Computational aspects of probit model

Eugene Demidenko*

Abstract. *Sometimes the maximum likelihood estimation procedure for the probit model fails. There may be two reasons: the maximum likelihood estimate (MLE) just does not exist or computer overflow error occurs during the computation of the cumulative distribution function (cdf). For example, the approximation explosive effect due to an inaccurate computation of the cdf for a large value of the argument occurs in a popular statistical package S-plus. The goal of the paper is to provide remedies for these two abnormalities. First, despite the availability of a criterion for the MLE existence, expressed in terms of a separation plane in the covariate space, there are no constructive criteria to verify whether such a separation exists. We develop constructive criteria for the MLE existence that are valid also for other link functions. Second, to avoid the overflow problem we suggest approximate formulae for the log-likelihood function and its derivatives in the case of possible large value of the argument. Standard algorithms of the log-likelihood maximization like Newton-Raphson or Fisher Scoring are very sensitive to large values of the linear predictor, particularly outliers. Five algorithms are compared by the time to converge and reliability via statistical simulations. The corrected algorithms, based on the approximate formulae are more reliable and almost as fast as the standard ones.*

Key words: *probit model, cut-off-error, maximum likelihood, existence, binary data, binary model*

AMS subject classifications: 49, 62

Received January 20, 2002 Accepted February 10, 2002

1. Introduction

Models with binary dependent variable are commonly used in statistical applications. The classic books by Cox (1970), Finney (1971), McCullagh and Nelder (1989) provide theoretical background with numerous examples of applications. Several link functions have been proposed in the literature; the most popular are logit and probit, McCullagh and Nelder (1989). Often, the procedure of maximum

*7 927 Rubin Building, Dartmouth College, New Hampshire 03756, USA, e-mail: `eugene.demidenko@dartmouth.edu`

likelihood estimation of the probit model, implemented in many statistical packages, runs smoothly. However, sometimes it fails. The goal of the present paper is to study possible reasons of that failure and provide some remedies.

There may be two reasons of failure to converge during the log-likelihood function maximization: a) maximum likelihood estimate (MLE) just does not exist, b) the argument of the normal cumulative distribution function is too large in absolute value, due to rare/frequent event data and/or the presence of an outlier. In this paper we suggest criteria for the MLE existence and provide some approximate formulae that work well even for a very large argument value.

Criteria for the existence and uniqueness of the maximum likelihood estimate (MLE). An important property of many popular link functions, including probit, is that the log-likelihood is a concave function of parameters. Consequently, under mild conditions the MLE is unique, if it exists. General criteria for the existence of the global optimum of a continuous function defined on a non-compact set are developed in Demidenko (1981, 1996, 2000) and Nakamura and Lee (1993). Haberman (1974) and Weddenburn (1976) investigated numerical aspects of the method of maximum likelihood estimation for binary data in detail. In particular, despite the concavity, it was realized that for some data the MLE may not exist. The issue of the MLE existence in logistic regression was considered by Silvapulle (1981) and Albert and Anderson (1984). Lesaffre and Kaufmann (1992) have suggested a necessary and sufficient condition for the MLE existence in the probit model which, in fact, coincides with that derived by Albert and Anderson for the logistic model. That criterion is formulated in terms of separation of observation points in the covariate space. As was mentioned by Albert and Anderson, in order to prove that the MLE exists one need further to apply some linear programming technique to implement that criterion in practice, i.e. demonstrate that the separation does not exist. Thus, despite the fact that the criterion for the MLE existence in the probit model is known, it is unclear how to realize that criterion in practice. In this paper we suggest a constructive procedure to check whether that separation exists that boils down to a criterion for the existence of a solution to a system of homogeneous linear inequalities. It may be too expensive to apply the necessary and sufficient criterion to every probit model because it is timely consuming; a simple sufficient criterion may work as well.

Large value of the argument. The estimation procedure for probit model may fail in the case of a rare or a frequent event, i.e. for an extreme argument value in the normal cumulative distribution function Φ. For example, Demidenko and Spiegelman (1997) describe an example of a binary model for the Nurses' Health Study, a prospective cohort of 89,538 white married women, 601 of whom developed breast cancer, a rare event. Then, the computation accuracy of the cumulative distribution function Φ is limited by 10^{-7} for single and by 10^{-14} for double precision. For example, if for some observation point the argument of Φ is large then $1-\Phi$ becomes close to zero and the maximum likelihood program crashes due to division by zero. To avoid possible overflow errors we suggest to use approximate formulae to compute quantities like Φ or $1-\Phi$ for a large argument.

The structure of the paper is as follows. In the next section main notations and the Feller approximation are introduced. In *Section 3* we suggest constructive criteria for the MLE existence in binary models. In *Section 4* we investigate the effect of poor approximation of the normal cumulative distribution function and suggest a reliable computational algorithm for the log-likelihood function maximization in case of the extreme argument value of function Φ. Finally, in *Section 5* we compare algorithms via statistical simulations.

