutoff2	5.7	5.7	5.7	5.7	5.7	5.7	10.8	10.8	10.8	5.9	5.9	5.9	5.9	4.5	5.3
Cutoff3	5.9	5.9	**NS**	**NS**	6.9	6.9	10.6	7.3	7.3	7.3	7.3	7.3	7.3	7.4	6.1
Cutoff4	5.5	5.5	**NS**	**NS**	5.7	5.7	10.4	10.4	10.4	6.8	6.8	6.8	6.8	7.6	6.6
Cutoff5	4.5	4.5	4.6	4.6	4.6	4.6	10.3	6.9	6.9	6.9	6.9	**2.8**	**2.8**	**3**	**2.9**
Cutoff6	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
Cutoff7	8.4	9	NS	NS	8.4	9.0	10.8	6.7	6.6	6.7	6.6	**3.1**	**3.1**	6.4	5.4
Cutoff8	5.9	5.7	6.1	5.8	6.1	5.8	10.4	7.7	7.6	7.7	7.6	**3.7**	**3.7**	5.1	6.0
Empirical1	6.0	10.7	6.0	10.7	6.0	10.7	10.7	10.7	10.7	6.4	6.6	6.4	6.3	5.5	5.4
Quasisep1	6.5	6.6	5.2	4.0	5.2	5.3	10.3	8.0	10.3	8.0	7.0	4.1	7.0	8.3	8.2
Quasisep2	**3.6**	**3.6**	**3.6**	**3.7**	**3.6**	**3.6**	**3.6**	**3.6**	**3.6**	10.2	9.2	**3.6**	**3.6**	**3.6**	**3.6**
Number of Datasets Meeting Minimum LRE Criterion															
LRE ≥ 4.0	26	26	23	23	26	26	26	26	25	27	27	19	18	24	24
LRE ≥ 6.0	15	18	17	16	18	18	26	25	24	24	26	8	8	17	16

Note: D = Default Settings; 1 = Starting Point 1; 2 = Starting point 2, NS= did not converge to a solution, MT: MATLAB, GF: GLMFIT. Bolded numbers indicate minimum parameter estimates with an LRE below a reliability threshold value of 4.

