International Journal of Mathematics and Statistics Studies

Vol.4, No.4, pp.21-31, August 2016

Published by European Centre for Research Training and Development UK (www.eajournals.org)

APPLICATION OF NEWTON RAPHSON METHOD TO NON – LINEAR MODELS

Bakari H.R, Adegoke T.M, and Yahya A.M

Department of Mathematics and Statistics, University of Maiduguri, Nigeria

ABSTRACT: *Maximum likelihood estimation is a popular parameter estimation procedure however parameters may not be estimable in closed form. In this work, the maximum likelihood estimates from different distributions were obtained after the failure of the likelihood approach. The targeted models are Non Linear models with an application to a Logistic regression model. Although, obtaining the estimate of parameters for non linear models cannot be easily obtained directly. That is the solution is intractable. So there is a need to look else where, so as to obtain the solutions . In this work, R statistical package was used in performing the analysis. The result shows that convergence was attained at the 18th iteration out of 21. This also provides the values and the maximum estimate for β_0 and β_1.*

KEYWORDS*:* Intractable Functions, Maximum Likelihood, Likelihood Function

INTRODUCTION

The problem of estimation is to devise a means of using sample observations to construct good estimates of one or more of the parameters. It is expected that the ”information” in the sample concerning those parameters will make an estimate based on the sample generally better than a sheer guess. How well the parameter is approximated can depend on the method, the type of data and other factors. The method of maximum likelihood corresponds to many well-known estimation methods in statistics (such as; maximum likelihood, moments, least squares, bayesian estimation etc) and finding particular parametric values that make the observed results the most probable (given the model). But in this study we are concentrating on maximum likelihood. In statistics, maximum-likelihood estimation (MLE) is a method of estimating the parameters of a statistical model. When applied to a data set and given a statistical model, maximum-likelihood estimation provides estimates for the model’s parameters. But with more complicated models, maximum likelihood alone may not result in a closed form solution. Analytic expressions for maximum likelihood estimators in complex models are usually not easily available, and numerical methods are needed. Newton’s method can be used to find solutions when no closed form exists and it can converge quickly, especially if the initial value of the iteration is close to the actual solution. Here the importance of an efficient estimator is reinforced since the platykurtic nature of an inefficient estimator diminishes the ability of the algorithm to converge. However, with the rapid increase of computer speed, maximum likelihood estimation has become easier and has increased in popularity. In this paper, interest is mainly focused on the estimation of parameters of some distributions which does not have a closed form solution.

LITERATURE REVIEW

Maximum-likelihood estimation was recommended, analyzed (with flawed attempts at proofs) and vastly popularized by R. A. Fisher between 1912 and 1922 Pfanzagl (1994)[**?**]. Although it had been used earlier by Gauss, Laplace, Thiele, and F. Y. Edgeworth (September 1908 , December 1908). Reviews of the development of maximum likelihood have been provided by a number of authors (Savage (1976) [**?**], Pratt (1976) [**?**], Stigler (1978 [**?**], 1986 [**?**], 1999 [**?**]) and

Aldrich (1997) [**?**]). Much of the theory of maximum-likelihood estimation was first developed for Bayesian statistics, and then simplified by later authors Pfanzagl (1994)[**?**]. Efron (1982) [**?**] explained the method of maximum likelihood estimation along with the properties of the estimator. According to Aldrich (1997) [**?**], the making of maximum likelihood was one of the most important developments in 20th century statistics. The method of moment (MM) is also a commonly used method of estimation. In this method, the sample moments are assumed to be estimates of population moments and thus moment estimates for the unknown values of population parameters are found (Lehman and Casella, 1998 [**?**]). Negative integer moments are useful in applications in several contexts, notably in life testing problems. Bock et al. (1984)[**?**] illustrated the examples of their use in the evaluation of a variety of estimators. With the particular reference to Chi-square distribution, in the inverse regression problem, Oman (1985)[**?**] gave an exact formula for the mean squared error of Kruutchkoffs inverse estimator by use of negative integer moments of the noncentral Chi-squared variable.

