DIGITALCOMMONS
— @WAYNESTATE —

Journal of Modern Applied Statistical Methods

Volume 13 | Issue 1

Article 29

5-1-2014

Robust Regression Analysis for Non-Normal Situations under Symmetric Distributions Arising In Medical Research

S S. Ganguly
Sultan Qaboos University, Muscat, Oman, ganguly@squ.edu.om

Recommended Citation
Ganguly, S S. (2014) "Robust Regression Analysis for Non-Normal Situations under Symmetric Distributions Arising In Medical Research," *Journal of Modern Applied Statistical Methods*: Vol. 13 : Iss. 1 , Article 29.
DOI: 10.22237/jmasm/1398918480
Available at: http://digitalcommons.wayne.edu/jmasm/vol13/iss1/29

Journal of Modern Applied Statistical Methods
May 2014, Vol. 13 No. 1, 446-462.

ISSN 1538 − 9472

Robust Regression Analysis for Non-Normal Situations under Symmetric Distributions Arising In Medical Research

S. S. Ganguly
Sultan Qaboos University
Muscat, Oman

In medical research, while carrying out regression analysis, it is usually assumed that the independent (covariates) and dependent (response) variables follow a multivariate normal distribution. In some situations, the covariates may not have normal distribution and instead may have some symmetric distribution. In such a situation, the estimation of the regression parameters using Tiku's Modified Maximum Likelihood (MML) method may be more appropriate. The method of estimating the parameters is discussed and the applications of the method are illustrated using real sets of data from the field of public health.

Keywords: Maximum likelihood, modified maximum likelihood, student's t-distribution, order statistics, delta method

Introduction

Often in medicine, a relationship is established between a response variable *y*, which depends on the r covariates x_1, x_2, …, x_r, which are independent of each other, so that, in total, there may be (r + 1) variables. In classical regression model, the response variable *y* is treated as a random variable whose mean depends upon fixed variables of the x_i's. The mean is assumed to be linear function of the regression coefficients α, β_1, β_2, … , β_r.

The linear regression model also arises in a different setting. Suppose all the variables y, x_1, x_2, …, x_r are random and have a joint distribution

$$f(y, x_1, x_2, ..., x_r),$$

which is not necessarily normal so that

S. S. Ganguly is a Professor in the Department of Family Medicine and Public Health. Email at: ganguly@squ.edu.om.

$$f(y,x_1,x_2,\ldots,x_r)=g(y\,|x_1,x_2,\ldots,x_r)\prod_{i=1}^{r}h(x_i). \tag{1}$$

It is assumed herein that the conditional distribution of y given , x_1, x_2, …, x_r is normal and is given by

$$g(y|x_1,x_2,\ldots,x_r)\approx\left\{\sigma_o^2(1-\sum_{i=1}^{r}\rho_{oi}^2)\right\}^{-\frac{1}{2}} \times\exp\left[-\frac{1}{2\sigma_o^2(1-\sum_{i=1}^{r}\rho_{oi}^2)}\left\{y-\mu_o-\sum_{i=1}^{r}\rho_{oi}\left(\frac{\sigma_o}{\sigma_i}\right)(x_i-\mu_i)\right\}^2\right] \tag{2}$$

with mean

$$E(y|\underline{x})=\mu_o-\sum_{i=1}^{r}\rho_{oi}\left(\frac{\sigma_o}{\sigma_i}\right)(x_i-\mu_i) \tag{3}$$

and variance

$$V\left(y|\underline{x}\right)=\sigma_o^2\left(1-\sum_{i=1}^{r}\rho_{oi}^2\right). \tag{4}$$

The marginal density corresponding to the covariate x_i is assumed to be symmetric about mean of the form:

$$\frac{1}{\sigma_i}f\left(\frac{x_i-\mu_i}{\sigma_i}\right) \tag{5}$$

Here $\mu_i=E(x_i),\sigma_i^2=V(x_i)\ and\ \rho_{oi}(i=1,2,\ldots,r)$ is the correlation coefficient between y and x_i. Relation (2) provides for the measurement of dependency of the response random variable on the random covariates x_i (i=1,2,…,r).

