AgEcon SEARCH
RESEARCH IN AGRICULTURAL & APPLIED ECONOMICS

The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search
http://ageconsearch.umn.edu
aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

On the Examination of the Reliability of Statistical Software for Estimating Logistic Regression Models

Jason S. Bergtold[1], Krishna Pokharel[2] and Allen Featherstone[3]

[1]Associate Professor, Department of Agricultural Economics, Kansas State University, Manhattan, KS 66506,bergtold@ksu.edu

[2] Graduate Research Assistant, Department of Agricultural Economics, Kansas State University, Manhattan, KS 66506, kpoharel@ksu.edu

[3] Professor, Department of Agricultural Economics, Kansas State University, Manhattan, KS 66506,afeather@ksu.edu

Selected Paper prepared for presentation at the 2015 Agricultural & Applied Economics Association and Western Agricultural Economics Association Annual Meeting, San Francisco, CA, July 26-28

Copyright 2015 by Bergtold, Pokharel and Featherstone. All rights reserved. Readers may make verbatim copies of this document for non-commercial purposes by any means, provided that this copyright notice appears on all such copies.

On the Examination of the Reliability of Statistical Software for Estimating Logistic Regression Models

Abstract

The numerical reliability of software packages was examined for the logistic regression model. Software tested include SAS 9.3, MATLAB R2012a, R 3.1.0, Stata/IC 13.1 and LIMDEP 10.5. Thirty benchmark datasets were created by simulating different conditional binary choice processes. To obtain certified values, this study followed the National Institute of Standards and Technology procedures when they generated certified values of parameter estimates and standard errors for the nonlinear logistic regression models used. The logarithm of the relative error was used as a measure of accuracy to examine the numerical reliability of these packages.

1 Introduction

The primary objective of statistical software is to analyze data or estimate models. Researchers use estimated results for different purposes (e.g. policy analysis, prediction, inference etc.) assuming the results are reliable, which also mean the results are numerically accurate.[1] Thus, the numerical accuracy of the results regardless of the use of the statistical software packages is one of the crucial factors for credible research. However, research has shown that different statistical packages may provide different results for the same problem (McCullough and Vinod 1999). Consider a scenario where two researchers are solving the same problem, but using different statistical packages, report 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 specify the statistical procedures (McCullough 1998; Odeh, Featherstone, and Bergtold 2010).

However, in most cases, researchers consider the results may be 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, Featherstone, and Bergtold 2010). Researchers mainly focus on user-friendliness and speed of software packages, often ignoring the numerical accuracy of software (McCullough 2000b). If the estimated result is not reliable, then it has strong negative implications for policy analysis, statistical inference, prediction etc., which weakens the work of applied economists (Tomek 1993). Estimates are more sensitive to starting values and algorithms in nonlinear models and a small change in model specifications may significantly

[1] Reliability describes the repeatability and consistency of a measure or test. Musa, Iannino, and Okumoto (1987) define software reliability as "the probability of failure-free operation of a computer program in a specified environment for a specified period of time".

change results (McCullough and Vinod 2003; Odeh, Featherstone, and Bergtold 2010).[2] The two main problems for many nonlinear models are: one algorithm may give solutions while other may fail and even if both algorithms give solutions, one may be more accurate. The reasons for different solutions are differences in step length, convergence criterion, and the way that derivatives are computed (McCullough and Renfro 2000).

Logistic regression is one of the widely used models using maximum likelihood in modeling a discrete choice outcome variable by agricultural economists and economists. The estimation of logistic regression is based on the assumption of asymptotic maximum likelihood inference. Greene (2002) showed that if nonlinearity in the parameters appear on the right-hand side of the equation, then the nonlinear least squares estimator is consistent. However, if parameters appear nonlinearly in functions of the dependent variable in a model, then the nonlinear least squares estimator is no longer consistent because it ignores the Jacobian of the transformation (a derivative of a coordinate transformation), but the maximum likelihood (ML) estimator is consistent and efficient for these type of models. Moreover, if parameters contain in the Jacobian matrix (the matrix of all first order partial derivative of a vector valued function), least squares is different than maximum likelihood. This indicates that separate benchmark tests may be required for maximum likelihood estimation routines for logistic regression.

[2] In any empirical work, researchers do not know true values, thus cross validation of research results become critical for nonlinear models.

The purpose of this research is to examine the numerical reliability of five statistical software packages for the estimation of logistic regression models. Software testing was conducted using SAS 9.3, MATLAB R2012a, R 3.1.0, Stata/IC 13.1 and LIMDEP 10.5, which are commonly used by both agricultural economists and economists. This paper focuses on the numerical accuracy of the results (parameter estimates and standard errors) estimated by these packages using the simulated datasets with several available algorithms, starting values, convergence criteria, and different commands and whether errors are correctly identified and reported during the estimation of logistic regression.

The main contribution of this study is to conduct a comparative examination of the numerical reliability of the above mentioned software packages with the simulated datasets, which have not been examined in this respect. The results of this study will be useful to researchers and software vendors as it shows the strengths and weaknesses of software packages for the logistic regression estimation. The software vendors may address inadequacy in later versions if it exists and researchers can choose a software package based on their problems because some packages perform well for particular datasets (e.g. multicollinarity, quasi separation etc.). Collinearity is a big challenge in environmental or ecological research, whereas quasi-complete separation or complete separation is a common problem in the area of health economics. For example, Heinze and Schemper (2002) found complete separation in the statistical analysis of endometrial cancer study. If there is a problem of complete separation in data, the maximum likelihood estimates for logistic regression do not exist.

Many past studies have examined the reliability of software packages including SAS, MATLAB, STATA, LIMDEP, etc. (Musa, Iannino, and Okumoto 1987; McCullough 1998, 1999b; Kolenikov 2001; Keeling and Pavur 2007; Odeh, Featherstone, and Bergtold 2010; M´elard 2014). These studies have mainly focused on linear and nonlinear regression models using the National Institute of Science and Technology (NIST) benchmark datasets. Because of the reliability studies, many software vendors correct inadequacies in a new version. For example, the results obtained from LIMDEP 8.0 were better than LIMDEP 7.0 for nonlinear regression (Odeh, Featherstone, and Bergtold 2010).

Although the logistic regression model is popular for choice decision modeling, there has not been any systematic examination of the numerical reliability of software packages for logistic regression estimation routines. There are some studies that focus on choice estimation using built-in commands from software packages. For example, Huber and Train (2001) examined the similarities and differences between classical and Bayesian methods for mixed logit. Likewise, Oster (2002, 2003) used exact methods to compare StatXact, LogXact, Stata, Testimate, and SAS based on hardware requirements, documentations, data entry, estimated results etc. However, these studies were not intended to examine the numerical reliability of software.[3]

2 Literature Review

This section provides a general background on sources and types of error due to data storage and computation by software. The first subsection introduces representation form of data on importing and exporting associating with errors. The remaining sections discuss past studies related to reliability of statistical software focusing on estimator, algorithm, and starting values.

2.1 Computational Error

Computational error may occur during data storage and the estimation of a model and the sources of error can be viewed from different perspectives. One way to view at the possible sources of error is from computing side, which includes the interaction among hardware, complier, and algorithm. If problem exists in either of the three components or all, then the ability to produce numerically accurate results decreases. An alternative approach is looking at statistical computing as a data flow process.

[3] An important point to be noted here is that reliable or numerically accurate estimates do not imply the estimates are consistent and unbiased, which are not the objective of the study.

Software package needs to import and export data for any statistical estimation (Sawitzki 1994). It imports and exports data in ASCII representation (a code for representing English characters as numbers, with each letter assigned a number from 0 to 127) with finite precision based on powers of ten, but the internal representation is binary.[4] The different representations rarely coincide in precision, which yields truncation errors (Sawitzki 1994).[5]

Simon and LeSage (1988) indicated that truncation, cancellation, and accumulation errors are the main errors for numerical inaccuracies. Truncation error occurs when computer stores data incorrectly, especially decimal numbers. The presence of truncation error leads to the less accurate result in subsequent calculations. Cancellation error occurs if the data have a low level of relative variations, which means the subtraction of two roughly equal numbers. Similarly, accumulation error occurs gradually from small errors to large errors in total number of computations during estimation process.