Objectives

The study is aimed at applying Newton Raphson method to non-linear models with a view to obtain of obtaining a maximum likelihood estimates for logistic regression models.

METHODOLOGY

Maximum Likelihood

The likelihood function of the random samples s the product of their respective probability

distributions

$$L(\phi; X_1, ..., X_2) = \prod_{1}^{n} f(x, \phi) \qquad (1)$$

To maximize the natural logarithm of the likelihood with respect to φ and equating to zero gives the score function as

$$\frac{\delta[lnL(\phi)]}{\delta\phi} = 0 \quad (2)$$

if (2) cannot be solve analytically then we need to adopt an iterative method to estimate the parameters of the distribution.

$$\frac{\delta^2[lnL(\phi)]}{\delta\phi^2} = 0 \qquad (3)$$

Newton Raphson Method

NEWTON'S RULE will be adopted. The optimum of the approximation (which is easy to calculate) gives a guess of the optimum for the actual function. If this guess is not adequately close to the optimum, a new approximation is computed and the process repeated. The Newton Rapson Method can be stated as

$$g(x) = X - \frac{f'(X)}{f''(X)} \quad (4)$$

The one-parameter situation

The derivative of the log-likelihood,

$$s(\varphi;X) = lik^{0}(\varphi,X) \quad (5)$$

is usually named the score function. Note that the score function is a random variable since it depends on the random observations x_i. It can be expressed by the likelihood function itself through

$$s(\phi;X)\frac{1}{lik(\phi;X))}lik'(\phi;X) \quad (6)$$

which is obtained from equation (6) by ordinary rules on differentiation. If $l(\varphi;X)$ is a smooth function in φ, the likelihood should have a derivative equal to zero at the max-point. A common approach in order to find maximum points is therefore to solve the scoring equation

$$s(\hat{\varphi};X) = 0 \quad (7)$$

In order to evaluate if the solution is actually a maximum point, the second derivative must be inspected. As is apparent from equation (6), $s(\varphi;X)$ is a stochastic quantity. An important property of the scoring function is that if X has probability density $f(X;\varphi)$, then

$$E[s(\varphi;X)] = 0$$

. A solution of the scoring equation can therefore be seen as a value of *hatφ* such that the scoring function is equal to its expectation. The variability of $S(\varphi;X)$ reflects the uncertainty involved in estimating φ. The variance of S is called the Fisher information (sometimes it is also called the expected information). The Fisher information $I(\varphi)$ of an experiment can be written as

$$I(\varphi) = V\,ar[s(\varphi;X)] = E[l(\varphi)] \quad (8)$$

The multi-parameter situation

Assume $\Phi = (\varphi_1,...,\varphi_p)^T$ is a vector of p, say, unknown parameters. The derivative of the log-likelihood, still named the score function, now is a vector:

$$s(\phi;X) = lik(\phi;X) = \frac{1}{lik(\phi;X)}lik(\phi;X) \quad (9)$$

The i^{th} element of the vector $S(\varphi;X)$,$S_i(\varphi;X)$, is the partial derivative of $lik(\varphi;X)$ with respect to φ_i. A more mathematically correct way of expressing $S(\varphi;X)$ would be $\frac{\delta}{\delta\phi}lik(\phi,X)$,but we will use the simpler form $lik(\varphi;X)$. As for the one-parameter case, each $s_i(\varphi;X)$ has expectation zero. Finding the MLE by solving the scoring equations

$$s(\hat{\varphi};X) = 0 \qquad (10)$$

now result in a set of p equations with p unknowns. The expected information is now generalized to be a matrix $I(\Phi)$ with the (i, j)th entry given by

$$I_{ij}(\Phi) = Cov[s_i(\Phi;X), s_j(\Phi;X)] = E[\frac{\delta^2}{\delta\phi_i\delta\phi_j}l(\Phi)] \; (11)$$