The linear relationship may also be written in the form of classical regression model as

$$E\left(y|\underline{x}\right)=\alpha+\sum_{i=1}^{r}\beta_i x_i \tag{6}$$

where

$$\alpha = \mu_o - \sum_{i=1}^{r} \rho_{oi} \left(\frac{\sigma_o}{\sigma_i} \right) \mu_i \tag{7}$$

and

$$\beta_i = \rho_{oi} \left(\frac{\sigma_o}{\sigma_i} \right), \quad i = 1, 2, \ldots, r \tag{8}$$

are the regression coefficients. It may be noted that $E\left(y|\underline{x}\right)$ is the best linear predictor of the response variable y where the population is $N_{r+1}(\underline{\mu}, \Lambda)$.

In medical epidemiology, one often encounters situations where some (if not all) covariates x_i have non-normal symmetric distributions. This article is restricted to a situation where the covariates have non-normal symmetric distributions. The objective, therefore, is to estimate the parameters $\left(\alpha, \underline{\beta}\right)^T$ from n sample values $\left(y_i, \underline{x}_i\right)$, 1 ≤ i ≤ n. For this, consider the family of student's t-distributions. The method, which has been developed here, is, of course, general and can be used for other families of location-scale distributions of the type (5).

Likelihood equations

Suppose that the covariate x_i (i=1,2,…,r) has the symmetric distribution with the density given by

$$h\left(x_i\right) \approx \left(k_i \sigma_i^2\right)^{-\frac{1}{2}} \left\{1 + \frac{(x_i - \mu_i)^2}{k_i \sigma_i^2}\right\}^{-p_i} \quad , -\infty \langle x_i \langle \infty \tag{9}$$

where $k_i = 2p_i - 3$, $p_i \geq 2$; $E(x_i) = \mu_i$ and $v(x_i) = \sigma_i^2$. Assume that p_i is known. For p_i = 5, (9) is almost indistinguishable from logistic distribution, because the two distributions are both symmetric and have first four moments common (Pearson, 1963). If the two distributions are plotted, it will be seen that one sits almost on top of the other. It may be noted that

$$t = \frac{\{\upsilon_i(x_i - \mu_i)\}^{\frac{1}{2}}}{\sigma_i(k_i)^{\frac{1}{2}}}$$

has Student's t – distribution with ($2p_i$–1) degrees of freedom. For $1 \le p_i \le 2$, k is equal to 1 in which case σ in (9) is simply a scale parameter.

Given the data matrix ($n > r$+1) of the form

$$(y_j; x_{j1}, \ldots., x_{ji}, \ldots.x_{jk}),\ j = 1, 2, \ldots, n \tag{10}$$

where y is the response variable and the x terms as explanatory variables or covariates. Then the likelihood function L based on relation (1) can be written as usual and is given by

$$L \approx \left\{\left(\prod_{i=o}^{r} \sigma_i^2\right)\left(1 - \sum_{i=1}^{r} \rho_{oi}^2\right)\right\}^{-\frac{n}{2}}$$
$$* \exp\left[-\frac{1}{2\sigma_o^2\left(1 - \sum_{i=1}^{r} \rho_{oi}^2\right)} \sum_{j=1}^{n} \left\{y_{[j]} - \mu_o - \sum_{i=1}^{r} \rho_{oi}\left(\frac{\sigma_o}{\sigma_i}\right)\left(x_{i(j)} - \mu_i\right)\right\}^2\right] \tag{11}$$
$$* \prod_{j=1}^{n} \prod_{i=1}^{r} \left\{1 + \frac{(x_i - \mu_i)2}{k_i \sigma_i^2}\right\}^{-p_i}$$

where $x_{i(j)}$, $(i = 1, 2, \ldots, r\,;\, j = 1, 2, \ldots, n)$ are the order statistics of x_i observations, and $y_{[j]}$ $(j = 1, .., n)$ are the corresponding concomitant y observations. The maximum likelihood estimators are the solutions of the likelihood equations, i.e, of the derivatives of ℓnL. These equations are, however, intractable. Solving them by iterative procedures may be problematic, for example, one may encounter multiple roots, slow convergence, or convergence to wrong values (see specifically Barnett, 1966; Lee et al., 1980; Tiku and Suresh, 1992; Vaughan, 1992). Instead the Tikus method of modified likelihood (MML) estimation was employed, which gives explicit estimators and involves replacing intractable terms by linear approximations. Because this method is already well established and is known to produce estimators which are fully efficient for large n (Tiku, 1970; Bhattacharyya, 1985) and almost fully efficient for small n (Tiku et al, 1986; Tiku and Suresh, 1992; Vaughan, 1992, 1994).