McCullough and Vinod (1999) reported the impact of truncation or round off error for Vancouver Stock Exchange index, which began with value of 1,000. The index was recalculated to four decimal places, but reported only with three decimal places truncating the last digit. The index value became 520 after some months in

[4] Computer stores numbers with finite precision. It cannot perform any computation with any arbitrary precision. However, the objective is to understand how the system work within the limitations rather than finding the limitations in a system.

[5] Altman, Gill, and McDonald (2004) define precision as the degree of agreement among a set of measurements of the same quantity- the number of digits that are the same under unchanged conditions for repeated measurements.

similar economic situations. The index value found to be 1098.892 when the index was recalculated properly. This case indicates that if the errors occur during intermediate computations, final estimation constitutes errors in a proportion. But to determine what proportion is correct for intermediate computation is difficult, and the potential error may still be significant (Odeh, Featherstone, and Bergtold 2010).

2.2 Studies on Linear and Nonlinear Regressions

Several past studies have examined the numerical reliability of statistical software (Koro¨si, Matyas, and Sz´ekely 1993; Sawitzki 1994; McCullough 1998; Kolenikov 2001; Odeh, Featherstone, and Bergtold 2010). Most of these studies used Statistical Reference Datasets (StRD) developed by NIST to examine the reliability of software for linear and nonlinear regressions, and so on. For example, McCullough (2004) used the Wilkinson's tests (entry level tests) to examine the numerical accuracy of E-Views 3, LIMDEP 7, RATS 4.3 (Regression Analysis of Time Series), SHAZAM 8, and TSP 4.4 (Time Series Processor). These tests are simple, thus they can be used to most of the statistical packages for examining the reliability and it is assumed that every reliable software should pass this test. However, some packages still have some flaws like dropping points from a graph, incorrect calculation of the sample variance, correlation coefficients in excess of unity, and incorrect and inconsistent handling of missing values. The author also argued that some of these statistical packages fixed flaws like incorrect calculation of standard error that were raised by the reliability test literature.

Veall (1991) performed a comprehensive study of SHAZAM 6.2 and found that the SHAZAM developer made significant improvement in version 6.2 than the version 6. The author also argued that SHAZAM's performance is well on the Longley benchmark regression.[6] However, the author also claimed that SHAZAM has problem on computation of correlations of very large and very small perfectly correlated numbers.

Silk (1996) complements the work of Veall (1991) by examining iterated SUR (Seemingly Unrelated Regression) and system of linear equations in SAS 6.10, SHAZAM 7.0, and TSP 4.3. The iterated SUR routines can give different parameter estimates and different test statistics for hypothesis of interest. The author found that each package provides identical results for two-stage least squares, three-stage least squares, and seemingly unrelated regression. However, these packages provide different result from the built-in routines for the limited information maximum likelihood and the full information maximum likelihood procedures. Reliability studies help to find bug and inadequacy in software packages, which motivate researcher to examine packages in a new version with different tests and benchmark datasets. For example, McCullough (1998, 1999a) proposed and conducted the intermediate level tests on linear and nonlinear regressions, random number generators, and statistical distributions in SAS

[6] Longley is a benchmark dataset provided by NIST.

6.12, SPSS 7.5, and S-PLUS 4.[7] All software packages produce reliable results on univariate summary statistics and linear regression test suites. However, the author found weaknesses in all three packages in random number generators due to deficient period length. SAS and SPSS had inadequate routines in the one-way analysis of variance (ANOVA) and nonlinear least squares. If inadequacy in any routine exists for this kind of average tests, it means the software packages do not meet the minimum requirements that is supposed to have for reliable estimation. The deficiencies in software packages imply that the reliability of statistical packages cannot be taken for granted (McCullough 1999a).[8]

The reliability of the statistical procedures in Microsoft Excel 97 for linear and nonlinear regressions, random number generation, and statistical distributions were assessed and the results exhibited that Excel's performance was not adequate in all three areas (McCullough and Wilson 1999). Similarly, Nerlove (2005) examined the Excel 2000 with the numerical algorithms group's (NAG's) add-in and found that the performance of the Excel was not significantly different with and without the add-in.[9] The author also suggested the correction method if the program does not provide the accurate results for high difficulty level of data for linear regression model. For example, the data set with high multicollinearity, it can be corrected with centering

[7] NIST ordered datasets by level of difficulty (lower, average, and higher) and the level of difficulty of a dataset depends on the algorithm.

[8] In subsequent versions of statistical software, these problems may not exist because software developers, generally, remedy deficiencies that were noted in earlier version.

[9] Add-ins in Microsoft Excel provide optional commands and features for data analysis, which may not be available by default. Thus, users must install or (in some cases) activate these add-ins to use them.

the data. It means subtracting the mean for each explanatory variable may help decrease the degree of multicollinarity.

Likewise, Keeling and Pavur (2007) conducted a comparative study of the reliability of several software packages (Excel 2000/*XP*, Excel 2003, SAS 9.1, JMP 5.0, Minitab 14, R 1.9.1, Splus 6.2, SPSS 12, Stata 8.1, and StatCrunch 3.0) with regard to univariate summary statistics, one-way ANOVA, linear regression, and nonlinear regression using the NIST data sets. These packages performed better than the earlier versions with regard to linear and nonlinear regressions. However, SAS 9.1 had difficulty on the univariate data in calculating the autocorrelation for the three data points. SAS 9.1 improved its performance for all of the average and high difficulty datasets whereas McCullough (1999a) reported LREs of zero for the above datasets. Similarly, Odeh, Featherstone, and Bergtold (2010) showed that some packages performed better with a new version. For example, Excel 2007 gave higher LRE value than was found in the previous literature. However, this does not imply that there are no issues with Excel. Likewise, STATA 10 gave slightly better results from those reported by STATA (2007) on its website. These studies show that software vendors improve the performance of a packages in a new version, which also justify the importance of reliability studies.

2.3 Maximum Likelihood Estimators

Nonlinear estimation routines also provide solutions for maximum likelihood and most of software packages provide t-statistics and confidence intervals (Wald intervals) for maximum likelihood. These statistics are based on quadratic approximations to the log-likelihood and they will be accurate only if the log-likelihood is approximately quadratic over the range of interest. Moreover, Wald intervals and likelihood intervals are quite similar if log-likelihood is approximately quadratic. However, when log-likelihood is not approximately quadratic- which is a more likely case for logistic regression- the two intervals are different and the Wald interval is not reliable (McCullough and Vinod 2003).

Many solvers merge the concepts of stopping rule and convergence criteria, which make it difficult to know whether optimality conditions meet at a reported maximum. Commonly used convergence criteria are relative parameter convergence, zero gradient, objective function, but some of these criteria may report a solution at false maximum. For instance, if relative parameter convergence and function value are used for the convergence criterion, solver can stop and report "convergence" even though the gradient is far from zero (Train 2003). Rose and Smith (2002) examined an ARCH model where the program reports convergence based on the value of the objective function, a very small change in log-likelihood from 243.5337516 to 243.5337567 changes a nonzero gradient to a zero gradient, but the problem is that a reported zero gradient does not imply an optimum or even a saddle point and for each component of gradient can be numerically zero, but the solution may not be a valid solution, it may due to a very flat region of the surface (McCullough and Vinod 2003).

These situations show that available software packages are not foolproof. Thus, users should not only rely on a software-generated message, they should carefully examine the solution. McCullough and Vinod (2003) suggest the following steps for verifying a solution obtained from nonlinear solver: (1) examine the gradient- is it zero? Relying only on the default setting may give false convergence; (2) examine the Hessian using an eigenvector analysis to determine whether it is positive (negative) definite for minimization (maximization). This helps to identify if the local maximum of the likelihood occurs in a flat region of the log-likelihood function; (3) profile the likelihood function to examine the adequacy of the quadratic approximation. The measure of variability in estimates are reflected in the standard errors that can be produced by nonlinear routines as t-statistics or Wald statistics. The Wald statistic is helpful to assess the adequacy of quadratic approximation, and (4) solution path: does the solution path show the expected rate of convergence? This information helps to find whether there is problem on nonlinear solvers. These are useful methods to guard against less accurate results, but these methods are not accommodated by many packages. Some packages do not allow the user to display gradient, though gradient is used for optimization.