2. The log-likelihood function

Let y_i be a binary variable and $\mathbf{x}_i$ be a $k \times 1$ fixed vector of the explanatory variables or covariates, $i = 1, ..., n$. Throughout the paper it is assumed

$$rank\,(\mathbf{x}_1, ..., \mathbf{x}_n) = k, \tag{1}$$

which is referred to as *a full rank condition.* The probit model is based on the binomial distribution of y_i with probability

$$\Pr(y_i = 1) = \Phi(\beta'\mathbf{x}_i), \quad i = 1, ..., n \tag{2}$$

where Φ is the normal cumulative distribution function (cdf)

$$\Phi(s) = \int_{-\infty}^{s} \phi(t)dt,$$

where $\phi(t) = (2\pi)^{-1/2}e^{-\frac{1}{2}t^2}$ is the standard normal density, and β is the $k \times 1$ parameter of interest. Assuming independence of $\{y_i,\ i = 1, ..., n\}$ the log-likelihood function for the probit model is written as

$$l(\beta) = \sum_{y_i=1} \log[\Phi(s_i)] + \sum_{y_i=0} \log[1 - \Phi(s_i)], \quad \beta \in \mathbb{R}^k, \tag{3}$$

where $s_i = \beta'\mathbf{x}_i$. The MLE, $\widehat{\beta}_{ML}$ maximizes function (3). To find the MLE we need the first and second derivatives of (3):

$$\partial l/\partial\beta = \sum_{y_i=1} \phi(s_i)\Phi^{-1}(s_i)\mathbf{x}_i - \sum_{y_i=0} \phi(s_i)(1 - \Phi(s_i))^{-1}\mathbf{x}_i\ , \tag{4}$$

$$\begin{aligned} \partial^2 l/\partial\beta^2 \quad = \quad & -\{ \sum_{y_i=1} [s_i\phi(s_i)\Phi^{-1}(s_i) + \phi^2(s_i)\Phi^{-2}(s_i)]\mathbf{x}_i\mathbf{x}_i' \\ & + \sum_{y_i=0} [\phi^2(s_i)(1 - \Phi(s_i))^{-2} - s_i\phi(s_i)(1 - \Phi(s_i))^{-1}]\mathbf{x}_i\mathbf{x}_i' \}. \end{aligned} \tag{5}$$

Taking the expectation we obtain the $k \times k$ information matrix

$$\mathbf{I}(\beta) = -E(\partial^2 l/\partial\beta^2) = \sum_{1}^{n} \frac{\phi^2(s_i)}{\Phi(s_i)(1 - \Phi(s_i))}\mathbf{x}_i\mathbf{x}_i'. \tag{6}$$

The covariance matrix may be estimated in two ways – as the inverse of the expected or observed information matrix:

$$cov_1(\widehat{\beta}_{ML}) = \mathbf{I}^{-1}(\widehat{\beta}_{ML}), \quad cov_2(\widehat{\beta}_{ML}) = \left(\left. \frac{\partial^2 l}{\partial \beta^2} \right|_{\beta=\widehat{\beta}_{ML}} \right)^{-1}.$$

For further approximations the following facts about the cdf and the standard normal density function will be used:

$$\lim_{s \longrightarrow -\infty} \frac{\phi(s)}{s\Phi(s)} = -1, \quad \lim_{s \longrightarrow \infty} \frac{\phi(s)}{s[1-\Phi(s)]} = 1 \tag{7}$$

$$s\Phi(s) + \phi(s) > 0, \ \phi(s) - s(1-\Phi(s)) > 0 \quad \forall s \in \mathbb{R}^1. \tag{8}$$

The proof of (7) is given by Feller (1957, p. 166). Inequalities (8) can be obtained by a slight modification of that proof. The approximation based on the limits (7) will be used in *Section 4* to calculate the cdf for large value of the argument.

3. Criteria for the existence

In many cases the maximization of the log-likelihood function (3) runs quite smoothly. However, for some data the estimation result seems suspicious: even the criterion of convergence met the maximum of the log-likelihood function is close to zero and the final estimate is unreasonably large in absolute value. Can one trust that result? Maybe the MLE does not exist at all and the obtained convergence is the result of computer inaccuracy and cut-off-errors? To answer this question one has to have a criterion to check whether the MLE exists for particular data. If that criterion shows that the MLE exists, then one could trust the estimation result, otherwise it would indicate that the resulted estimate is false. We stress that the failure of the maximum likelihood procedure itself does not mean that the MLE does not exist because that failure may occur due to the overflow error during computation of the cdf at the large value of the argument (see the next section).

The uniqueness of the MLE, if it exists, for probit model under assumption (1) was proven by Haberman (1974, p. 309) and follows from the fact that functions $\log(\Phi(\cdot))$ and $\log(1-\Phi(\cdot))$ are strictly concave. Notice that for linear regression condition (1) implies the existence of the MLE, unlike probit model. The Hessian of the log-likelihood function (5) under (1) is positive definite, as follows from inequalities (8). Therefore, the log-likelihood is a strictly concave function of β – an important feature of the probit model. Also, the log-likelihood function of the probit model must be negative because $0 < \Phi(\beta' \mathbf{x}_i) < 1$ for all β. Haberman (1974, p. 320) and Weddenburn (1976) formulated the condition for the MLE existence for the probit and logistic models in general terms; Albert and Anderson (1984) did it for the logistic, and Lesaffre and Kaufmann (1992) for the probit model. That necessary and sufficient criterion was formulated in terms of separation of the data points in covariate space. Further, S_0 denotes the index set $\{i : y_i = 0\}$ and S_1 denotes $\{i : y_i = 1\}$.

Necessary and sufficient criterion for the MLE existence of the probit model (Lesaffre and Kaufmann, 1992). The MLE for the probit model exists if and only if there is no $\beta \neq \mathbf{0} \in \mathbb{R}^k$ such that

$$\beta'\mathbf{x}_i \geq 0, \quad i \in S_0 \qquad \text{and} \qquad \beta'\mathbf{x}_i \leq 0, \quad i \in S_1, \tag{9}$$

which will be referred to as a *separation condition.* Geometrically, the MLE does not exist if and only if there is a plane which separates points $\{\mathbf{x}_i, i = 1, ..., n\}$ into two groups according to occurrence/non-occurrence of event y. After Albert and Anderson (1984) points $\{\mathbf{x}_i, i = 1, ..., n\}$ are called *overlapped* if there is no separation plane. If (9) is not true for any β, then, as it can be easily shown, $l(\beta) \to -\infty$ if $\parallel \beta \parallel \to \infty$. Otherwise, $\lim_{\lambda \to +\infty} l(\lambda\beta_0'\mathbf{x}_i) = 0$ where $\beta_0 \in \mathbb{R}^k$ is such that $\beta_0'\mathbf{x}_i \geq 0$ for $i \in S_0$ and $\beta_0'\mathbf{x}_i \leq 0$ for $i \in S_1$, $\beta \neq \mathbf{0}$.