Modified Maximum Likelihood

Consider the i^{th} covariate of a random sample of size n denoted by $x_{1i}, x_{2i}, \ldots, x_{ni}$ from any location-scale distribution with density given by

$$\frac{1}{\sigma_i} f\left[\frac{x_{ji}-\mu_i}{\sigma_i}\right], \quad i = 1, 2, \ldots, r .$$

For simplicity of notation, suppress the suffix i and consider f to be a student t density. Then the likelihood equations for estimating μ and σ corresponding to each covariate are

$$\frac{\partial \ell nL}{\partial \mu} = \frac{2p}{k\sigma}\sum_{j=1}^{n} g(z_j) = 0$$

and

$$\frac{\partial \ell nL}{\partial \sigma} = \frac{-n}{\sigma} + \frac{2p}{k\sigma}\sum_{j=1}^{n} z_j g(z_j) = 0$$

where (12)

$$z_j = \frac{(x_j - \mu)}{\sigma}$$

and

$$g(z_j) = \frac{z_j}{\left\{1 + (1/k) z_j^{\,2}\right\}} .$$

Equations (12) do not provide explicit solutions. Following Tiku-Suresh (1992); Vaughan and Tiku (2000), the first step is to express these equations in terms of order statistics $x_{(1)} \le x_{(2)} \ldots \le x_{(n)}$. Because complete sums are invariant to ordering

$$\frac{\partial \ell nL}{\partial \mu}=\frac{2p}{k\sigma}\sum_{j=1}^{n}g(z_{(j)})=0$$

and

$$\frac{\partial \ell nL}{\partial \sigma}=\frac{-n}{\sigma}+\frac{2p}{k\sigma}\sum_{j=1}^{n}z_{(j)}g(z_{(j)})=0 \tag{13}$$

where

$$z(j)=\frac{(x_{(j)}-\mu)}{\sigma},\ j=1,2,...,n.$$

Under appropriate regularity considerations which are very general in nature, $g(z_{(j)})$ can be replaced by linear approximations given by the first two terms of Taylor series expansions (Tiku, 1967, 1968; Tiku and Suresh, 1992; Tiku and Kambo, 1992, Vaughan, 1992; Vaughan and Tiku, 2000), so that

$$\begin{aligned} g\left\{z_{(j)}\right\} &\approx g\left\{t_{(j)}\right\}+\left[z_{(j)}-t_{(j)}\right]\left\{\frac{d}{dz}g(z)\right\}_{z=t_{(j)}} \\ &=\alpha_j+\beta_j z_{(j)},\ j=1,2,...,n \end{aligned} \tag{14}$$

where

$$t_{(j)}=E\left\{z_{(j)}\right\}.$$

Thus, the modified equations are obtained, i.e.

$$\frac{\partial \ell nL}{\partial \mu}\approx\frac{\partial \ell nL^*}{\partial \mu}=\frac{2p}{k\sigma}\sum_{j=1}^{n}\left\{\alpha_j+\beta_j z_{(j)}\right\}=0$$

and (15)

$$\frac{\partial \ell nL}{\partial \sigma}\approx\frac{\partial \ell nL^*}{\partial \sigma}=\frac{-n}{\sigma}+\frac{2p}{k\sigma}\sum_{j=1}^{n}z_{(j)}\left\{\alpha_j+\beta_j z_{(j)}\right\}=0$$

Equations (15) have explicit solutions, which are called modified maximum likelihood (MML) estimators. Note that the ML and MML estimators are asymptotically equivalent.