McKenzie and Takaoka (2003) used two different tests to examine the numerical accuracy of LIMDEP 8 using StRD following the tests suggested by McCullough (1997). The authors reported only two potential problem areas out of twelve tests. LIMDEP has problem on identification of probit/logit models (Test E9) and picking up perfectly collinear datasets (Test E13). In Test E9, an error message is necessary because not all the probit model's parameters are identified. The LIMDEP version 7 produces the error message ‘0/1 choice model is inestimable, bad variable = SE’. For the Test E13 that uses singular matrix, LIMDEP provides a warning stating a condition number, but it still estimates some parameters. To some extent, this error message is helpful to understand identification issue. A reliable algorithm should give error message if it cannot estimate a model.

Research has shown that analysts who often run solvers with the default setting, the result are found to be less accurate and software differs in ability in producing accurate results. For example, Stokes (2004) showed the quite different results for solving a probit model in SAS, RATS, LIMDEP, and STATA. The reason for the different results is the default convergence level built into the packages.

Most software packages use approximation for hypothesis testing for discrete choice analysis assuming the test statistics follow a normal or a chi-square distribution. When the assumptions on the test statistics are satisfied, exact and approximate methods give similar results and one can reach to the same conclusions. However, when the assumption of a normal or a chi-square distribution on the test statistics are not satisfied, the exact test should be applied because the approximate test may yield invalid results (Oster 2002).[10] Conventional approximation does not work well when data are unbalanced or thinly distributed into many categories. Conventional approach for estimating logistic regression depends on asymptotic

[10] Exact test can be done in the following ways: permute data in all possible ways under the null hypothesis that is being tested and compute the value of the test statistic for each permutation. Finally, compare the test statistics between the observed value and the permuted distribution. Based on the associated p values, one can conclude whether the results are statistically significant.

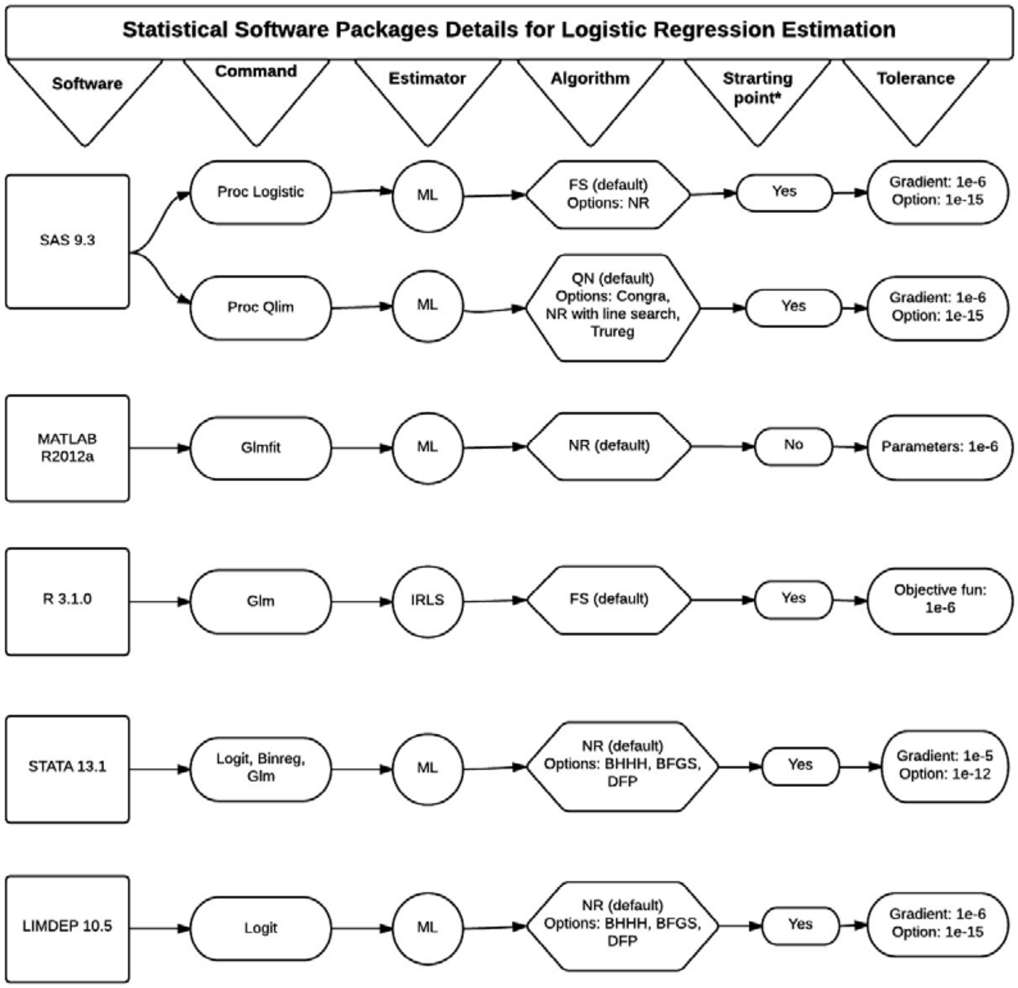
maximum likelihood inference. But software packages based on this kind of inference may provide incorrect results, or may fail to report results, especially if mutlicollinearity exists or when there are many independent variable relative to sample size (Oster 2002). Oster (2002, 2003) used exact methods covering the one-sample, two dependent samples, two independent samples, stratified sample etc. to examine StatXact 5, LogXact 4.1, Stata 7, Testimate 6, and SAS 8.2. The author compared the packages based on hardware requirements, documentations, categorical data analysis etc. The author recommends StatXact for categorical data analysis. Likewise, Huber and Train (2001) examined the similarities and differences between classical and Bayesian methods for mixed logit and found that Bayesian approach has benefits for numerical accurate results for the small sample. However, this study was not intended to examine the reliability of software.

A recent study by Chang and Lusk (2011) compared the maximum likelihood estimator for SAS 9.2, LIMDEP 9 (contain NLOGIT 4), and Hole's model (a user written add- in module) for STATA 11 for the mixed logit model using Monte Carlo simulation. They used the default algorithm and tolerance level for each software package. The results show that SAS and NLOGIT always converges across the 500 iterations, but Stata had poor non-convergence when sample sizes were below 200 observations. They reported a tendency for bias for small size ($N = 200$) due to a great variability in the ML parameter estimates. In addition, the results for root mean squared error (RMSE) are similar among the three packages for large samples, but RMSEs are different for small samples. Overall, among these packages, NLOGIT generated the smallest RMSE. However, their procedures are unable to examine the accuracy of the different mixed logit procedures due to the lack of variation in data to identify the parameter estimates empirically (Hole 2011). In addition, this study did not use certified values to compare estimates obtained from the software packages.

The reliability study does motivate software developers to improve packages, but many deficiencies still exist. Several past studies have examined these packages (SAS 9.3, MATLAB R2012a, R 3.1.0, Stata/IC 13.1, and LIMDEP 10.5.) along with random number generator, linear and nonlinear regressions, but the numerical reliability of logistic regression estimation with the benchmark datasets has not been examined to the authors' knowledge.

3 Research Methods

The paper examines the numerical reliability of these software packages: SAS 9.3, MATLAB R2012a, R 3.1.0, STATA/IC 13.1, and LIMDEP 10.5, for both default and optimal user settings within the software packages (algorithm, starting value, convergence level) examined for each benchmark dataset created. The logistic regression estimation commands/procedures for each software package are illustrated in Figure 1. The figure shows the command/procedure examined; the type of estimator used; algorithms available; if the user can specify starting values for the parameter estimation; and available tolerance criteria and settings. The remaining sub-sections present the logistic regression model; benchmark dataset creation; and procedures for testing software reliability.