Table 4. Minimum LRE for parameter estimates with user optional settings

	STATA						MT	R		LIMDEP		SAS			
	LOGIT		BINREG		GLM		GF	GLM		LOGIT		LOGISTIC		QLIM	
Dataset	U1	U2	U1	U2	U1	U2	D	D1	D2	U1	U2	U1	U2	U1	U2
Base	10.9	8.4	10.9	8.4	10.9	8.4	10.8	10.9	10.8	10.9	10.6	10.9	10.8	10.6	6.6
Multico1	**0.7**	**0.7**	**0.7**	**0.7**	**0.7**	**0.7**	**0.7**	**0.7**	**0.7**	10	9.1	**0.7**	**0.7**	**0.7**	**0.7**
Multico2	9.7	9.3	9.7	9.3	9.7	9.4	10.5	9.7	10.7	9.7	10.7	9.7	10.7	10.2	10.1
Multico3	8.7	8.2	7.9	9.1	7.9	9.1	10.7	8.4	10.8	8.9	10.8	8.4	10.8	8.9	9.6
Multico4	7.0	7.1	7.0	7.1	7.0	7.1	10.2	10.2	10.2	10.2	10.2	10.2	10.2	9.7	9.5
Multico5	7.8	7.1	7.2	7.2	7.2	7.2	10.5	9.3	9.0	9.3	9.0	9.3	9.0	10.4	9.4
Multico6	7.0	7.0	7.0	7.2	7.0	7.0	10.2	6.9	6.9	10.2	10	10.2	10.2	8.1	8.9
Multico7	6.5	6.4	6.7	6.2	6.7	6.2	10.5	9.7	10.5	9.7	7.9	9.7	10.5	7.9	6.9
Multico8	10.1	9.6	10.1	9.5	10.1	9.5	10.1	10.1	8.2	10.1	9.1	10.1	8.2	8.6	8.2
Multico9	5.9	5.8	6.4	5.8	6.4	5.8	10.3	6.0	9.6	9.0	9.6	6.0	9.6	6.7	9.3
Multico10	5.4	5.0	5.6	5.6	5.4	5.0	10.4	9.1	10.4	9.1	10.4	9.1	10.4	8.0	9.0
Multico11	10	7.5	10	7.5	10	7.5	10.2	10.2	7.5	10.2	7.9	10.2	7.5	6.7	8.9
Multivar1	8.3	8.1	8.3	8.1	8.3	8.1	10.6	9.3	10.6	10.8	7.4	9.1	7.2	9.2	9.8
Multivar2	7.5	7.5	7.5	7.5	7.5	7.5	10.6	9.5	8.6	**1.7**	**2.5**	9.5	8.6	8.7	8.1
Multivar3	10.3	7.7	10.3	7.7	10.3	7.7	10.3	10.3	7.7	10.4	9.9	10.3	7.7	10.2	8.4
Multivar4	**2.6**	**2.6**	**2.5**	**2.5**	**2.5**	**2.5**	**2.4**	**2.4**	**2.4**	**2.4**	**2.4**	**2.4**	**2.4**	7.8	**2.4**
Multivar5	5.5	5.3	5.5	5.3	5.5	5.3	10.3	8.0	**0.0**	8.0	7.5	8.0	8.1	9.1	9.0
Multivar6	6.7	6.3	6.6	6.3	6.6	6.3	10.3	7.7	7.5	7.7	8.3	7.7	7.5	8.6	7.5
Multivar7	7.1	7.2	7.6	7.6	7.9	7.6	10.6	10.6	11	7.7	11	10.6	11	6.8	7.3
Cutoff1	8.5	8.1	8.5	8.1	8.5	8.1	10.2	5.8	5.7	10.2	10.2	5.8	5.7	6.8	6.8
Cutoff2	6.5	6.4	6.5	6.4	6.5	6.4	10.8	10.8	10.8	10.8	10.8	10.8	10.8	8.1	7.6
Cutoff3	6.3	6.3	6.9	6.9	6.9	6.9	10.6	7.3	7.3	8.0	7.7	7.3	7.3	7.5	7.5
Cutoff4	5.5	5.5	5.7	5.7	5.7	5.7	10.4	10.4	10.4	8.0	8.0	10.4	10.4	9.0	9.0
Cutoff5	4.5	4.6	4.6	4.7	4.6	4.7	10.3	6.9	6.9	6.9	10.6	6.9	6.9	8.4	7.3
Cutoff6	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
Cutoff7	8.4	9.0	8.4	9	8.4	9.0	10.8	6.7	6.6	9.8	10.2	6.7	6.7	7.7	7.9
Cutoff8	6.2	6.4	6.5	6.3	6.5	6.3	10.4	7.7	7.6	9.7	7.7	7.7	7.6	7.7	7.6
Empirical1	10.7	10.7	10.7	10.7	10.7	10.7	10.7	10.7	10.7	7.4	7.2	10.7	10.7	10.6	10.5
Quasisep1	6.5	6.6	5.4	6.6	6.6	6.6	10.3	8.0	10.3	9.1	9.1	8.0	10.3	8.3	8.2
Quasisep2	**3.6**	**3.6**	**3.6**	**3.7**	**3.6**	**3.6**	**3.6**	**3.6**	**3.6**	10.2	9.2	**3.6**	**3.6**	**3.6**	**3.6**
Number of Datasets Meeting Minimum LRE Criterion															
LRE ≥ 4.0	26	26	26	26	26	26	26	26	25	27	27	26	26	27	26
LRE ≥ 6.0	21	21	21	21	21	21	26	25	24	27	27	25	25	27	26

Note: D = Default Settings; U = User Optimized Setting; 1 = Starting Point 1; 2 = Starting point 2, NS= did not converge to a solution, MT: MATLAB, GF: GLMFIT, MATLAB and R have no user optional settings. Bolded numbers indicate minimum parameter estimates with an LRE below a reliability threshold value of 4.