Criterion (9) is illustrated by a one-covariate probit model with the intercept term, $\Pr(y_i = 1) = \Phi(\beta_1 + \beta_2 x_i)$, see *Figure 1.* For this problem $k = 2$ and $\mathbf{x}_i = (1, x_i)'$. Denote

$$M_0 = \max_{y_i=0} x_i, \quad m_0 = \min_{y_i=0} x_i, \quad M_1 = \max_{y_i=1} x_i, \quad m_1 = \min_{y_i=1} x_i.$$

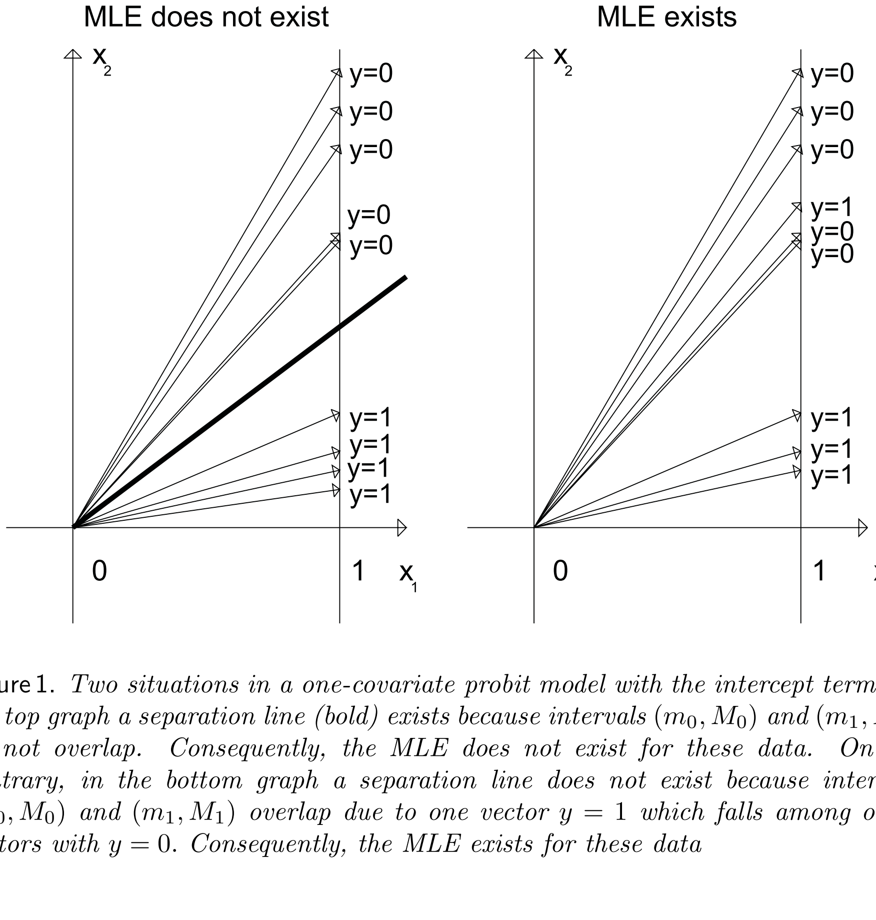

Figure 1. *Two situations in a one-covariate probit model with the intercept term. In the top graph a separation line (bold) exists because intervals* (m_0, M_0) *and* (m_1, M_1) *do not overlap. Consequently, the MLE does not exist for these data. On the contrary, in the bottom graph a separation line does not exist because intervals* (m_0, M_0) *and* (m_1, M_1) *overlap due to one vector* $y = 1$ *which falls among other vectors with* $y = 0$*. Consequently, the MLE exists for these data*

Vectors $\{\mathbf{x}_i, i = 1, ..., n\}$ can be separated by a line if and only if intervals (m_0, M_0) and (m_1, M_1) do not overlap which occurs if $M_0 \leq m_1$ or $M_1 \leq m_0$. Therefore, the MLE for the probit model with one covariate and intercept exists if and only if either $m_0 < m_1 < M_0$ or $m_0 < M_1 < M_0$. In particular, there should be at least one data point with $y = 0$ or $y = 1$ for the MLE to exist. The MLE does not exist if for all data points $y_i = 0$ (or $y_i = 1$), $i = 1, ..., n$. Note that this statement is true only for probit models with the intercept term. The condition on separation is harder to verify for higher dimensions. We aim to construct criteria for the MLE existence in the probit model for an arbitrary parameter dimension in the rest of this section.

To simplify further considerations we introduce the following vectors:

$$\mathbf{v}_i = (1 - 2y_i)\mathbf{x}_i = \left\{ \begin{array}{rl} \mathbf{x}_i \text{ if} & i \in S_0 \\ -\mathbf{x}_i \text{ if} & i \in S_1 \end{array} \right. , \qquad i = 1, ..., n. \tag{10}$$

Most of the statements of the following theorem can be viewed just as certain reformulations of the separation condition (9) in terms of existence of the solution of a system of homogeneous inequalities. The last statement is the well known Gordan's theorem (1873), see also Cottle et al. (1992).