Here the second equality can be proven by a simple generalization of the argument in the one-parameter case. In the multi-parameter situation we usually name $I(\Phi)$ by the expected information matrix or the Fisher information matrix. An important property of $I(\Phi)$ is that it is always positive semi-definite. Where matrix I is positive semi-definite if $aIa \geq 0$ for all vectors a. Note that $I(\Phi)$ depends on the unknown quantity Φ. Common practice is to insert the estimated value $\Phi\hat{}$ for Φ giving an estimate $I(\Phi)\hat{}$ of $I(\Phi)$. A further complication is that the expectations in (12) are not always possible to compute. Then an alternative is to use the observed information matrix $J(\Phi)$ with (i, j)th entry given by

$$J_{ij}(\Phi) = \frac{\delta^2}{\delta\phi_i\delta\phi_j}l(\Phi) \quad (12)$$

As for the expected information matrix, an estimate $\Phi\hat{}$ needs to be inserted for Φ in order to evaluate $J(\Phi)$. The i^{th} diagonal element of $J^1(\Phi)$ can then be used as an approximation for the variance of $\hat{\varphi}_i$ instead of the i^th diagonal element of $I^1(\Phi)$. Both these approximations will equally be valid in the sense that as the number of observations increases, the approximation error will decrease to zero. If possible to calculate, the expected information is preferable, since the observed information in some cases can be unstable. Note that in many standard models used in statistics, $I(\Phi) = J(\Phi)$.

Algorithm of Newton Raphson Method

Consider the newton raphson iteration given as

$$X_{n+1} = X_n + \frac{f_n}{f'_n} \qquad (13)$$

Repacing X with φ , X_{n+1} with φ_1, $f(x)$ with $s(\varphi;X)$ and $f^0(x)$ with $J(\varphi)$ in equation (14) we will obtain the algorithm for Newton Raphson Method

$$\phi^{(n+1)} = \phi^{(n)} + \frac{s(\phi^{(n)})}{J(\phi^{(n)})} \qquad (14)$$

Non-Linear Regression

The general equation of a non-linear regression model can be expressed as

$$Y_i = f(x_i\beta \;) + \epsilon_i, \qquad i = 1, ...n \qquad (15)$$

where x_i is a vector of explanatory variables, β is a p-dimensional vector of unknown regression parameters and $_i$ is a noise term. We will make the standard assumptions about these noise terms:

- $E[\epsilon_i] = 0$
- $Var[\epsilon_i] = \sigma^2$
- $\epsilon_1, \ldots, \epsilon_n$ are uncorrelated
- $\epsilon_i' s$ are normally distributed

Multiple linear regression is the special case where

$f(xi,\beta) = \beta + \beta 1xi,1 + ... + \beta p-1xi,p-1$(16)

We will however in this work allow for nonlinear g functions. Assume that $\{(y_i,x_i), i = 1,2,...,n\}$ are observed (y_i is the observed value of Y_i). Under the assumptions above, the likelihood function is given by

$$\begin{aligned} L(\beta, \sigma^2) &= \prod_{i=1}^{n} g(y_i; x_i\beta, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi\sigma^2)}} e^{-\frac{1}{2\sigma^2}(y - f(x_i\beta))} \qquad (17) \end{aligned}$$

while the log-likelihood is

$$\begin{aligned} l(\beta, \sigma^2) &= \sum_{i=1}^{n} \left[-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(y_i - f(x_i\beta))^2 \right] \\ &= -\frac{n}{2} - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - f(x_i\beta))^2 \qquad (18) \end{aligned}$$

not possible to obtain, and numerical methods have to be applied. For notational simplicity, dene

$$f\prime(x_i\beta) = \frac{\partial}{\partial\beta_j} f(x_i\beta) \qquad (19)$$

and

$$f\prime\prime(x_i\beta) = \frac{\partial^2}{\partial\beta_j \partial\beta_k} f(x_i\beta) \qquad (20)$$