Table 5. Minimum LRE for standard errors using default settings

	STATA						MT	R		LIMDEP		SAS			
	LOGIT		BINREG		GLM		GF	GLM		LOGIT		LOGISTIC		QLIM	
Dataset	D1	D2	D1	D2	D1	D2	D	D1	D2	D1	D2	D1	D2	D1	D2
Base	11.3	7.3	11.3	7.3	11.3	7.3	11.3	6.3	7.3	11.3	7.3	6.3	7.3	5.6	5.6
Multico1	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	10.5	9.7	**0.1**	**0.1**	**0.1**	**0.1**
Multico2	9.9	9.5	9.9	9.5	9.9	9.5	10.7	5	5.4	9.9	11	5.0	5.4	4.7	4.7
Multico3	6.7	7.2	6.7	7.2	6.7	7.2	6.0	4.1	5.5	8.2	10.7	4.1	5.5	3.1	3.1
Multico4	8.4	7.9	8.4	7.9	8.4	7.9	10.1	6.0	5.8	10.1	10.1	5.9	5.8	4.8	4.8
Multico5	7.9	7.3	7.9	7.3	7.9	7.3	8.7	4.9	4.7	9.7	9.3	4.9	4.7	4.6	4.7
Multico6	7.1	6.8	7.1	6.8	7.1	6.8	6.8	**3.7**	**3.7**	7.3	7.2	7.3	7.2	**2.4**	**2.4**
Multico7	5.1	5.1	5.1	5.1	5.1	5.1	6.9	4.7	6.4	9.4	6.4	4.7	6.4	**3.4**	**3.4**
Multico8	10.5	10.1	10.5	10.1	10.5	10.1	7.1	6.6	4.4	6.6	8.8	6.6	4.4	5.7	5.7
Multico9	5.9	7.2	7.3	7.2	5.9	7.2	10.4	4.3	5.4	8.2	10.5	8.2	8.2	5.1	5.1
Multico10	7.1	7.1	7.1	7.1	7.1	7.1	10.2	5.6	6.4	10.2	10.2	5.6	6.4	4.1	4.1
Multico11	7.3	9.0	7.3	9.0	7.3	9.0	10.2	7.3	4.5	7.3	9.0	7.3	4.5	5.7	5.7
Multivar1	8.4	8.4	9.5	8.4	8.4	8.4	9.7	4.9	7.8	9.6	7.7	4.9	7.7	5.4	5.4
Multivar2	7.9	7.9	7.9	7.9	7.9	7.9	9.2	5.2	4.7	**2.5**	**2.5**	5.2	4.7	4.5	4.5
Multivar3	10.7	8.7	10.7	8.7	10.7	8.7	10.7	6.9	4.4	6.9	8.7	6.9	4.4	5.7	5.7
Multivar4	**3.7**	**3.7**	**3.7**	**3.7**	**3.7**	**3.7**	9.3	4.2	4.0	8.4	8.0	4.2	4.0	5.7	**3.9**
Multivar5	7.3	7.3	7.3	7.3	7.3	7.3	9.9	5.0	**0.0**	9.7	**0.0**	5.0	5.4	5.1	5.2
Multivar6	6.7	6.4	6.7	6.5	6.7	6.4	8.9	4.1	4.1	8.1	8.0	4.1	4.1	4.6	4.6
Multivar7	6.2	6.2	6.2	6.2	6.2	6.2	10.3	7.2	5.4	7.2	10.5	7.2	5.4	**3.7**	**3.7**
Cutoff1	10.6	10.4	**NS**	**NS**	10.6	10.4	10.0	**3.9**	**3.9**	7.7	10.3	**3.9**	**3.9**	4.3	4.3
Cutoff2	6.5	6.6	6.5	6.6	6.5	6.6	8.4	6.7	6.7	6.7	6.7	6.7	6.7	5.0	4.9
Cutoff3	5.8	5.8	**NS**	**NS**	5.8	5.8	9.5	**3.8**	**3.8**	7.5	7.5	7.5	7.5	4.1	4.1
Cutoff4	5.2	5.2	**NS**	**NS**	5.2	5.2	8.6	6.7	6.7	6.7	6.7	6.7	6.7	4.6	4.6
Cutoff5	5.7	5.7	5.7	5.7	5.7	5.7	10.2	4.3	4.3	8.5	8.5	4.3	4.3	4.0	4.0
Cutoff6	**0.9**	**0.9**	**0.9**	**0.9**	**0.9**	**0.9**	**0.9**	**0.9**	**0.9**	**1.1**	**1.1**	**0.9**	**0.9**	**0.9**	**0.9**
Cutoff7	8.8	10.1	**NS**	**NS**	8.8	10.1	9.2	**3.6**	**3.5**	7.1	7.1	**3.6**	**3.5**	4.8	4.8
Cutoff8	6.3	6.1	6.3	6.1	6.3	6.1	10.0	4.0	4.0	8.1	8.0	4.0	4.0	4.9	4.9
Empir1	6.5	10.3	6.5	10.3	6.5	10.3	9.1	6.9	6.8	6.9	6.8	6.9	6.8	5.2	5.2
Quasisep1	4.0	4.0	4.0	4.0	4.0	4.0	6.8	**3.9**	6.8	7.7	6.8	**3.9**	6.8	**2.7**	**2.7**
Quasisep2	4.1	4.1	4.1	4.1	4.1	4.1	4.1	4.1	4.1	9.5	9.7	4.1	4.1	**0.0**	**0.0**