Theorem 1. *The following statements are equivalent:*

(i) MLE for the probit model exists;

(ii) there are no $\beta \neq \mathbf{0}$ such that $\beta'\mathbf{v}_i \geq 0$ for all $i = 1, ..., n$;

(iii) for any $\beta \neq \mathbf{0}$ there is vector $\mathbf{v}_j$ from $\{\mathbf{v}_i,\ i = 1, ..., n\}$ such that $\beta'\mathbf{v}_j < 0$;

(iv) the system of homogeneous linear inequalities $\beta'\mathbf{v}_i > 0$ for β has no solution;

(v) there exist $\gamma_1 \geq 0, ..., \gamma_n \geq 0$ not all equal to zero that $\sum_{i=1}^n \gamma_i \mathbf{v}_i = \mathbf{0}$.

Consequently, if a vector is a linear combination of other vectors with nonnegative coefficients it can be removed from the MLE existence consideration because it does not affect the existence of solution to a system of homogeneous inequalities. The following simple existence criterion is formulated which sometimes works. Before formulating the criterion let us make some comments on the geometry of the Euclidean space $\mathbb{R}^k$. We define an ort-vector as a vector with coordinates $0, 1$ or -1. Then, the space can be divided into 2^k quadrants. An open quadrant can be defined as an open cone spanned by the k neighboring ort-vectors.

Theorem 2. *(Sufficient Criterion I). The MLE exists if every open ortant of $\mathbb{R}^k$ contains a vector from $\{\mathbf{v}_i,\ i = 1, ..., n\}$, or algebraically, for any k dimensional vector $\mathbf{e} = (e_1, ..., e_k)'$ consisting of 1 or -1 there exists a vector $\mathbf{v}_j = (v_{j1}, ..., v_{jk})'$ such that $e_r v_{jr} > 0$ for all $r = 1, ..., k$.*

Proof. We prove that for any vector $\beta \neq \mathbf{0}$ there exists a vector $\mathbf{v}_j$ from $\{\mathbf{v}_i,\ i = 1, ..., n\}$ such that $\beta'\mathbf{v}_j > 0$, using (iii) of *Theorem 1.* Let $\beta \neq \mathbf{0}$ be given. We construct vector $\mathbf{e} = (e_1, ..., e_k)'$ using the following rule

$$e_r = \left\{ \begin{array}{cl} 0, & \text{if } \beta_r = 0 \\ \beta_r/|\beta_r|, & \text{if } \beta_r \neq 0. \end{array} \right.$$

Coordinates of this vector are 0 or ±1 and $\mathbf{e} \neq \mathbf{0}$. Hence there exists vector $\mathbf{v}_j$ such that $e_r v_{jr} > 0$ for all $r = 1, ..., k$ and $\mathbf{e}'\mathbf{v}_j = \sum_{r=1}^k e_r v_{jr} > 0$. But $\beta'\mathbf{v}_j = \sum_{r=1}^k e_r v_{jr}|\beta_r| > 0$, i.e. the MLE exists. □

This criterion is simple but quite restrictive; for instance, for a one-covariate probit model with the intercept term it works only if x takes positive and negative values for $y = 0$ and $y = 1$. A better sufficient criterion is formulated below. To simplify the notation vectors $\mathbf{x}_i$ are supplied with the superindex 0 or 1; thus $\mathbf{x}_i^0$ corresponds to the covariate vector with $y_i = 0$ and $\mathbf{x}_j^1$ corresponds to $y_j = 1$; also $\mathbf{x}^0$ are supplied with the subindex i and $\mathbf{x}^1$ are supplied with the subindex j. As follows from (9) the MLE does not exist if and only if $\{\mathbf{x}_i^0\}$ and $\{\mathbf{x}_j^1\}$ can be separated by a plane.

Theorem 3. *(Sufficient Criterion II). The MLE exists if there exists* $\mathbf{x}_p^0$ *which can be represented as a linear combination of some vectors* $\{\mathbf{x}_j^1\}$ *with positive coefficients, or there exists* $\mathbf{x}_p^1$ *which can be represented as a linear combination of some vectors* $\{\mathbf{x}_j^0\}$ *with positive coefficients.*

Proof. Let vector $\mathbf{x}_p^0$ can be represented as a positive linear combination $\sum \lambda_j \mathbf{x}_j^1$ where $\lambda_j > 0$. We use (iv) of *Theorem 1* to prove that the MLE exists. On the contrary, if β exists such as $\beta'\mathbf{x}_j^1 > 0$ and $\beta'\mathbf{x}_i^0 < 0$, then for vector $\mathbf{x}_p^0$ we have

$$\beta'\mathbf{x}_p^0 = \beta' \sum \lambda_j \mathbf{x}_j^1 = \sum \lambda_j \beta' \mathbf{x}_j^1 > 0,$$

a contradiction. Analogously, we prove the MLE existence if $\mathbf{x}_p^1$ can be expressed as a positive linear combination of vectors $\mathbf{x}_i^0$. □

The following algorithm determines whether the MLE exists, based on *Theorem 3*.

Algorithm 1.

1. Pick a vector $\mathbf{x}_p^0$ from $\{\mathbf{x}_i^0\}$.

2. Pick k linearly independent vectors from $\{\mathbf{x}_j^1\}$.

3. Express $\mathbf{x}_i^0$ as a linear combination of k vectors from step 2 solving the according system of linear equations. If all coefficients of the solution are positive then the MLE exists, and quit. Otherwise, return to step 2 until all k linearly independent vectors from $\{\mathbf{x}_j^1\}$ are enumerated.

4. Return to 1 and pick another vector.

The necessary and sufficient criterion for the MLE existence is formulated next.

Theorem 4. *(Necessary and Sufficient Criterion). The MLE does not exist if and only if there are* $k-1$ *vectors* $\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_{k-1}$ *among* $\{\mathbf{v}_i, i = 1, ..., n\}$ *that all* n *determinants,*

$$D_i = \det \begin{bmatrix} v_{i1} & v_{i2} & \cdots & v_{ik} \\ w_{11} & w_{12} & \cdots & w_{1k} \\ \cdots & & & \\ w_{k-1,1} & w_{k-1,2} & \cdots & w_{k-1,k} \end{bmatrix}, \quad i = 1, 2, ..., n \tag{11}$$

have the same sign.