NR (Newton-Raphson), BHHH (Berndt-Hall-Hall-Hausman), BFGS (Broyden-Fletcher-Goldfarb-Shanno), DFP (Davidon-Fletcher-Powell), FS (Fisher's Scoring), IRLS (Iteratively Reweighted Least Squares) , QN (Quasi Newton), CONGRA(Conjugate Gradient), TRUREG (Trust Region Optimization), ML (Maximum likelihood)
* This option indicates that if the specified procedure allows the user to specify the starting point. The convergence criteria are (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. e.g. STATA default tolerance setting for NR algorithm is nrtol (1e-5) and options are qtol (1e-5).

Figure 1: Study Design for the Numerical Reliability of Software

3.1 Logistic Regression Model

The generalized linear model (GLM) provides a broad family for statistical analysis. The family of models allows for the estimation of a regression with response data (Y) from the exponential family, which includes the binomial, Poisson, geometric, negative binomial, exponential, gamma, normal, and inverse normal distributions (Myers et al. 2012). This paper focuses on the binomial distribution (binary response variable). A binary response variable (Y) has the probability of success $P(Y = 1) = p$ and the probability of failure $P(Y = 0) = 1 - p$ and Y is assumed to be Bernoulli distributed, which it the binomial distribution with one trial.

The logistic regression model utilizes the logistic CDF to ensure that the conditional mean or $P(Y = 1)$ is bounded between 0 and 1. In addition, following the approach in GLM it is assumed the linear predictor is linear in the parameters, but can include nonlinear covariate terms. Thus, the logistic regression takes the following general functional form:

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

where $\eta(\boldsymbol{X_i}; \beta)$ is the linear predictor (or index function); $\boldsymbol{X}_i$ is a set of explanatory variables; β is a set of parameters to be estimated; and u_i is an IID error term (Bergtold et al., 2010). Greene (2002) states that the error term for the logistic regression model has the following properties:

- Mean of error is zero;
- Error terms are independent but not normally distributed; and

- Variance of the errors depends on the explanatory variables, and by construction are heteroskedastic.

3.2 Model Estimation

The logistic regression model can be estimated using the method of maximum likelihood (Greene 2002). The likelihood function for the logistic regression can be given as follows:

$$L(Y, \boldsymbol{X}; \beta) = \prod_{i=1}^{N} \left[\Lambda\left(\eta(\boldsymbol{X_i}; \beta)\right)\right]^{Y_i} \left[1 - \Lambda\left(\eta(\boldsymbol{X_i}; \beta)\right)\right]^{1-Y_i} \tag{1}$$

Where $L(.,.;.)$ is the likelihood function, Λ represents the logistic cumulative density function, and N is the sample size. We maximize the log-likelihood function rather than the likelihood function because the log function is monotonically increasing and easier to work with it (Greene 2002). The simplified log-likelihood function ($LL(\beta)$) in terms of parameters can be written as follows:

$$LL(Y, \boldsymbol{X}; \beta) = \sum_{i=1}^{N} \left[Y_i \ln\left(\Lambda\left(\eta(\boldsymbol{X_i}; \beta)\right)\right) + (1 - Y_i) ln\left(1 - \Lambda\left(\eta(\boldsymbol{X_i}; \beta)\right)\right)\right] \tag{2}$$

where. In model estimation, we find the estimated value of β that maximizes $LL(Y, \boldsymbol{X}; \beta)$. The value of log-likelihood is always negative because the likelihood is a probability between 0 and 1 and the logarithm of any number between 0 and 1 is negative (Train 2003).

As the log-likelihood function is non-linear in the parameters, a closed-form solution for the esitmators of the parameters is not available. Thus, iterative numerical methods mut be used to maximize the log-likelihood function and obtain parameter estimates. Numerical methods employed will utilize optimization algorithm for estimation, which have various components, including starting value, choice of optimization algorithm, gradient calculation procedures, and termination or stopping rule criteria (Train 2003). Each of these components may have an affect on the numerical accuracy and reliability of model estimation.

3.3 Optimization Algorithms

Many estimation procedures maximize some kind of function, such as the log-likelihood function, the simulated likelihood function, or squared moment conditions (Train 2003). In logistic regression, an optimization algorithm maximizes the log-likelihood function and the optimization is achieved iteratively. An optimization algorithm chooses an initial value and generates a sequence of better values than the initial one until the algorithm finds an optimal solution determined by the termination criteria or stopping rule. The strategy used to move from one value to the next differs from one algorithm to another. Most strategies use the values of the objective function, the constraints and the first and second derivative of these functions to reach a solution. Some algorithms accumulate information gathered at previous iterations, whereas others use only local information from at the current point. Likewise, some algorithms reach an optimal solution in a few iterations, while other may take many iterations. For example, if the log-likelihood function is quadratic in β, then the Newton Raphson algorithm will attain the maximum in one step for any starting value. However, most log-likelihood functions are not quadratic, and hence the Newton Raphson algorithm needs more than one step to find an optimal solution (Train 2003). Regardless of these specifics, a good algorithm is supposed to have the properties of robustness, efficiency, and accuracy. A robust algorithm is one that should perform well on a broad variety of problems for many reasonable starting values. An efficient algorithm means it should not take an excessive amount of computer time and memory (storage). An accurate algorithm means it should not be over-sensitive to data problems or arithmetic from rounding errors. However, there is usually a trade-off between efficiency and accuracy, as well as between storage requirements and convergence rates (Nocedal and Wright 2000).

There are many numerical methods available for maximizing the log Likelihood function for in the method of maximum likelihood for the logistic regression model. The most widely used optimization method (algorithm) is Newton Raphson (NR) (Train 2003), with other populat algorithms including the Broyden-Fletcher-Goldfarb-Shanno (BFGS), Berndt-Hall-HallHausman (BHHH), Davidon-Fletcher-Powell (DFP) and Fisher Scoring (FS). Each algorithm requires stopping or terminating criterion to optimize the objective function. Researcher can set one or more of the following convergence criteria. For example, in the PROC LOGISTIC and PROC QLIM commands in SAS, the default gradient convergence is set equal to $1E-8$, but researchers can change it, but in some packages like MATLAB and R, the convergence criteria cannot be changed. In general, an algorithm uses one (or more than one) of the following stopping criteria:

1. $|\log(L(\beta_{n+1}) - \log(L(\beta_n))| < \varepsilon$

The successive change in the log-likelihood values should be less than the convergence criterion (ε) - a very small number.

2. max ($|\beta_{n+1} - \beta_n|$) < ε

   The successive change in the parameter values is less than the convergence level.

3. $g^{\mathrm{T}}(-H^{-1})g < \varepsilon$ where g is the gradient and H is the Hessian of $LL(Y, \boldsymbol{X}; \beta)$. This criterion measures the size of the gradient using the Hessian and is assumed equal to zero (met) when it is less than the set convergence level.

4. $||g(\beta_n)|| < \varepsilon$

   The magnitude (as measured by the norm) of the gradient is less than the set convergence level.

First derivative and second derivative methods are used in finding the optimal solution. The main difference between them is that first derivative methods do not calculate the Hessian matrix, whereas second derivate methods need to calculate Hessian matrix or approximations of it. In addition, there are derivative free methods (also referred to as grid search methods) that can utilized, as well. First derivative methods, in general, need less computer time than second derivative and grid search methods to find the optimal solution. However, methods that do not use the Hessian tend (or second derivatives) tend to be less reliable. We examine a number alternative algorithms in this study.

3.4 Data Generation and Data Properties

This section explains Statistical Reference Datasets, data generation process, and the process to obtain certified values. Data were generated in MATLAB and certified values for parameter estimates and associated standard errors were created in Mathematica.[11]

The National Institute of Standards and Technology (NIST) develops Statistical Reference Datasets (StRd), which are designed for benchmarking statistical packages. The purpose of StRD is to provide reference datasets to objectively examine the accuracy of statistical software (National Institute of Standards and Technology 2014). The StRD dataset archives include the five main suites.[12] In each suite, datasets are ordered by level of difficulty (low, average, and high) and the level of difficulty depends on algorithm (technique).[13]

The nonlinear benchmark tests can also be used to test maximum likelihood estimator (MLE), however, they are not designed for them (Altman, Gill, and McDonald 2004). Thus, we created thirty benchmark datasets for the logistic regression model using maximum likelihood estimation. Thirty benchmark datasets were created by simulating different conditional binary choice processes in MATLAB

[11] Data descriptions and models are reported in Appendix.