For distribution $(p\geq 2, k=2p-3)$

$$h(x_j)\infty(k\sigma)^{-\frac{1}{2}}\left\{1+\frac{x_{(j)}-\mu}{k\sigma^2}\right\}^{-p} ,-\infty\angle x_j\angle\infty \tag{16}$$

This method gives the following MML estimators (see Tiku and Suresh, 1992; Tiku and Kambo, 1992; Vaughan, 1992; Vaughan and Tiku, 2000; Tiku et al, 2008)

$$\hat{\mu}=\frac{1}{m}\sum_{j=1}^{n}\beta_j x_{(j)} \qquad (m=\sum_{j=1}^{n}\beta_j) \tag{17}$$

and

$$\hat{\sigma}=\frac{\left\{B+(B^2+4nc)^{\frac{1}{2}}\right\}}{2\left\{n(n-1)\right\}} \tag{18}$$

where

$$B=\frac{2p}{k}\sum_{j=1}^{n}\alpha_j x_{(j)}\ and\ C=\frac{2p}{k}\left\{\sum_{j=1}^{n}\beta_j y^{2}_{(j)}-m\hat{\mu}^{2}\right\} \tag{19}$$

The coefficients α_j and β_j are obtained from the equations

$$\alpha_j=\frac{(2/k)\,t^3_{(j)}}{\left\{1+\left(\frac{1}{k}\right)t^2_{(j)}\right\}^2}\ \text{and}\ \beta_j=\frac{1-(1/k)\,t^2_{(j)}}{\left\{1+(1/k)\,t^2_{(j)}\right\}^2}, j=1,2,...,n \tag{20}$$

For $p=\infty$ (i.e. for normal distribution), $\alpha_j=0$ and $\beta_j=1$, because k=2p−3. Note that $\alpha_j=-\alpha_{n-j+1}$, $\beta_j=\beta_{n-j+1}$ $(1\le j\le n)\,and\sum_{j=1}^{n}\alpha_j=0$. Tables of the value of $t_{(j)}$ are available for p=2(.5) 10 and $n \le 20$ (Tiku and Kumra, 1985). For $n > 20$, $t_{(j)}$ are obtained from the equation

$$\int_{-\infty}^{t_{(j)}} f(z)dz = \frac{j}{n+1} \qquad (1 \le j \le n). \tag{21}$$

In evaluating (21), it should be noted that $\{(k/\upsilon)z\}^{1/2}$ has student's t-distribution with $\upsilon = 2p-1$ degrees of freedom.

It may be of interest to note that in deriving the estimators μ and σ given by the equations (17)-(20), the method of MML estimation for $p < \infty$ automatically gives small weights to extreme order statistics close to the center. It is precisely due to this fact these estimators are robust to reasonable departures from the true value of *p* in (16). In most applications, therefore, it is not very important to pinpoint the true value of *p* and use it in all derivatives. Any reasonable value of *p* gives almost optimal results.

A Q-Q plot can be employed to give a reasonable value closure (if not exactly) the true value of *p* corresponding to covariate *x* (Tiku et al, 1986, p.277). The order statistic $x_{(j)}$ is plotted against the values $t_{(j)} = E(z_{(j)}),\ z_j = (x_{(j)} - \mu)/\sigma,\ j = 1,2,...,n$, under the assumed model, i.e. for a particular value of *p* in (16). If the plot gives a straight line (or nearly so), the model is taken to be valid for the MML estimation.

Following the above procedure, the parameters μ_i and $\sigma_i (i = 1,2,...,r)$ are estimated. In order to estimate the remaining parameters viz., $\mu_o, \sigma_o, \rho_{oi}\ (i = 1,2,...,r)$, the likelihood function (11) is considered. Because $\frac{\partial \ell nL}{\partial \mu_i}$ and $\frac{\partial \ell nL}{\partial \sigma_i}$, (*i*=1,2,…,*r*) are expressed in terms of $g\{z_{i(j)}\}$, the likelihood equations $\frac{\partial \ell nL}{\partial \mu_i} = 0,\ \frac{\partial \ell nL}{\partial \sigma_i} = 0\ \ (i = 1,2,...,r)$ and $\frac{\partial \ell nL}{\partial \rho_{oi}} = 0\ \ (i = 1,2,...,r)$ have no explicit solutions. The modified likelihood equations are $\frac{\partial \ell nL^*}{\partial \mu_i} = 0, \frac{\partial \ell nL^*}{\partial \sigma_i} = 0$, (*i*= 0,1,…,*r*) and $\frac{\partial \ell nL^*}{\partial \rho_{oi}} = 0\ (i = 1,2,...,r)$, and are obtained by replacing $g\{z_{i(j)}\}$ with the linear approximations given by (14). The solutions of these equations are the following MML estimators:

$$\hat{\mu}_o = \bar{y} - \sum_{i=1}^{r} \hat{\rho}_{oi} \left(\frac{\hat{\sigma}_o}{\hat{\sigma}_i} \right) (\bar{x}_i - \hat{\mu}_i) \tag{22}$$

$$\hat{\sigma}_o = \left[s_o^2 + \sum_{i=1}^{r} \left\{ \frac{s_{oi}^2}{s_i^2} \left(\frac{\sigma_i^2}{s_i^2} - 1 \right) \right\} \right]^{1/2} \tag{23}$$

$$\hat{\rho}_{oi} = \frac{s_{oi}^2}{s_i^2} \left[\frac{\hat{\sigma}_1}{\hat{\sigma}_o} \right], \quad i = 1, 2, ..., r \tag{24}$$

Here,

$$n\bar{y} = \sum_{j=1}^{n} y_{[j]} = \sum_{j=1}^{n} y_j \text{ , } n\bar{x} = \sum_{j=1}^{n} x_{i(j)} = \sum_{j=1}^{n} x_{ij} \tag{25}$$

$$(n-1)s_o^2 = \sum_{j=1}^{n} \left[y_{[j]} - \bar{y} \right]^2 = \sum_{j=1}^{n} \left(y_j - \bar{y} \right)^2 \tag{26}$$

$$(n-1)s_i^2 = \sum_{j=1}^{n} \left[x_{i(j)} - \bar{x}_i \right]^2 = \sum_{j=1}^{n} \left(x_{ij} - \bar{x}_i \right)^2 \tag{27}$$

and

$$\begin{aligned} (n-1)s_{oi}^2 &= \sum_{j=1}^{n} \left[y_{[j]} - \bar{y} \right] \left[x_{i(j)} - \bar{x}_i \right] \\ &= \sum_{j=1}^{n} \left[y_j - \bar{y} \right] \left[x_{ij} - \bar{x}_i \right], i = 1, 2, ..., r. \end{aligned} \tag{28}$$

Relation (22) provides for the measurement of dependency of the response random variable on the random covariates x_i $(i = 1, 2, ..., r)$. The linear relationship is also represented in the form of classical model (6).

The asymptotic variances and covariances of the estimators $\hat{\mu}_o, \hat{\mu}_1, \hat{\sigma}_o, \hat{\sigma}_i$ and $\hat{\rho}_{oi}$ $(i = 1, 2, ..., r)$ are obtained with the use of the second partial derivatives of the likelihood function (11). The matrix formed by the negative of the expected values of the second partial derivatives gives the information matrix, which may be expressed as the partitioned matrix

$$\underline{V} = \begin{pmatrix} \underline{V}_1 & \underline{O} \\ \underline{O} & \underline{V}_2 \end{pmatrix} \tag{29}$$

where the matrix is of the order (3r+2) × (3r+2) and

$$\underline{V}_1 = E\left[-\frac{\partial^2 \ell n L^*}{\partial\mu_i \partial\mu_{i'}}\right]$$

of order (r+1) × (r+1) and

$$\underline{V}_2 = E\left[-\frac{\partial^2 \ell n L^*}{\partial\theta_i \partial\theta_{i'}}\right] \quad , \quad i, i' = 1, 2, ..., r$$

of order (2r+1) × (2r+1) with $(\theta_1 = \sigma_o, \theta_2 = \sigma_1, ..., \theta_{2r+1} = \rho_{ok})$.

The inverse of $\underline{V}_1$ and $\underline{V}_2$ matrices provides the elements of the precision and covariance structure of the estimated coefficients.

The estimated values of the parameters obtained above are used in relation (7) and (8) which give the estimated values of the regression coefficients α and β_i $(i = 1, 2, ..., r)$ of the model (6). The asymptotic covariance structure of the estimated regression coefficients $\overset{\wedge}{\alpha}$ and $\overset{\wedge}{\beta}_i$ $(i = 1, 2, .., r)$ are obtained using delta method (Serfling, 1980) as:

Let $g(\alpha, \beta)$ and $\underline{\theta} = (\underline{\mu}, \underline{\sigma}, \underline{\rho})$, then

$$\begin{pmatrix} \overset{\wedge}{\alpha} \\ \overset{\wedge}{\beta} \end{pmatrix} \cong N_{(r+1)}\left[\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \quad n^{-1} G^T \overset{\wedge}{\Sigma} G\right], \tag{30}$$

where

$$G = \left[\frac{\partial \underline{g}}{\partial \underline{\theta}}\right]_{\overset{\wedge}{\mu}_o, ..., \overset{\wedge}{\rho}_{ok}}$$

of order (3r+2) × (r+1) and

$$\hat{\Sigma} = \begin{pmatrix} \underline{\hat{V}}_1^{-1} & \underline{O} \\ \underline{O} & \underline{\hat{V}}_2^{-1} \end{pmatrix}$$

of order (3r+2) × (3r+2). Note that when $p = \infty$ the distribution (16) reduces to the ideal normal distribution in which case $\hat{\mu} = \bar{x}$ (sample mean) and $\hat{\sigma}^2 = s^2$ (sample variance), $\bar{x}$ *and* s^2 being optimal under the assumption of normality.

Examples

Example 1

Consider the part of the data set pertaining to 20 male insulin-dependent diabetic patients as provided in Dobson (1990, p. 69), which is reproduced in Table 1.

Table 1. Carbohydrate, age and weight for twenty insulin-dependent diabetics

y = Carb. (gm)	x_1 = Age (yrs)	x_2 = Wgt (kg)	y = Carb. (gm)	x_1 = Age (yrs)	x_2 = Wgt (kg)
33	33	100	50	31	108
40	47	92	51	61	85
37	49	135	30	63	130
27	35	144	36	40	127
30	46	140	41	50	109
43	52	101	42	64	107
34	62	95	46	56	117
48	23	101	24	61	100
30	32	98	35	48	118
38	42	105	37	28	102

In this sample, the goal is to establish the relationship between the response variable y (amount of carbohydrate) and the two covariates x_1 (age) and x_2 (body weight, relative to “ideal” weight for height) using the linear regression model (6) which takes the form

$$E\left(y \middle| x_1, x_2\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 \tag{31}$$

Here, it is assumed that, in relation (1), the conditional distribution of the response random variable y is normal; however, the covariates follow independently non-normal symmetric distribution. The model (31) is fitted using above described modified maximum likelihood method.

First obtain the values of p_1 and p_2 corresponding to the two covariates x_1 and x_2 using Q-Q plots, where the order statistics $x_{1(j)}$ and $x_{2(j)}$ were plotted separately against $t_{1(j)}$ and $t_{2(j)}$ respectively, $j = 1,\ldots,n$ for different values of p as given in Tiku and Kumra (1985). The values of $p_1 = 5$ and $p_2 = 7$ provided an approximate straight line patterns which determined the appropriate types of densities in (16). Once p_1 and p_2 are known, then using the equations (17)-(20), the MML estimates of the parameters μ_1, σ_1 and μ_2, σ_2 are obtained. Using these values in equations (22)-(28) the rest of the parameters $\mu_o, \sigma_o, \rho_{o1}$ and ρ_{o2} are estimated. Solutions of the information matrix (29) provided the elements of the precision and covariance structure of the estimated parameters. The estimated values and their standard errors are presented in Table 2.

Table 2. MML estimates of the parameters and their standard errors for the data set in Table 1

Param.	Est.	Std. Err.
μ_o	37.732	1.848
μ_1	46.437	3.008
μ_2	109.936	3.776
σ_o	7.635	1.411
σ_1	13.989	1.789
σ_2	17.265	2.351
ρ_{o1}	-0.064	0.228
ρ_{o2}	-0.420	0.243

Table 3. MML and ML estimates of the parameters and their standard errors for the data set in Table 1

	Param.	Est.	Std. Err.	W
	Constant (α)	59.783	12.469	
MML	Coefficient (β_1)	-0.035	0.124	-0.282
	Coefficient (β_2)	-0.186	0.099	-1.879
	Constant (α)	60.432	13.017	
ML	Coefficient (β_1)	-0.046	0.131	-0.351
	Coefficient (β_2)	-0.187	0.101	-1.851

Using the estimated values in Table 2 in relation (7) and (8), obtain MML estimates of the regression parameters α, β_1 and β_2. Use of delta method as described in (30) provided the asymptotic standard errors; also these parameters based on usual maximum likelihood method were estimated. The results, obtained under the two methods are summarized in Table 3.