The partial derivatives of $l(\beta,\sigma^2)$ with respect to β and σ^2 are then given by the score function $s(\beta,\sigma^2)$ with element

$$s_j(\beta, \sigma^2) = \frac{\partial}{\partial\beta_k} l(\beta, \sigma^2) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - f(x_i\beta))f\prime_j(x_i\beta)$$

$$s_{p+1}(\beta, \sigma^2) = \frac{\partial}{\partial \sigma^2} l(\beta, \sigma^2)$$

$$= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (y_i - f(x_i,\beta))^2 \quad (21)$$

and the observed information matrix $j(\beta,\sigma^2)$ with elements

$$J_{j,k}(\beta, \sigma^2) = -\frac{\partial^2}{\partial \beta_j \partial \beta_k} l(\beta, \sigma^2)$$

$$= \frac{1}{\sigma^2} \sum_{i=1}^{n} [f\prime_j(x_i,\beta) f\prime_k(x_i,\beta) - (y_i - f(x_i,\beta)) f\prime\prime_{j,k}(x_i,\beta)] \quad (22)$$

$$J_{j,p+1}(\beta, \sigma^2) = -\frac{\partial^2}{\partial \beta_j \partial \sigma^2} l(\beta, \sigma^2)$$

$$= \frac{1}{\sigma^4} \sum_{i=1}^{n} (y_i - f(x_i,\beta)) f\prime_k(x_i,\beta) \quad (23)$$

$$J_{p+1,p+1}(\beta, \sigma^2) = -\frac{\partial^2}{\partial \sigma^2 \partial \sigma^2} l(\beta, \sigma^2)$$

$$= -\frac{n}{\sigma^4} + \frac{1}{\sigma^6} \sum_{i=1}^{n} (y_i - f(x_i,\beta))^2 \quad (24)$$

where k, l = 1, ..., p. These quantities can be directly imputed into the general Newton Raphson algorithm **??**. A more efficient algorithm can be obtained by utilizing that for given β, an analytical expression for the maximum likelihood estimate $\hat{\sigma}^2$ for σ^2 can be obtained. From **??**

$$\frac{\partial}{\partial \sigma^2} l(\beta, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (y_i - f(x_i,\beta)) \quad (25)$$

and the solution $\hat{\sigma}^2$ to the equation $\frac{\partial}{\partial \sigma^2} l(\beta, \sigma^2 = 0$ is given by

$$\hat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i,\beta))^2 \quad (26)$$

Logistic Regression

Suppose that $(y_i|x_i) i = 1,...,n$ represent a random sample from the Binomial distribution.

Then,

$$y_i \sim binorm(1, p(x_i,\beta)) \quad (27)$$

$$p(x_i,\beta) = \frac{\exp\{\beta_0 + \beta_1 x_i\}}{1 + \exp\{\beta_0 + \beta_1 x_i\}}$$

(28)

By making the usual assumption that all observations are independent, the likelihood function becomes

$$L(\beta) = \prod_{i=1}^{n} p(x_i,\beta)^{y_i}(1 - p(x_i,\beta))^{1-y_i} \tag{29}$$

The log-likelihood can be express as

$$l(\beta) = \sum_{i=1}^{n} [y_i \log(p(x_i,\beta)) + (1 - y_i) \log(1 - p(x_i,\beta))] \tag{30}$$

Since,

$$p(x_i,\beta) = \frac{\exp\{\beta_0 + \beta_1 x_i\}}{1 + \exp\{\beta_0 + \beta_1 x_i\}} \tag{31}$$

then

$$\begin{aligned} 1 - p(x_i,\beta) &= 1 - \frac{\exp\{\beta_0 + \beta_1 x_i\}}{1 + \exp\{\beta_0 + \beta_1 x_i\}} \\ &= \frac{1}{1 + \exp\{\beta_0 + \beta_1 x_i\}} \end{aligned} \tag{32}$$