[12] The five suites of datasets are Analysis of Variance, Linear Regression, Nonlinear Regression, Markov Chain Monte Carlo, and Univariate Summary Statistics.

[13] Datasets classified in to different difficulty levels provide a rough guide to users. According to the NIST, a package that can solve a dataset (or the model) with the high level of difficulty does not guarantee the given package can solve the dataset with the low and average levels of difficulty and vice- versa.

following Bergtold, Spanos, and Onukwugha (2010). A table of summarizing the datasets and parameters used are presented in the Appendix at the end of the paper. 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 in the covariates. 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 nonlinearities into the predictor/index function; and creating datasets that quasi-separable. For example, for near mutlicollinearity, datasets were generated with different numbers of observations (e.g. 50 to 1000) changing the degree of correlation between covariates from 0.75 to 0.995. Collinearity is a significant challenge in environmental or 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%. For example, the cutoff value for the cutoff4 data was set at 0.015%, which may represent a situation of modeling credit card fraud.

To obtain certified values to test the numerical reliability of logistic regression estimation, we followed National Institute of Standards and Technology procedures when they generated certified values of parameter estimates and standard errors for nonlinear models estimated using nonlinear least squares procedures (National Institute of Standards and Technology 2014). Certified values were generated through self-coded logistic regression estimation procedures with analytic derivatives using the Mathematica software optimization algorithms, given their high degree of reliability and accuracy. McCullough (2000a) showed that Mathematica can obtain perfect accuracy on the NIST StRD. Certified values were obtained simulating quad precision (128 bits) by doing all computations to 50 significant digits to minimize the overflow and round-off errors during intermediate computations.[14]For each benchmark dataset, the associated logistic regression model was estimated using three alternative algorithms (a Quasi-Newton, conjugate gradient, and grid search algorithm). Standard errors were estimated using the common estimator found in most software packages. Following Greene (2002), the asymptotic covariance matrix is estimated by using the inverse of the Hessian evaluated at the maximum likelihood estimates of the parameters. Certified values for parameter estimates and standard errors were confirmed when two different algorithms for two different starting values in Mathematica agreed on the first 11 significant digits for each parameter estimate.[15] Analytic derivatives were used in estimation because it is more accurate than their finite difference approximation (McCullough 1998).

The two starting values used are the null vector with the value for intercept replaced by the unconditional log odds and the parameter estimates from the linear probability model. A set of starting values can be changed by users in some logistic regression commands, for example, PROC LOGISTIC and PROC QLIM in SAS, whereas

[14] Overflow error occurs when the computer handles a number that is too large for it because each computer has well-defined range of values that it can store or represent. Rounding error occurs when the calculated approximation of a number and its exact value are different.

[15] The datasets descriptions, certified and estimated values for the logistic regression model for a sample dataset are given in Appendix.

for GLM in MATLAB only the default starting values can be used in estimation. Since convergence is sensitive to starting value in logistic regression estimation, changing starting values might give higher LREs, which indicates that a default setting is not the optimal setting for that problem.

3.5 Measure of Accuracy

To examine the numerical reliability of statistical software, logarithm of relative error (LRE) can be used as a measure of accuracy (McCullough 1998; McCullough and Wilson 1999; Odeh, Featherstone, and Bergtold 2010). The LRE measure is used to assess the reliability of estimated results for the above mentioned statistical packages. The LRE measure is given by the following formula:

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

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 ($LAE = -\log_{10}|q|$) is used. The LRE value measures the number of significant digits of the estimated results in comparison to the certified values.[16]For example, an LRE value of 6.5 exhibits that the estimated result is accurate to the six significant digits. If the program reports the negative LRE value, then it is considered

[16] The significant digits in a number are defined as the first nonzero digit and all succeeding digits. For example, 5.314 has four significant digits whereas 0.00029 has only two significant digits.

as zero. For a non-linear model, it is expected to have a minimum LRE score of four (McCullough 1998; Odeh, Featherstone, and Bergtold 2010), which is used as a measure of accuracy of significant digits of the estimated results.

For each benchmark dataset generated, a number of parameter estimates and standard errors are obtained, we calculate LRE for each parameter and standard error. Then we choose the minimum LRE for parameter estimate and standard error for each setting (e.g. algorithm, starting value, convergence level), and compare the minimum LRE to different settings. We first examine the default settings and then examine the user settings that can be changed to find an optimal user setting by examining different algorithms and tolerance criteria, which may be different from the default. If a package's default setting gives lower LRE values than optimal user setting, then it indicates that a default setting did provide as robust estimation results compared to optimal user settings, which is a very likely the case for logistic regression (McCullough and Vinod 2003). Users can obtain more numerically accurate results by changing default settings of a package.

Based on a minimum accuracy (an LRE score) for thirty datasets, this study reports how many datasets meet the minimum accuracy of four digits (an LRE score of four) for parameter estimates and standard errors. An accuracy of four digits (an LRE score of 4) is assumed here to provide minimum reliability for estimation of nonlinear models (McCullough 1998; Odeh, Featherstone, and Bergtold 2010). It may be the case that higher accuracy is needed, and a minimum LRE of 6 could be considered instead. When default settings fail to provide the minimum requirement or low reliability, then users should consider changing algorithmic settings, such as use of an alternative algorithm, change in starting value, lowering of tolerance criteria or use of analytic derivatives. A program that fails to give accuracy to four digits for simple problem may give less accurate results for more difficult problems (for this study, multicollinearity or quasi-separation datasets). These results (comparison of an LRE score for different setting) provide information about the strengths and weaknesses of statistical software, which give flexibility to researchers to choose optimal settings and right software based on their problems, and allow software vendors to potential reliability issues in newer versions of the software.

4 Results

We estimated the thirty benchmark datasets for each software package at two alternative starting points at the default and optimal user settings (as determined by the authors). If a software package has more than one procedure that estimates logistic regression models, then report results for each procedure. Results for minimum LRE for default and optimal user settings are reported in Tables 1 to 4. In Tables 2 and 4, which present the results for the user optimal settings, the default values are again reported as there are no setting to adjust by the end user. The default in this case is the user optimal. A value of NS indicates that an algorithm achieved false convergence, or no optimal solution was obtained for the particular setting (i.e. the model did not estimate). An LRE score of four or greater is required to meet the minimum standard of reliability of nonlinear regression estimation following McCullough (1998).

For parameter estimates, the LOGIT command in STATA, R, and MATLAB with default setting met McCullough's (1998) criteria of minimum LRE of four in twenty-six out of thirty benchmark datasets. Similarly, twenty and seventeen datasets for starting values one and two met the minimum LRE criteria for PROC LOGISTIC in SAS. Likewise, the LOGIT command in LIMDEP met the criteria for twenty-seven and twenty-six models using the default setting for starting values one and two, respectively. Similar results were found for the minimum LRE scores for the estimated standard errors.

On average, the optimal user settings were able to improve algorithmic performance for estimation of models. In some cases, it allowed a procedure for a given statistical package to meet the minimum LRE score. Consider the Cutoff5 dataset for PROC LOGISTIC in SAS. For both starting points, the minimum LRE for the parameter and standard error estimates were 2.8 and 4.3 respectively. When optimal user settings were used by lowering the tolerance criteria, the minimum LRE for CUTOFF 5 for both starting points for the parameters and standard errors increased to 7.4 and 8.5, respectively.

The results for each statistical package examined are discussed in more detail below.

STATA

All thirty datasets were estimated with LOGIT, BINREG, and GLM in STATA. Users can change convergence criteria, algorithm, and starting value in STATA for each procedure. Four algorithms (NR, BHHH, BFGS, and DFP), two tolerance criteria (1e-5 (default) and 1e-12) and two starting values (start one and start two) were examined for to determine optimal user settings.