The analysis in Table 3 reveals that the MML estimates of the regression parameters for the data set in Table 1 are very close to the values obtained using maximum likelihood method, as expected. Moreover, the two methods gave approximately the same results for the Wald statistics *W*, which permits to test the null hypothesis $H_o : \beta_1 = 0 \; and \; \beta_2 = 0$. For large n, the null distribution of W is referred to a standard normal distribution.

Example 2

Consider another data set from Murray (1937), reproduced in El-Saidi (1995, p. 214) as shown in Table 4. The data provides 11 observations on the number of male flies died after twenty minutes exposure to pyrethrum at various concentrations.

The main objective is to describe the probability of success p_j as a function of dose x_j. In literature, such type of analysis are carried out usually considering either probit or logit models (Cox, 1970). However, the logit model is preferred to a probit model due to two primary reasons (Hosmer and Lameshow, 1989): from mathematical point of view, it is an easily used function, and it leads to itself to a biological meaningful interpretation.

Table 4. Mortality of male flies after twenty minutes exposure to pyrethrum

Concentration ($\log_{10}$)	Number of flies		Proportions Died
	Exposed	Died	
1.6020	462	109	0.2359
1.7782	500	199	0.3980
1.9031	467	298	0.6381
2.0000	515	370	0.7185
2.0792	561	459	0.8182
2.1461	469	400	0.8529
2.2041	550	495	0.9000
2.2553	542	499	0.9207
2.3010	479	450	0.9395
2.3979	497	476	0.9577
2.4771	453	442	0.9757

The logit model is a family of Generalized Linear Models (GLMs) with link function $g(p_j)$ as $\ell n\left(\frac{p_j}{1-p_j}\right)$ (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989). The link function $g(p_j)$ is continuous and maps the $\left[0,1\right]$ range of probabilities onto $\left(-\infty,\infty\right)$ and is represented by

$$g(p_j) = \ell n\left(\frac{p_j}{1-p_j}\right) = \alpha + \beta x_j,\ j = 1,2,\dots,n \tag{32}$$

so that

$$p_j = \frac{\exp(\alpha + \beta x_j)}{1 + \exp(\alpha + \beta x_j)} \quad ,\ j = 1,2,\dots,n \tag{33}$$

The relation (33) is known as binary logistic model with probability of success p_j, this belongs to the standardized logistic distribution which is symmetric in nature (Rao and Toutenburg, 1995, p. 263).

In order to estimate the unknown parameters α and β in (32), usually ML method is used. The technique involves the solution of the likelihood equations, which have no explicit solutions and have to be solved by interactive procedures. Solving these equations is, therefore, tedious and time consuming. Therefore, these parameters are estimated using MML method.

For this, consider the link function i.e. log odds as a response variable and x_j as a covariate. First estimate μ_1 and σ_1 for p=5 in distribution (16). Using these values in equations (22)-(28), the rest of the parameters μ_o, σ_o *and* ρ_{o1} involved in the likelihood function (11) were obtained. The estimated values of the variances and co-variances were obtained using these values in second partial derivatives of the likelihood function (11) and solving for the inverse of the information matrix (29). The estimated values of the parameters and their standard errors involved in the likelihood function (11) with p =5 for the data set in Table 4 are shown in Table 5.

Using these estimated values of the parameters in relation (7) and (8), obtain the MML estimates of the parameters $\hat{\alpha}$ and $\hat{\beta}$ of the logistic model (33). The use of delta method (30) gave the asymptotic variances of $\hat{\alpha}$ and $\hat{\beta}$. The ML estimates of these parameters and their variances under the logit model (32) were also obtained using iterative procedures viz; Newton-Raphson method (Cox, 1970, Chapter 2). The results obtained under the two procedures are summarized in Table 6.

These analyses also reveal that the MML estimates of the regression parameters α and β for the data set in Table 4 are very close to the values obtained using maximum likelihood method, as expected.

Table 5. MML estimates of the parameters and their standard errors for the data set in Table 4

Param.	Est.	Std. Err.
μ_o	1.642	0.459
μ_1	2.115	0.082
σ_o	1.593	0.198
σ_1	0.284	0.035
ρ_{o1}	0.999	0.003

Table 6. MML and ML estimates of the parameters and their standard errors for logit model (32)

	Param.	Est.	Std. Err.
MML	Constant (α)	-10.219	0.186
	Coefficient (β)	5.608	0.087
ML	Constant (α)	-10.329	0.343
	Coefficient (β)	5.661	0.172

This study used Tiku's modified maximum likelihood method for carrying out regression analysis when the underlying distributions of the data set have non-normal symmetric distributions. The method yields estimators which are explicit functions of sample observations and are numerically very close to the maximum likelihood estimators and equally efficient.