Number of Datasets Meeting Minimum LRE Criterion

LRE ≥ 4.0	27	27	23	23	27	27	28	23	23	28	27	25	26	22	21
LRE ≥ 6.0	20	21	19	19	20	21	27	9	8	28	27	11	11	0	0

Note: D = Default Settings; U = User Optimized Setting; 1 = Starting Point 1; 2 = Starting point 2, NS= did not converge to a solution, MT: MATLAB, GF: GLMFIT. Bolded numbers indicate minimum parameter estimates with an LRE below a reliability threshold value of 4.

Table 6. Minimum LRE for standard errors with user optional settings

	STATA						MT	R		LIMDEP		SAS			
	LOGIT		BINREG		GLM		GF	GLM		LOGIT		LOGISTIC		QLIM	
Dataset	U1	U2	U1	U2	U1	U2	D	D1	D2	U1	U2	U1	U2	U1	U2
Base	11.3	9.7	11.3	9.7	11.3	9.9	11.3	6.3	7.3	11.3	11.1	11.3	11.3	5.6	5.6
Multico1	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	10.5	9.7	**0.1**	**0.1**	**0.1**	**0.1**
Multico2	9.9	9.4	9.9	9.5	9.9	9.5	10.7	5.0	5.4	9.9	11	9.9	11	4.7	4.7
Multico3	7.2	7.3	7.3	7.2	7.3	7.2	6.0	4.1	5.5	8.8	10.7	8.2	10.8	3.1	3.1
Multico4	8.4	**1.4**	8.4	**1.4**	8.4	**1.4**	10.1	6.0	5.8	10.1	10.1	10.1	10.1	4.8	4.9
Multico5	7.9	8.1	7.9	7.9	7.9	7.9	8.7	4.9	4.7	9.7	9.3	8.8	8.8	4.6	4.6
Multico6	7.1	7.1	7.1	**1.6**	7.1	7.1	6.8	**3.7**	**3.7**	9.8	9.9	9.9	10.6	**2.4**	**2.5**
Multico7	5.1	5.6	**0.2**	**0.2**	**0.2**	**0.2**	6.9	4.7	6.4	9.4	8.4	9.4	10.5	**3.4**	**3.4**
Multico8	10.5	10.1	10.5	10.1	10.5	10.1	7.1	6.6	4.4	10.5	10.1	10.5	8.8	6.1	5.7
Multico9	7.3	7.2	7.3	7.3	7.3	7.3	10.4	4.3	5.4	10.1	10.5	8.2	8.2	5.1	5.1
Multico10	6.8	**0.5**	7.1	7.1	6.8	0.5	10.2	5.6	6.4	10.2	10.2	10.3	10.2	4.1	4.1
Multico11	10.2	9.0	10.2	9.0	10.2	9.0	10.2	7.3	4.5	10.2	9.7	10.2	9.0	5.8	5.8
Multivar1	8.7	8.4	8.7	8.4	8.7	8.4	9.7	4.9	7.8	10.6	8.2	9.6	7.7	5.4	5.4
Multivar2	7.9	7.9	7.9	7.9	7.9	7.9	9.2	5.2	4.7	**2.5**	**2.5**	10.4	9.3	4.5	4.5