This software packages performed well for parameter estimates based on the LRE criteria in twenty six out of the thirty benchmark datasets models for the LOGIT and GLM commands for starting values one and two. However, for the BINREG command using the default setting with starting values one and two, only fourteen and fifteen datasets met the criteria, respectively. Finding user optimal setting, twenty six models out of the 30 benchmark datasets met the minimum LRE criteria. Likewise, for standard errors with PROC LOGISTIC, twenty five and twenty six models met the LRE criteria for start one and two. And with PROC QLIM for starting values one and two, eight and nine datasets did not meet the minimum LERE criteria.

MATLAB

All the models were estimated with GLMFIT in MATLAB. Users cannot change algorithm, convergence criteria, or starting values. This software package reliably estimated twenty-six and twenty-seven models out of thirty for the parameter estimates and associated standard errors. One of the shortcomings of the GLMFIT procedure in MATLAB is that it only provides the use of one algorithm (Newton Raphson) to estimate the logistic regression model.

R

This software packages uses the GLM command for estimation of logistic regression models. For starting values one and two, twenty-six and twenty-five models met the minimum LRE criteria for parameter estimates, respectively. However, for standard errors only twenty-three datasets met the criteria for both starting values. R provides a limited number of user options to control estimation. For the GLM command, users can only change starting values.

LIMDEP

LIMDEP uses the LOGIT/BLOGIT command to estimate logistic regression. This software package reliably estimated twenty-seven and twenty-six of the thirty models for starting values one and two, respectively using the default settings. LIMDEP estimated twenty-seven models of the thirty models reliably with user optimal setting for the both starting values. Likewise, twenty-eight models met the minimum LRE criteria in user optimal setting for estimation of standard errors for both starting points.

SAS

All the models were estimated with PROC LOGISTIC and PROC QLIM in SAS. For parameter estimates obtained using PROC LOGITISTIC with default settings, out of the thirty benchmark models, only twenty and seventeen models met the minimum LRE of four for starting points one and two, respectively. In contrast, twenty-six models were reliably estimated using user optimal settings for both starting values. Similarly, with PROC QLIM, twenty-five models of the thirty models met the criteria in default setting for both starting values, whereas twenty-six and twenty-five models met the criteria with user optimal setting for starting points one and two, respectively.

The results for LRE score show that for most of the packages, more datasets met the reliability of software criteria in user optimal setting than default setting. Even if the same number of models were estimated reliably in default and user optimal settings, software gives higher LREs when using user optimal setting. For example, with the Base benchmark dataset, the LRE score for starting value two using the default setting in LIMDEP was 6.6, but with user optimal setting it increased to 10.6.

5 Conclusion

The reliability of estimating logistic regression models for five widely selected software packages used by applied economists and researchers in other disciplines was examined. The packages were SAS 9.3, MATLAB R2012a, R 3.1.0, STATA/IC 13.1, and LIMDEP 10.5. To test reliability we developed 30 benchmark datasets following the procedures established by NIST for their nonlinear regression benchmark datasets. For each statistical package, we estimated the 30 benchmark datasets to estimate the associated logistic regression models for different procedures in each software package. The reliability of the software packages and associated procedures was assessed using the minimum LRE of the parameter and asymptotic standard error 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 a user optimal setting to try and get closer to the certified values generated for each benchmark dataset. In reality, the certified benchmark values will be unknown, thus modelers and researchers should follow the suggestions by McCullough (2004) to verify their results.

Software reliability testing results suggest that many of the logistic regression estimation procedures in the software packages tested are reliable, meeting the minimum LRE requirement of 4. STATA, MATLAB, R, LIMDEP, and SAS for the most part provided consistent reliable results. Users should be careful when relying on default settings. The BINREG procedure in STATA and PROC LOGISTIC in SAS both performed comparatively worse when using the default settings. When user optimal settings were determined by changing tolerance criteria and algorithm choice, reliability results significantly improved for both estimation commands. Overall, user optimal 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. This was the case for both 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 30 unique benchmark datasets with certified parameter and standard error estimates for reliability testing that can be used to test future versions of and other statistical software packages in the future.

Table 1. Minimum LRE for Parameter Estimates with Default Settings

	STATA						MATLAB	R		LIMDEP		SAS			
	LOGIT		BINREG		GLM		GLMFIT	GLM		LOGIT		LOGISTIC		QLIM	
Datasets	D1	D2	D1	D2	D1	D2	D	D1	D2	D1	D2	D1	D2	D1	D2
Base	10.9	6.7	7.3	8	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	5.2	5.4	9.7	9.4	10.5	9.7	10.7	9.7	10.7	4.7	5.2	6.8	10.3
Multico3	7.1	7.9	5.1	4.6	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	4.1	3.3	7	7.1	10.2	10.2	10.2	10.2	10.2	5.5	5.3	8.6	7.8
Multico5	7.1	6.9	3.9	4.5	7.1	6.9	10.5	9.3	9	9.3	9	4.6	4.4	8.9	8.1
Multico6	7	6.6	5.4	4.8	7	6.6	10.2	6.9	6.9	6.9	6.9	6.9	6.9	8.1	8.8
Multico7	6.4	6.4	5.8	5	6.3	6	10.5	9.7	10.5	9.7	6.6	5	6.6	6.6	6.3
Multico8	10.1	9.6	5.4	4.3	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	4.3	3.3	5.9	5.8	10.3	6	9.6	6	9.6	6	6	6.7	6.3
Multico10	4.8	4.8	1.3	0.2	4.8	4.8	10.4	9.1	10.4	9.1	10.4	4.8	4.6	3.9	4.4
Multico11	5.4	7.5	3.9	4.4	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	5	4.8	8.2	8.1	10.6	9.3	10.6	9.1	7.2	4.4	7.2	7.4	8
Multivar2	7.5	7.5	4.1	4.2	7.5	7.5	10.6	9.5	8.6	1.7	1.7	4.5	4	8	7.1
Multivar3	10.3	7.7	5.1	5.3	10.3	7.7	10.3	10.3	7.7	6	7.7	6	3.4	7.5	8.6
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	2.8	2	5.1	5.1	10.3	8	0	8	0	3.4	3.7	5.5	5.2
Multivar6	6.7	6	3	3.2	4.5	6.1	10.3	7.7	7.5	7.7	7.5	3.7	3.6	4.4	4.4
Multivar7	7.1	7.1	5.9	5.5	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	2	3.3	3.3
Cutoff2	5.7	5.7	3.7	4.2	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	6.1	7.4
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	5.7	5.7	4.6	4.6	10.3	6.9	6.9	6.9	6.9	2.8	2.8	9.6	5.7
Cutoff6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cutoff7	8.4	9	NS	NS	8.4	9	10.8	6.7	6.6	6.7	6.6	3.1	3.1	6.5	5.4
Cutoff8	5.9	5.7	2.9	2.9	6.1	5.8	10.4	7.7	7.6	7.7	7.6	3.7	3.7	5.1	6
Empirical1	6	10.7	2.6	2.9	6	10.7	10.7	10.7	10.7	6.4	6.6	6.4	6.3	5.7	5.7
Quasisep1	6.5	6.6	5.3	5.3	5.2	5.3	10.3	8	10.3	8	7	4.1	7	8.3	8.2
Quasisep2	3.6	3.6	0	0	3.6	3.6	3.6	3.6	3.6	10.2	9.2	3.6	3.6	3.6	3.6

Note: D = Default Settings; 1 = Starting Point 1; 2 = Starting point 2, NS= did not converge to a solution