Proof. Let the MLE not exist, i.e. there exists $\beta \neq \mathbf{0}$ as the solution to n homogeneous inequalities $\beta'\mathbf{v}_i \geq 0$, $i = 1, ..., n$. It is well known that the solutions to the system of homogeneous inequalities is a polyhedral cone C^+ conjugate to the cone C spanned by the vectors $\{\mathbf{v}_i, i = 1, ..., n\}$, e.g. Hoffman (1999). Each edge of C^+ is orthogonal to at least $k-1$ vectors of $\{\mathbf{v}_i, i = 1, ..., n\}$. Hence it is possible to pick a $\beta_* \neq \mathbf{0}$ and $k-1$ vectors $\mathbf{w}_1, ..., \mathbf{w}_{k-1}$ among $\{\mathbf{v}_i\}$ such that $\beta_*'\mathbf{v}_i \geq 0$ for all $i = 1, ..., n$ and $\beta_*'\mathbf{w}_j = 0$ for $j = 1, ..., k-1$. Thus, in searching a plane which separates $\{\mathbf{x}_i^0\}$ and $\{\mathbf{x}_j^1\}$, without loss of generality, we can restrict ourselves to planes which go through $k-1$ points from $\{\mathbf{v}_i\}$. Further, it is well known that the position of vector $\mathbf{v} = (v_1, ..., v_k)'$ about the plane defined by k points $(\mathbf{0}, \mathbf{w}_1, ..., \mathbf{w}_{k-1})$ is determined by the sign of

$$\det \begin{bmatrix} v_1 & v_2 & \cdots & v_k \\ w_{11} & w_{12} & \cdots & w_{1k} \\ \cdots & & & \\ w_{k-1,1} & w_{k-1,2} & \cdots & w_{k-1,k} \end{bmatrix}.$$

Therefore, the MLE does not exist if and only if all D_i have the same sign for all $\mathbf{v}_i$ and a certain group of $k-1$ vectors from $\{\mathbf{v}_i\}$. □

The following algorithm determines whether the MLE exists.

Algorithm 2.

1. Pick any $k-1$ different vectors $\mathbf{w}_1, ..., \mathbf{w}_{k-1}$ from $\{\mathbf{v}_i, i = 1, ..., n\}$; there are $\binom{k-1}{n}$ ways to pick $k-1$ different vectors $\mathbf{w}_1, ..., \mathbf{w}_{k-1}$.
2. Compute (11) for $v_1, ..., v_n$. If all D_i have the same sign the MLE does not exist, and we quit. Otherwise go to step 1 and pick another group of $k-1$ vectors.

If for any $k-1$ group of vectors $\{D_i\}$ do not have the same sign, the MLE exists.

It is worthwhile to note that the above criteria for the MLE existence are applicable to general binary model $P(y_i = 1) = \mu(\beta'\mathbf{x}_i)$ where $\mu(\cdot)$ is a link-function such that: (i) $0 < \mu(\cdot) < 1$, (ii) $\log(\mu(\cdot))$ and $\log(1 - \mu(\cdot))$ are strictly concave functions. In particular, these criteria are valid for the logistic model.

4. Algorithm for the log-likelihood maximization

The most popular general iterative algorithms for the log-likelihood function maximization are Newton-Raphson (NR) and Fisher Scoring (FS) with iterations:

$$\mathbf{b}_{r+1} = \mathbf{b}_r + \lambda_r \mathbf{H}_r^{-1}\mathbf{g}_r, \ r = 0, 1, 2, ... \tag{12}$$

where r is the iteration index, $\mathbf{b}_r$ is the MLE approximation at the rth iteration, $\lambda_r > 0$ is the step length in the direction $\mathbf{H}_r^{-1}\mathbf{g}_r$, and $\mathbf{g}_r$ is the gradient of the log-likelihood function (4). For the Newton-Raphson algorithm $\mathbf{H} = -\partial^2 l/\partial\beta^2$ calculated by formula (5) and for the Fisher Scoring algorithm $\mathbf{H} = \mathbf{I}$ calculated by formula (6); all quantities are calculated at $\beta = \mathbf{b}_r$. The important feature of the probit model is that the matrices are positive definite under the full rank condition. As follows from the general theory of optimization at each iteration, where $\mathbf{g}_r \neq \mathbf{0}$, there exists such a positive step length λ_r that $l(\mathbf{b}_{r+1}) > l(\mathbf{b}_r)$. A common practice is to start with $\lambda = 1$ and reduce it by half until $l(\mathbf{b}_{r+1}) > l(\mathbf{b}_r)$.

4.1. The unit step algorithm

It is possible to avoid matrix $\mathbf{H}$ recalculation and its inverse at each step of iteration (12) considering the upper bound approximation of the Hessian. In fact, as follows from (8)

$$0 < \frac{s\phi(s)}{\Phi(s)} + \frac{\phi^2(s)}{\Phi^2(s)} < 1, \quad 0 < \frac{\phi^2(s)}{[1-\Phi(s)]^2} - \frac{s\phi(s)}{[1-\Phi(s)]} < 1$$

for any $s \in (-\infty, \infty)$. Therefore, as follows from (5)

$$\partial^2 l / \partial \beta^2 > -\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' = -\mathbf{X}'\mathbf{X}, \tag{13}$$

where $\mathbf{X}$ is the $n \times k$ matrix with $\mathbf{x}_i$ as the ith row (the matrix inequality means that the difference between the left and the right side of the inequality is a positive definite matrix). Based on inequality (13) the following Unit Step (US) algorithm for the log-likelihood maximization can be proposed:

$$\mathbf{b}_{r+1} = \mathbf{b}_r + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{g}_r, \ r = 0, 1, 2, ... \tag{14}$$