Multivar3	10.7	8.7	10.7	8.7	10.7	8.7	10.7	6.9	4.4	10.8	10.7	10.7	8.7	5.7	5.7
Multivar4	**0.0**	**0.0**	**0.0***	**0.0**	**0.0**	**0.0**	9.3	4.2	4.0	6.3	6.8	4.2	4.0	5.7	**3.8**
Multivar5	7.2	6.8	7.2	6.8	7.2	6.8	9.9	5.0	**0.0**	9.7	9.3	9.8	9.8	5.1	5.1
Multivar6	6.7	6.6	6.7	6.7	6.7	6.7	8.9	4.1	4.1	8.1	8.5	8.1	8.0	4.9	4.9
Multivar7	6.2	6.2	6.1	**0.0**	6.1	**0.0**	10.3	7.2	5.4	7.6	10.5	10.3	10.5	**3.7**	**3.7**
Cutoff1	10.6	10.4	10.6	10.4	10.6	10.4	10	**3.9**	**3.9**	10.3	10.3	7.7	7.7	4.3	4.4
Cutoff2	7.4	7.1	7.4	7.1	7.4	7.1	8.4	6.7	6.7	10.1	10.1	10.1	10.1	4.9	4.9
Cutoff3	6.5	5.7	5.8	5.8	5.8	5.8	9.5	**3.8**	**3.8**	9.3	8.0	7.5	7.5	4.1	4.1
Cutoff4	5.2	5.1	5.2	5.2	5.2	5.2	8.6	6.7	6.7	8.2	8.2	10.2	10.2	4.6	4.6
Cutoff5	6.4	6.4	5.7	5.7	5.7	5.7	10.2	4.3	4.3	8.5	10.2	8.5	8.5	4.0	4.0
Cutoff6	**0.9**	**0.9**	**0.9**	**0.9**	**0.9**	**0.9**	**0.9**	**0.9**	**0.9**	**1.1**	**1.1**	**0.9**	**0.9**	**0.9**	**0.9**
Cutoff7	8.8	10.1	8.8	10.2	8.8	10.1	9.2	**3.6**	**3.5**	10.5	10.0	7.1	7.1	4.8	4.9
Cutoff8	5.8	5.8	7.0	6.7	7.0	6.7	10	4.0	4.0	9.9	8.1	8.1	8.0	4.9	5.0
Empir1	10.3	10.3	10.3	10.3	10.3	10.3	9.1	6.9	6.8	9.8	8.4	10.3	10.3	5.2	5.2
Quasisep1	5.7	5.7	4.0	**3.9**	**0.0**	**3.9**	6.8	**3.9**	6.8	9.4	9.4	7.7	10.4	**2.7**	**2.7**
Quasisep2	4.1	4.1	4.1	4.1	4.1	4.1	4.1	4.1	4.1	9.5	9.7	4.1	4.1	**0.0**	**0.0**
Number of Datasets Meeting Minimum LRE Criterion															
LRE ≥ 4.0	27	25	26	22	25	22	28	23	23	28	28	28	28	23	22

LRE ≥ 6.0	22	19	21	18	21	18	27	9	8	28	28	26	26	1	0

Note: D = Default Settings; U = User Optimized Setting; 1 = Starting Point 1; 2 = Starting point 2, NS= did not converge to a solution, MT: MATLAB, GF: GLMFIT, MATLAB has no user optional settings. Bolded numbers indicate minimum parameter estimates with an LRE below a reliability threshold value of 4.