Also,

$$\begin{aligned} \frac{p}{1-p} &= \exp \beta_0 + \beta_1 x_i \\ \log\left[\frac{p}{1-p}\right] &= \beta_0 + \beta_1 x_i \end{aligned} \tag{33}$$

The log-likelihood can now be expressed as

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{n} \left[y \log \frac{p}{1-p} + \log(1-p) \right] \\ &= \sum_{i=1}^{n} [y \log(\exp \beta_0 + \beta_1 x_i) + \log(1 + \exp \beta_0 + \beta_1 x_i)] \end{aligned} \tag{34}$$

and then calculate the gradient and the Hessian of $l(\beta)$ with respect to β directly using chain rule $\frac{\partial l}{\partial \beta_j} = \frac{\partial l}{\partial p(x_i,\beta)} \times \frac{\partial p(x_i,\beta)}{\partial \beta_j}$ j = 0; 1 to calculate the partial from the original model as follows:

From equation 33

$$\begin{aligned} \log p(x_i,\beta) - \log(1 - p(x_i,\beta)) &= \beta_0 + \beta_1 x_i \\ \frac{\partial p(x_i,\beta)}{\partial \beta_0} &= p(x_i,\beta)(1 - p(x_i,\beta)) \\ \frac{\partial p(x_i,\beta)}{\partial \beta_1} &= p(x_i,\beta)(1 - p(x_i,\beta)) \end{aligned} \tag{35}$$

So, after substitution, it follows that:

$$\frac{\partial l}{\partial\beta_0} = \sum_{i=1}^{n}\left[\frac{y_i}{p(x_i\beta)} - \frac{(1-y_i)}{1-p(x_i\beta)}\right]\frac{\partial p(x_i\beta)}{\partial\beta_0} \qquad (36)$$

$$= \sum_{i=1}^{n}\left[\frac{y_i}{p(x_i\beta)} - \frac{(1-y_i)}{1-p(x_i\beta)}\right]p(x_i\beta)(1-p(x_i\beta))$$

$$= \sum_{i=1}^{n}\frac{y_i}{p(x_i\beta)} - \frac{(1-y_i)}{1-p(x_i\beta)} = \sum_{i=1}^{n}(y_i - p(x_i\beta))$$

$$\frac{\partial l}{\partial\beta_0} = \sum_{i=1}^{n}\left[\frac{y_i}{p(x_i\beta)} - \frac{(1-y_i)}{1-p(x_i\beta)}\right]\frac{\partial p(x_i\beta)}{\partial\beta_1}$$

$$= \sum_{i=1}^{n}\left[\frac{y_i}{p(x_i\beta)} - \frac{(1-y_i)}{1-p(x_i\beta)}\right]p(x_i\beta)(1-p(x_i\beta_1))x_i$$

$$= \sum_{i=1}^{n}x_i(y_i - p(x_i\beta))$$

(37)

(38)

(39)

(40)

(41)

$$\frac{\partial^2 l}{\partial\beta_0^2} = -\sum_{i=1}^{n}p(x_i\beta)(1-p(x_i\beta))$$

$$\frac{\partial^2 l}{\partial\beta_1^2} = -\sum x_i p(x_i\beta)(1-p(x_i\beta))x_i = -\sum p(x_i\beta)(1-p(x_i\beta))x_i^2$$

$$\frac{\partial^2 l}{\partial\beta_0\partial\beta_1} = -\sum_{i=1}^{n}p(x_i\beta)(1-p(x_i\beta))x_i$$

(42)

(43)

(44)

(45)

In the following implementations of Newton-Raphson, a negative sign is inserted in front of the log likelihood, the gradient, and the hessian, as these routines are constructed for minimizing nonlinear functions.