There are several advantages of the US algorithm: (i) it does not require the trial of step length, (ii) the matrix $(\mathbf{X}'\mathbf{X})^{-1}$ is constant and can be calculated only once at the initial iteration, (iii) it avoids the overflow error problem for large s when calculating (5) or (6), (iv) at each iteration it guarantees the increase of the log-likelihood function. The increase of the log-likelihood function follows from the Taylor series expansion and inequality (13):

$$\begin{aligned} l(\beta) - l_r &= \mathbf{g}_r'(\beta - \beta_r) + \frac{1}{2}(\beta - \beta_r)' \frac{\partial^2 l}{\partial \beta^2}(\beta_*)(\beta - \beta_r) \\ &> \mathbf{g}_r'(\beta - \beta_r) - \frac{1}{2}(\beta - \beta_r)' \mathbf{X}'\mathbf{X}(\beta - \beta_r). \end{aligned}$$

The next approximation vector $\mathbf{b}_{r+1}$ calculated by formula (14) maximizes the right side of the above inequality. Since for $\mathbf{g}_r \neq \mathbf{0}$ that maximum is positive it follows that $l_{r+1} > l_r$ for all $r = 0, 1, \ldots$ which proves (iv).

Theorem 5. *If the MLE $\widehat{\beta}_{ML}$ exists then the US algorithm converges to $\widehat{\beta}_{ML}$ starting from any initial parameter vector* $\mathbf{b}_0$.

Proof. Since the MLE exists $\lim_{\|\beta\| \to \infty} l(\beta) = -\infty$. Therefore, the sequence generated by the US algorithm (14) is bounded because $l(\mathbf{b}_r) \geq l(\mathbf{b}_0)$. Let $\mathbf{b}_*$ be any limit point of $\{\mathbf{b}_r\}$, there exists at least one limit point since $\mathbf{b}_r$ are bounded. Then $\lim_{p \to \infty} \mathbf{b}_{r_p} = \mathbf{b}_*$ and letting $p \to \infty$ in (14) we obtain $\mathbf{b}_* = \mathbf{b}_* + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{g}_*$ where $\mathbf{g}_*$ is the gradient at $\beta = \mathbf{b}_*$. It implies $\mathbf{g}_* = \mathbf{0}$ at any limit point of $\{\mathbf{b}_r\}$.

But there is only one point where the gradient of the log-likelihood vanishes, $\widehat{\beta}_{ML}$. Hence the sequence generated by the US algorithm $\{\mathbf{b}_r\}$ has a unique limit point which is $\widehat{\beta}_{ML}$. □

Especially effective the US algorithm might be at initial steps of the maximization process; after few iterations one can switch to algorithms of the second order such as Newton-Raphson (see the next section).

4.2. Initialization

To start a maximization process (14) an initial vector $\mathbf{b}_0$ is needed. For well defined problems, when the range of $\mid s_i \mid = \mid \beta'\mathbf{x}_i \mid$ is fairly small, the initial guess is not so important. However, to reduce the number of iterations and to avoid a possible computer overflow problem, in the presence of outliers, a better initial guess is required. Usually the probit model contains the intercept term. So, let us assume that the first column of matrix $\mathbf{X}$ contains only 1's and $\mathbf{b}_0 = (b_{00}, \mathbf{b}'_{01})'$ where b_{00} is the intercept term and $\mathbf{b}_{01}$ is the $(k-1)-$vector of coefficients at explanatory variables. Then a reasonable initial guess would be $\mathbf{b}_{01} = \mathbf{0}$ and $b_{00} = \Phi^{-1}(r/n)$ where r is the number of $y_i = 1$ and Φ^{-1} is the inverse cdf. This choice is referred to as 'go-through-origin' guess.

Another initial vector can be derived via linear regression y_i on $\mathbf{x}_i$, i.e. $\mathbf{b}_0 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

4.3. Approximation explosive effect

For some observation points, outliers, the value $s = \beta'\mathbf{x}$ may be quite large and consequently may create troubles during the log-likelihood maximization. In this case the computer program usually stops due to an overflow error. The core of this trouble is in the computation of $1 - \Phi(s)$ and related quantities. For example, let us consider the computation of functions

$$\frac{\phi(s)}{1 - \Phi(s)} \tag{15}$$

and

$$\frac{\phi^2(s)}{(1 - \Phi(s))^2} - \frac{s\phi(s)}{1 - \Phi(s)} \tag{16}$$

for large s, presented in the gradient (4) and Hessian matrix (5) formulae, respectively. As follows from the right limit (7) the quantity (15) must be close to s and the quantity (16) must be close to zero for $s \simeq \infty$. *Figure 2* illustrates the *approximation explosive effect* for that takes place in the neighborhood of $s = 7$ in a popular statistical package S-plus. In fact, this effect has little to do with the accuracy of $\Phi(s)$ computation, and even very accurate algorithms of $\Phi(s)$ approximation for large positive s, such as described in Kennedy and Gentle (1980) or Vedder (1993), cannot help. The approximation explosive effect happens because for large positive s the value $\Phi(s)$ is close to 1 and therefore the accuracy of $1-\Phi(s)$ cannot be better than 10^{-7} for single and 10^{-14} for double precision. Thus, inaccuracy of (15) or (16) is limited by computer float-point arithmetic (cut-off-error), not inaccuracy of

$\Phi(s)$ approximation. Probably, the easiest way to avoid these cut-off-errors is to substitute $1-\Phi(s)$ by $\Phi(-s)$. However, to avoid the cut-off-error problem in a more comprehensive way one should use another formulae for the log-likelihood function and its derivatives which are presented in the following subsection.