Table 2. Minimum LRE for Parameter Estimates with User Optimal Settings

	STATA						MATLAB	R		LIMDEP		SAS			
	LOGIT		BINREG		GLM		GLMFIT	GLM		LOGIT		LOGISTIC		QLIM	
Datasets	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	10.3
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	7.1	7	7.1	7	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	9.3	9	9.3	9.3	10.4	9.4
Multico6	7	7	7	7.2	7	7	10.2	6.9	6.9	10.2	10	10.2	10.2	8.1	8.8
Multico7	6.5	7.4	6.7	6.2	6.7	6.2	10.5	9.7	10.5	9.7	7.9	10.6	10.5	7.9	7.2
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	8.3
Multico9	5.9	5.8	6.4	5.8	6.4	5.8	10.3	6	9.6	9	9.6	6	6	6.7	9.3
Multico10	5.4	5	5.6	5.6	5.4	5	10.4	9.1	10.4	9.1	10.4	9.1	10.4	8	9
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	9.7	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	8.6
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	8	7.5	8	8.1	9.1	9
Multivar6	6.7	6.3	6.6	6.5	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	5.8	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	3.3	3.7
Cutoff2	6.5	6.4	6.5	7.3	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	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	8	10.4	10.4	9	9
Cutoff5	4.5	4.6	5.7	5.7	4.6	4.7	10.3	6.9	6.9	6.9	10.6	7.4	7.4	9.6	7.3
Cutoff6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cutoff7	8.4	9	8.4	9	8.4	9	10.8	6.7	6.6	9.8	10.2	6.7	6.6	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	10.8	10.7
Quasisep1	6.5	6.6	5.4	6.6	6.6	6.6	10.3	8	10.3	9.1	9.1	8	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

Note: D = Default Settings; U = User Optimized Setting; 1 = Starting Point 1; 2 = Starting point 2, NS= did not converge to a solution, MATLAB has no user optimal setting.

Table 3. Minimum LRE for Standard Errors with Default Settings

	STATA						MATLAB	R		LIMDEP		SAS			
	LOGIT		BINREG		GLM		GLMFIT	GLM		LOGIT		LOGISTIC		QLIM	
Datasets	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	9.4	9.9	9.5	9.9	9.5	10.7	5	5.4	9.9	11	5	5.4	4.7	4.7
Multico3	7	7.4	6.7	7.2	6.7	7.2	6	4.1	5.5	8.2	10.7	4.1	5.5	3.1	3.1
Multico4	8.6	7.9	8.4	7.9	8.4	7.9	10.1	6	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	7	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	5.6	5.6	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	7.3	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	7.3	9	7.3	9	10.2	7.3	4.5	7.3	9	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	4.6	4.6	3.7	3.7	3.7	3.7	9.3	4.2	4	8.4	8	4.2	4	5.7	3.9
Multivar5	7.3	7.3	7.3	7.3	7.3	7.3	9.9	5	0	9.7	NS	5	5.4	5.1	5.2
Multivar6	0.1	0.1	6.7	6.5	6.7	6.4	8.9	4.1	4.1	8.1	8	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	10.6	10.4	10.6	10.4	10	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	4.9
Cutoff3	5.7	5.7	5.8	5.8	5.8	5.8	9.5	3.8	3.8	7.5	7.5	7.5	7.5	4.1	4.1
Cutoff4	5.1	5.1	5.2	5.2	5.2	5.2	8.6	6.7	6.7	6.7	6.7	6.7	6.7	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	8.5	4.3	4.3	4	4
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	7.1	7.1	3.6	3.5	4.8	4.8
Cutoff8	6.1	6.2	6.3	6.1	6.3	6.1	10	4	4	8.1	8	4	4	4.9	4.9
Empirical1	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	5.7	5.7	4	4	4	4	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

Note: D = Default Settings; U = User Optimized Setting; 1 = Starting Point 1; 2 = Starting point 2, NS= did not converge to a solution

Table 4. Minimum LRE for Standard Errors with User Optimal Settings

	STATA						MATLAB	R		LIMDEP		SAS			
	LOGIT		BINREG		GLM		GLMFIT	GLM		LOGIT		LOGISTIC		QLIM	
Datasets	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	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	4.1	5.5	8.8	10.7	8.2	10.8	3.1	3.1
Multico4	8.6	1.4	8.4	1.4	8.4	1.4	10.1	6	5.8	10.1	10.1	10.1	10.1	4.8	4.8
Multico5	8.3	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	2.4
Multico7	5.6	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	7.8	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	10.2	9	10.2	9	10.2	7.3	4.5	10.2	9.7	10.2	9	5.8	5.8
Multivar1	8.7	8.4	9.5	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	9.3	4.2	4	6.3	8	4.2	4	5.7	3.8
Multivar5	7.2	6.8	7.2	6.8	7.2	6.8	9.9	5	0	9.7	9.3	9.8	9.8	5.1	5.1
Multivar6	0.1	0.1	6.7	6.7	6.7	6.7	8.9	4.1	4.1	8.1	8.5	8.1	8	4.9	4.9
Multivar7	6.2	6.2	6.1	0	6.1	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	5.7	5.7	5.8	5.8	5.8	5.8	9.5	3.8	3.8	9.3	8	7.5	7.5	4.1	4.1
Cutoff4	5.1	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	4.7	2.4	5.7	5.7	10.2	4.3	4.3	8.5	10.2	8.5	8.5	4	4
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	7.1	7.1	4.8	4.8
Cutoff8	5.8	5.8	7	6.7	7	6.7	10	4	4	9.9	8.1	8.1	8	4.9	4.9
Empirical1	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	3.9	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

Note: D = Default Settings; U = User Optimized Setting; 1 = Starting Point 1; 2 = Starting point 2, NS= did not converge to a solution, MATLAB has no user optimal setting.

References

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

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* 3:1–28.

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

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

Heinze, G., and M. Schemper. 2002. "A solution to the problem of separation in logistic regression." *Statistics in medicine* 21:2409–2419.

Hole, A.R. 2011. "A comment on mixed logit models: accuracy and software choice." Working paper.

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

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

Kolenikov, S. 2001. "Review of STATA 7." *Journal of Applied Econometrics* 16:637–

646.

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

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

—. 2000b. "Is it safe to assume that software is accurate?" *International Journal of Forecasting* 16:349–357.

—. 1997. "A review of RATS v4. 2: Benchmarking numerical accuracy." *Journal of Applied Econometrics* 12:181–190.

—. 2004. "Wilkinson's tests and econometric software." *Journal of economic and social measurement* 29:261–270.

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

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

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

—. 1999b. "Econometric software reliability: EViews, LIMDEP, SHAZAM and TSP." *Journal of Applied Econometrics* 14:191–202.

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* 93:873–892.

McCullough, B.D., and B. Wilson. 1999. "On the accuracy of statistical procedures in Microsoft Excel 97." *Computational Statistics & Data Analysis* 31:27–37.

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

M´elard, G. 2014. "On the accuracy of statistical procedures in Microsoft Excel 2010." *Computational statistics* 29:1095–1128.

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

Myers, R.H., D.C. Montgomery, G.G. Vining, and T.J. Robinson. 2012. *Generalized linear models: with applications in engineering and the sciences*. John Wiley & Sons.

National Institute of Standards and Technology. 2014. "Statistical Reference Datasets." http://www.itl.nist.gov/div898/strd, (accessed on April 15, 2014).

Nerlove, M. 2005. "On the numerical accuracy of Mathematica 5.0 for doing linear and non-linear regression." *Mathematica Journal* 9:824–851.

Nocedal, J., and S.J. Wright. 2000. *Numerical Optimization*. Springer Series in Operation Research and Financial Engineering.

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

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

—. 2003. "An examination of statistical software packages for categorical data analysis using exact methods—Part II." *The American Statistician* 57:201–213.

Rose, C., and M.D. Smith. 2002. *Mathematical statistics with Mathematica*. Springer New York.

SAS Manual. 2009. "SAS/STAT(R) 9.22 User's Guide." http://goo.gl/GhTdCA, (accessed on July 15, 2014).

Sawitzki, G. 1994. "Report on the numerical reliability of data analysis systems." *Computational statistics & data analysis* 18:289–301.

Silk, J. 1996. "Systems estimation: A comparison of SAS, SHAZAM and TSP." *Journal of Applied Econometrics* 11:437–450.

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

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* 29:307–320.

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

Train, K.E. 2003. *Discrete Choice Methods with Simulation*. Cambridge University Press.