To illustrate logistic regression, we will analyze the data given in Table **??** below . The table contain a data set is given where the response is whether a beetle given a dose of poison has died or not, i.e., a binary response. The explanatory variable is the amount of poison. The data are grouped since many beetles are given the same dose.

Table 1: The mortality of beetles against dose of poison

Dose	Number of Insect	Number Killed
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60

Implementation of Numerical Method to Logistic Regression Model

Table 2 below gives the the summary of the iteration result from the analysis performed:

```
$minimum

[1] 18.71513

$estimate

[1] -60.71786 34.27055

$gradient
```

Table 2: Implementation of Newton Raphson Method with real life data

Interation	β_0	β_1	Gradient β_0	Gradient β_1	Function
0	2	1	179.3920	311.6497	553.8446
1	0.8662236	-0.9696594	-149.2270	-278.27354	264.512
2	1.3813254	-0.04068787	87.65981	147.04080	196.5479
3	1.1854755	-0.3589378	13.07235	13.09508	159.6459
4	1.1475771	-0.3880613	2.875724	-5.212983	159.2272
5	1.132853	-0.377521	3.354139	-4.349829	159.1309
6	-0.3256731	0.5412218	24.48201	33.95548	152.4527
7	-3.281333	2.289232	43.05080	67.96144	141.7666
8	-11.99262	7.29436	64.65333	108.43084	114.4357
9	-28.44790	16.58482	66.02012	113.63712	71.60655
10	-48.36249	27.69799	44.05701	77.1071	36.39559

11	-59.49891	33.83316	25.61607	45.36968	24.41009
12	-64.54647	36.54798	12.15607	21.83472	20.35624
13	-65.01072	36.72506	3.952493	7.294694	19.19656
14	-63.32287	35.73540	-0.0009565687	0.1641202589	18.83827
15	-61.42613	34.66077	-0.8270989	-1.4254833	18.72998
16	-60.74230	3.171	-0.2219827	-0.3933852	18.71557
17	-60.071244	34.26740	-0.009238042	-0.016760403	18.71514
18	-60.71714	34.27013	-4.215717e-05	8.400441e-05	18.71513
19	-60.71727	34.27021	7.017393e-07	1.206174e-06	18.71513
20	-60.71786	34.27055	-7.932261e-06	-1.403221e-05	18.71513
21	-60.71786	34.27055	-4.71224e-09	-7.985442e-09	18.71513

```
[1] -4.717224e-09 -7.985442e-09

$hessian

            [,1]       [ , 2]

[1,] 58.48343 103.9776

[2,] 103.97757 184.9640

$code

[1] 1

$iterations

[1] 21
```

RESULTS AND CONCLUSION

In this work, we consider the use of Newton Raphson method to obtain the estimate of parameters for a logistic regression model

From the result obtained from applying newton raphson method to obtaining a maximum likelihood estimates for loggistic regression model. A total of 21 iterations were performed to obtain the maximum likelihood estimate. Convergence was reached at the 18th returning 18.71513 as the value of the log-likelihood and the value of the estimate which maximizes the function is -60.71786 with gradient -4.717224×10^{-9} for β_0 and 34.2705 as the value of the estimate which maximizes the function with gradient -7.985442×10^{-9} for β_1 . This means for every unit change in the dosage of poison, the log odd for a insect to die increase by 34.27 . The hessian matrix which is the value of the second derivativesand is also known as variance-covariance matrix is

$$\begin{bmatrix} 58.8343 & 103.9776 \\ 103.97757 & 184.9640 \end{bmatrix} \qquad (46)$$

Since the hessian matrix is positive and its determinant is also greather than zero, then we can conclude that the estimates obtained are local minimum.

Acknowledgements

The authors are grateful to the Editors, Associate Editor, and referees whose comments and suggestions were valuable to improve the exposition of the paper.