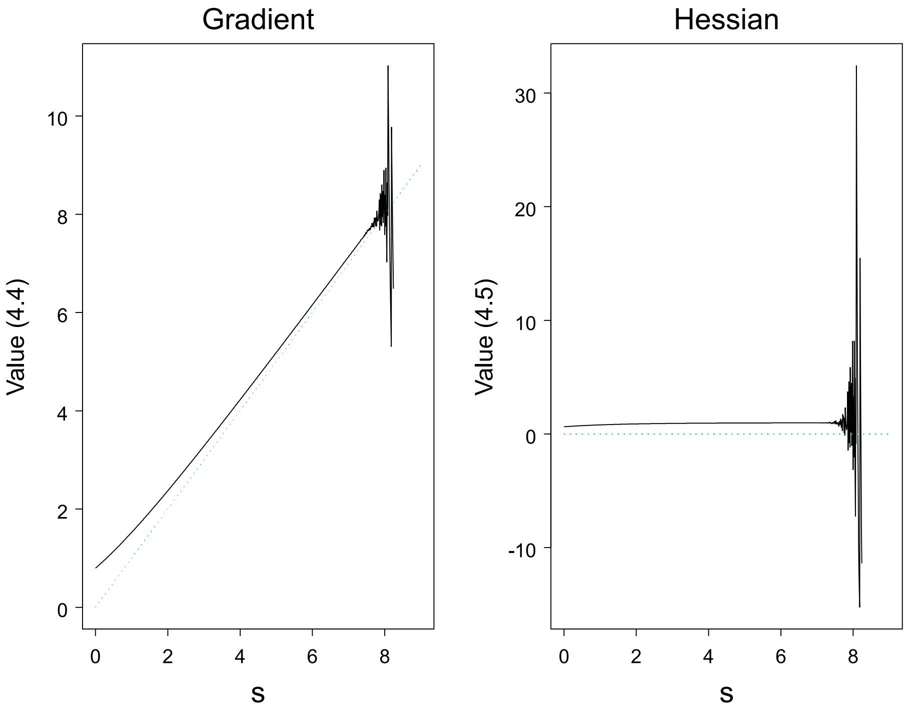

Figure 2. *Approximation explosive effect in statistical package S-plus. For s in the neigborhood of* 8 *the computation of the gradient and the Hessian of the log-likelihood function becomes unstable. This is driven by the fact that for large positive s the value* $\Phi(s)$ *is close to 1 and therefore* $1-\Phi(s)$ *cannot have accuracy less than* 10^{-7} *for single precision and* 10^{-14} *for double precision arithmetic. Dashed lines correspond to the approximation based on the limits (7): for large s (15) can be well approximated by s and (16) can be well approximated by zero*

4.4. Approximate formulae

As follows from the previous section a straightforward computation of the log-likelihood function and its derivatives by formulae (3), (4) and (5) becomes unreliable for large values $s = \beta'\mathbf{x}_i$. We suggest to use the Feller approximation (7) for large s.

Thus, a more reliable computation of the log-likelihood function is

$$l(\beta) = \sum_{y_i=1} \theta_1(\beta'\mathbf{x}_i) + \sum_{y_i=0} \theta_2(\beta'\mathbf{x}_i)$$

where

$$\theta_1(s)=\begin{cases}\log(-\phi(s)/s), & \text{if } s<-5\\ \log\Phi(s), & \text{if } \mid s\mid\leq 5\\ -\phi(s)/s, & \text{if } s>5,\end{cases}\qquad \theta_2(s)=\begin{cases}\phi(s)/s, & \text{if } s<-5\\ \log(1-\Phi(s)), & \text{if } \mid s\mid\leq 5\\ \log(\phi(s)/s), & \text{if } s>5.\end{cases}\qquad(17)$$

The gradient should be computed by the formula

$$\partial l/\partial\beta=\sum_{y_i=1}\theta_3(\beta'\mathbf{x}_i)\mathbf{x}_i-\sum_{y_i=0}\theta_4(\beta'\mathbf{x}_i)\mathbf{x}_i$$

where

$$\theta_3(s)=\begin{cases}-s, & \text{if } s<-5\\ \phi(s)\Phi^{-1}(s), & \text{if } \mid s\mid\leq 5\\ \phi(s), & \text{if } s>5,\end{cases}\qquad \theta_4(s)=\begin{cases}\phi(s), & \text{if } s<-5\\ \phi(s)\left(1-\Phi(s)\right)^{-1}, & \text{if } \mid s\mid\leq 5\\ s, & \text{if } s>5.\end{cases}\qquad(18)$$

For the Hessian matrix we recommend to use the following formulae

$$\partial^2 l/\partial\beta^2=-\{\sum_{y_i=1}\theta_5(\beta'\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i'+\sum_{y_i=0}\theta_6(\beta'\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i'\}$$

where

$$\theta_5(s)\quad=\quad\begin{cases}0, & \text{if } s<-5\\ s\phi(s)\Phi^{-1}(s)+\phi^2(s)\Phi^{-2}(s), & \text{if } \mid s\mid\leq 5\\ s\phi(s)+\phi^2(s), & \text{if } s>5,\end{cases}\qquad(19)$$

$$\theta_6(s)\quad=\quad\begin{cases}\phi^2(s)-s\phi(s), & \text{if } s<-5\\ \phi^2(s)(1-\Phi(s))^{-2}-s\phi(s)(1-\Phi(s_i))^{-1}, & \text{if } \mid s\mid\leq 5\\ 0, & \text{if } s>5.\end{cases}\qquad(20)$$

For the expected information matrix the following formula should be used

$$\sum_1^n\theta_7(\beta'\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i'$$

where

$$\theta_7(s)=\begin{cases}-s\phi(s), & \text{if } s<-5\\ \phi^2(s)\Phi^{-1}(s)(1-\Phi(s))^{-1}, & \text{if } \mid s\mid\leq 5\\ s\phi(s), & \text{if } s>5.\end{cases}\qquad(21)$$

The effectiveness of these approximative formulae is demonstrated in the following section where several algorithms are compared via statistical simulation. These formulae provide a good approximation and are very reliable for large s; the threshold is chosen 5. This value was found empirically and is subject to change.