References

Barnet, V. D. (1966). Order statistics estimators of the location of the Cauchy distribution. *Journal of America Statistical Association*, *61*(316): 1205-1218.

Bhattarcharyya, G. K. (1985). The asymptotics of maximum likelihood and related estimators based on type II censored data. *Journal of American Statistical Association*, *80*(390): 398-404.

Cox, D. R. (1970). *The analysis of binary data.* Methuen: London.

Dobson, A. J. (1990). *An introduction to generalized linear models*. Chapman and Hall: New York.

El-Saidi, M. A. (1995). A symmetric extended logistic model with applications to experimental toxicity data. *Biometrical Journal*, *37*(2)*,* 205-216.

Hosmer, D. W. and Lemeshow, S. (1989). *Applied logistic regression.* John Wiley: New York.

Lee, K. R., Kapadia, C. H. and Dwight, B. B. (1980). On estimating the scale parameters of the Rayleigh distribution from doubly censored samples. *Statistische Hefte*, *21*(1): 14-29.

McCullagh, P and Nelder, J. A. (1989). *Generalized linear models.* Chapman and Hall: London.

Murray, C.A. (1937). A statistical analysis of fly mortality data. *Soap, 13*(8): 89-105.

Nelder, J. A. and Wedderburn, R. W. N. (1972). Generalized linear models. *Journal of Royal Statistical Society*, *Series A, 135*(3): 370-384.

Pearson, E. S. (1963). Some problems arising in approximating to probability distributions using moments. *Biometrika*, *50*, 95-112.

Rao, C. R. and Toutenburg, H. (1995). *Linear models: least squares and alternatives*. Springer-Verlag: New York.

Serfling, R. J. (1980). Approximation theorems of mathematical studies. Wiley: New York.

Tiku, M. L. (1967). Estimating the mean and standard deviation from a censored normal sample. *Biometrika*, *54*(1/2): 155-165.

Tiku, M. L. (1968). Estimating the parameters of normal and logistic distributions form censored samples. *Australian Journal of Statistics*, *10*(2): 64-74.

Tiku, M. L., Islam, M. Q. and Sazak, H.S. (2008). Estimation in bivariate non-normal distributions with stochastic variance functions. *Computational Statistics & Data Analysis*, *52*(3): 1728-1745.

Tiku, M. L. (1970) Monte Carlo study of some simple estimators in censored normal samples. *Biometrika*, *57*(1): 207-211.

Tiku, M. L. and Kambo, N. S. (1992) Estimation and hypothesis testing for a new family of bivariate non normal distributions. *Communications in Statistics – Theory and Methods, 21*(6): 1683-1705.

Tiku. M. L. and Kumra, S. (1985). Expected values and variances and covariances of order statistics for a family of symmetric distributions (Student's *t*). In B. J. Trawinski, R. E. Bechhofer, S. Kumra, M. L. Tiku, & A. C. Tahmane (Eds.) *Selected tables in mathematical statistics, Vol. 8.* Providence, R.I.: American Mathematical Society: pp. 141-270.

Tiku, M. L. and Suresh, R. P. (1992). A new method of estimation for location and scale parameters. *Journal of Statistical Planning and Inference*, *30*(2): 281-292.

Tiku, M. L., Tan, W. Y. and Balakrishnan, N. (1986). *Robust Inference*. Marcel Dekker : New York.

Vaughan, D. C. (1992). On the Tiku-Suresh method of estimation. *Communications in Statistics – Theory and Methods, 21*(2): 451-469.

Vaughan, D. C. (1994). The exact values of the expected values, variances and covariances of the order statistics from the Cauchy distribution. *Journal of Statistical Computation and Simulation*, *49*(1-2): 21-32.

Vaughan, D. C. & Tiku, M. L. (2000). Estimation and hypothesis testing for a non-normal bivariate distribution with applications. *Mathematical and Computer Modelling*, *32*: 53-67.