Veall, M.R. 1991. “SHAZAM 6.2: a review.” *Journal of Applied Econometrics* 6:317–320.

Appendix: Description of Benchmark Datasets

Dataset	P (Y_t =1), Cutoff Point	Predictor Functional Form	Multi-collinearity (ρ)	Number of Observation	Variance
Base	P = 0.6	η(X;b) = b0 + b1*X	-	200	$\sigma_1 = 1$
Multicollinearity1	P = 0.6	η(X;b) = b0 + b1*X1 + b2*X2	$\rho_{12} = 0.75$	500	$\sigma_1 = 1.5$ $\sigma_2 = 1.5$
Multicollinearity2	P = 0.6	η(X;b) = b0 + b1*X1 + b2*X2	$\rho_{12} = 0.995$	500	$\sigma_1 = 1.5$ $\sigma_2 = 1.5$
Multicollinearity3	P = 0.6	η(X;b) = b0 + b1*X1 + b2*X2	$\rho_{12} = 0.995$	500	$\sigma_1 = 1.5$ $\sigma_2 = 1.5$
Multicollinearity4	P = 0.6	η(X;b) = b0 + b1*X1 + b2*X2 + b3*X1^2 + b4*X1*X2 + b5*X2^2	$\rho_{12} = 0.75$	1000	$\sigma_{10} = 1$ $\sigma_{11} = 1.5$ $\sigma_{20} = 1.5$ $\sigma_{21} = 2$
Multicollinearity5	P = 0.6	η(X;b) = b0 + b1*X1 + b2*X2 + b3*X1^2 + b4*X1*X2 + b5*X2^2	$\rho_{12} = 0.95$	1000	$\sigma_{10} = 1$ $\sigma_{11} = 1.5$ $\sigma_{20} = 1.5$ $\sigma_{21} = 2$
Multicollinearity6	P = 0.6	η(X;b) = b0 + b1*X1 + b2*X2 + b3*X1^2 + b4*X1*X2 + b5*X2^2	$\rho_{12} = 0.995$	1000	$\sigma_{10} = 1$ $\sigma_{11} = 1.5$ $\sigma_{20} = 1.5$ $\sigma_{21} = 2$
Multicollinearity7	P = 0.6	η(X;b) = b0 + b1*X1 + b2*X2 + b3*X3 + b4*X4	$\rho_{ij} = 0.985 \forall$ i, j= 1..4, $i \neq j$	1000	σ = 1
Multicollinearity8	P = 0.4 Q = 0.70-0.80	η(X;b) = b0 + b1*X1 + b2*X2 + b3*X1*X2	$\rho_0 = 0.3$ $\rho_1 = 0.7$	50	-
Multicollinearity9	P = 0.4	η(X;b) = b0 + b1*X1 + b2*X2 + b3*X3 + b12*X1*X2 + b13*X1*X3+ b23*X2*X3+b123*X1*X2*X3	$\rho_0 = 0.3$ $\rho_1 = 0.7$	400	σ = 1
Multicollinearity10	P = 0.4	η(X;b) = b0 + b1*X1 + b2*X2 + b3*X3 + b11*X1^2 + b12*X1*X2 + b13*X1*X3 + b23*X2*X3 + b112*x1^2*X2 + b113*X1^2*X3 + b123*X1*X2*X3 + b1123*X1^2*X2*X3	$\rho_0 = 0.3$ $\rho_1 = 0.7$	325	σ = 1 - 3

Multicollinearity11	P = 0.4 Q = 0.99	η(X;b) = b0 + b1*X1 + b2*X2 + b3*X1*X2	$\rho_0 = 0.8$ $\rho_1 = 0.4$	89	-
Multivariate 1	P = 0.6	η(X;b) = b0 + b1*X1 + b2*Ln(X2) + b3*X3	-	300	$\sigma = 1$
Multivariate2	P = 0.6	η(X;b) = b0 + b1*X1 + b2*X1^2 + b3*X2 + b4*ln(X2) + b5*X3	-	300	$\sigma_1 = 0.9$ $\sigma_2 = 1.8$
Multivariate 3	P = 0.4	η(X;b) = b0 + b1*X1 + b2*X2 + b3*X3 + b4*X4 + b5*X5	$\rho_{ij} = 0.5 \forall$ i, j = 1-5, $i \neq j$	1000	$\sigma_i = 15$
Multivariate4	P = 0.4	η(X;b) = b0 + b1*X1 + b2*X2 + b3*X3 + b4*X4 + b5*X5	ρ_{ij} =0.3- 0.7$\forall$ i, j = 1-5, $i \neq j$	1000	$\sigma_i = 1$
Multivariate5	P = 0.5	η(X;b) = b0 + b1*X1 + b2*X2 + b3*X3 + b4*X4 + b5*X5 + b11*X1^2 + b12*X1*X2 + b13*X1*X3 + b14*X1*X4 + b15*X1*X5 + b22*X2^2 + b23*X2*X3 + b24*X2*X4 + b25*X2*X5 + b33*X3^2 + b34*X3*X4 + b35*X3*X5 + b44*X4^2 + b45*X4*X5 + b55*X5^2	ρ_{ij} =0.15- 0.35 $\forall$ i, j = 1-5, $i \neq j$	100	$\sigma = 1$
Multivariate6	P = 0.4	Same as the Multivariate5's functional form	ρ_{ij} =0.1- 0.6, i, j = 1-5, $i \neq j$	200	σ = 0.5 - 0. 77
Multivariate 7	P = 0.4	η(X;b) = b0 + b1*X1 + b2*X2	$\rho = 0.75$	9	$\sigma_1 = 0.25$ $\sigma_2 = 0.40$
Cutoff1	P= 0.05	η(X;b) = b0 + b1*X1 + b2*X2	$\rho_{12} = 0.99$	50	$\sigma = 1$
Cutoff2	P = 0.15	η(X;b) = b0 + b1*X1 + b2*Ln(X2) + b3*X3	-	32	$\sigma = 1$
Cutoff3	P = 0.0005	η(X;b) = b0 + b1*X1 + b2*X2 + b3*X3 + b4*X4	$\rho_{12} = 0.96$ $\rho_{34} = 0.96$	5000	σ = 0.2- 2.5
Cutoff4	P = 0.00015	η(X;b) = b0 + b1*X1 + b2*X2 + b3*X3	ρ = 0.7 – (- 0.85)	20000	$\sigma = 1$

Cutoff5	P = 0.05	Same as the Multivariate5's functional form	ρ_{ij} =0.3 – 0.5∀ i, j = 1-5, i ≠ j	500	σ = 1- 2
Cutoff6	P = 0.10	η(X;b) = b0 + b1*X1 + b2*X2 + b12*X1*X2	ρ_0 = 0.2,ρ_1 = 0.9	65	-
Cutoff7	P = 0.10	η(X;b) = b0 + b1*X1 + b2*X2 + b12*X1*X2	ρ_0 = 0.3, ρ_1 = 0.8	17500	-
Cutoff8	P= 0.19	η(X;b) = b0 + b1*X1 + b2*X2 + b3*X3 + b11*X1^2 + b12*X1*X2 + b13*X1*X3 + b23*X2*X3 + b112*x1^2*X2 + b113*X1^2*X3 + b123*X1*X2*X3 + b1123*X1^2*X2*X3	ρ_0 = 0.4, ρ_1 = 0.7	200	-
Empirical1	On-Farm Conservation Practice Adoption Bergtold (2005)	η(X;b) = b0 + b1*X1 + b2*X2 + b3*X3 + b4*X4 + b5*X5 + b6*X6 + b7*X7 + b8*X8 + b9*X9 + b10*X10 + b11*X11 + b12*X12 + b13*X13	-	1081	-
Quasisep1	Generated data following an example out of Ryan (1997)	η(X;b) = b0 + b1*X	-	100	-
Quasisep2	Generated data following an example out of Ryan (1997)	η(X;b) = b0 + b1*X + b2*X^2 + b3*X^3 + b4*X^4 + b5*X^5	-	60	-