5. Algorithms comparison

To evaluate the speed and reliability of different algorithms of the log-likelihood function maximization five algorithms in the form (6) were compared:

1. FS - Fisher Scoring. The log-likelihood function is calculated by formula (3), the gradient is calculated by formula (4) and the expected information matrix is calculated by formula (6).
2. NR - Newton-Raphson, formulae (3), (4) and (5) were used.
3. FSA - Fisher Scoring with Approximations for large s: formulae (17), (18) and (21) were used.
4. NRA - Newton-Raphson with Approximations for large s: formulae (17), (18) and (19), (20).
5. NRUS - the same as NRA with the first iteration calculated by the Unit Step algorithm.

The sample size was taken $n = 500$ and the number of parameters, $k = 3$. Three types of data were generated. In the first case 'Small' the range of the linear predictor $s_i = \beta'\mathbf{x}_i$ was $(-2, 2)$. In the second type of experiments a 'Moderate' range $(-4, 4)$ was taken. At last in 'Large' the range of $\beta'\mathbf{x}_i$ corresponds $(-6, 6)$. All algorithms started from the 'go-through-origin' guess. The results are reported in *Table 1* where 'Time' is the average time in seconds to converge, 'Nonc.' is the percentage of cases with failed convergence (the number of iteration exceeded 100, convergence met if five digits in all parameters coincide in two subsequent iterations), 'Error' is the percentage of cases when the program was stopped due to overflow error during calculations.

As follows from *Table 1* for experiments with relatively small s_i all algorithms do the job well and are similar in terms of speed and reliability. Only the FS algorithm failed to converge in one experiment. For moderate values s_i standard algorithms FS and NR failed almost in one third of all experiments. They got worse for a wider range of s_i: computer program crashed in 83 experiments out of 100 due to overflow error. On contrary, the corrected algorithms FSA, NRA and NRUS based on approximate formulae worked very reliable and fast.

	Small			Moderate			Large		
	$(-2, 2)$			$(-4, 4)$			$(-6, 6)$		
Algorithm	Time	Nonc.	Error	Time	Nonc.	Error	Time	Nonc.	Error
FS	3.85	1%	0	4.65	0	29%	6.47	0	83%
NR	3.05	0	0	3.48	0	29%	3.86	0	83%
FSA	3.97	0	0	5.15	0	0	5.63	0	0
NRA	3.24	0	0	3.72	0	0	4.34	0	0
NRUS	3.94	0	0	4.41	0	0	4.95	0	0

Table 1. *Algorithms comparison for a different linear prediction range,* $n = 500, k = 3$, *100 simulation experiments*

References

[1] A. Albert, J. A. Anderson, *On the existence of maximum likelihood estimates in logistic regression*, Biometrika **71**(1984), 1-10.

[2] L. Baker, *More C Tolls for Scientists and Engineers*, McGraw-Hill, New York, 1991.

[3] R. W. Cottle, J-S. Pang, R. E. Stone, *The Linear Complementary Problem*, Academic Press, New York, 1992.

[4] D. R. Cox, *The Analysis of Binary Data,* Chapman and Hall, London, 1970.

[5] E. Demidenko, *Linear and Nonlinear Regression* (in Russian), Nauka, Moscow, 1981.

[6] E. Demidenko, *On the existence of the least squares estimate in nonlinear growth curve models of exponential type*, Communications in Statistics, Theory and Methods **25**(1996), 159-182.

[7] E. Demidenko, D. Spiegelman, *A Paradox: More measurement error can lead to more efficient estimates*, Communications in Statistics, Theory and Methods **26**(1997), 1649-1675.

[8] E. Demidenko, *Is this the least squares estimate?* Biometrika **87**(2000), 437-452.

[9] W. Feller, *An Introduction to Probability Theory and Its Application*, Wiley, New York, 1957.

[10] D. J. Finney, *Probit Analysis*, Cambridge University Press, Cambridge, 1971.

[11] P. Gordan, *Über die Auflösung linearer Gleichungen mit reelen Coefficienten*, Mathematische Annalen **6**(1873), 23-8.

[12] S. J. Haberman, *The Analysis of Frequency Data*, University of Chicago, Chicago, 1974.

[13] K. Hoffman, *Linear Algebra*, 3rd Ed., Prentice Hall, London, 1999.

[14] W. J. Kennedy, J. E. Gentle, *Statistical Computing*, Marcel Dekker, New York, 1980.

[15] E. Lesaffre, H. Kaufmann, *Existence and uniqueness of the maximum likelihood estimator for a multivariate probit model*, Journal of American Statistical Association **87**(1992), 805-11.

[16] P. McCullagh, J. A. Nelder, *Generalized Linear Models*, Chapman and Hall, London, 1989.

[17] T. Nakamura, C.-S. Lee, *On the existence of minimum contrast estimates in binary response model*, Annals of Institute of Statistical Mathematics **45**(1993), 741-58.

[18] W. H. Press, S. A. Teulolsky, W. T. Vetterling, B. P. Flannery, *Numerical Recipes in C*, Cambridge University Press, Cambridge, 1992.

[19] M. J. Silvapulle, *On the existence of maximum likelihood estimates for the binomial response models*, J. of Royal Statistical Society, ser. B **43**(1981), 310-313.

[20] J. D. Vedder, *An invertible approximation to the normal distribution function*, Computational Statistics & Data Analysis **16** (1993), 119-23.

[21] R. W. M. Weddenburn, *On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models*, Biometrika **63** (1976), 27-32.