 

# Fitting a non-linear model with errors in both variables and its application*

**M. SINGH**[1] **& G. K. KANJI,**[2] [1]*Statistics Unit, International Crops Research Institute for the Semi-Arid Tropics, Patancheru, A.P. India and* [2]*Department of Applied Statistics & Operational Research, Sheffield City Polytechnic*

SUMMARY *Estimation of the parameters of a non-linear model is considered when both measured variables have random errors. The maximum likelihood estimates with the asymptotic variance and covariance matrix are presented. Real data are used to illustrate the procedure discussed.*

## 1 Introduction

In field experiments, there are numerous situations where interest lies in obtaining the functional relationship between two variables, $X$ and $Y$, of the form $Y=f(X)$. When $X$ is free from measurement error, i.e. the observations recorded on $X$ are true values of $X$, we have the usual regression model $Y=f(X)+\varepsilon$, $\varepsilon$ being sum of the random error unaccounted for by the model $f(X)$ and the error of measurement of $Y$. When $X$ is measured with errors (considered random), the linear and quadratic functional relations between $Y$ and $X$ have been considered by several authors (Madansky, 1959; Kendall & Stuart, 1979; Causton & Venus, 1981; Wolter & Fuller, 1982). In many biological experiments the functional relationships are non-linear in parameters and variable $X$. For instance, Fig. 1 shows percent light interception with leaf area index (area of leaves per unit area of land) for a pearl millet cultivar. The records of both variables contain measurements errors. The functional relationship between these two variables conceptualised by plant physiologists, and indicated by the graph is sigmoid.

$$y=\alpha(1-e^{-\beta x}) \tag{1}$$

*Submitted as Journal Article No. JA 738 by ICRISAT, and presented at the 46th Session of the International Statistical Institute Conference, 8–16 September 1987, held in Tokyo, Japan.

This relationship provides an understanding of the mechanism of the vegetative growth and hence the total biomass production of the crop. Another example is that relating drought recovery response with proline content (measurement with error) in plant tissues at the end of a drought period.

Non-linear functional relationship, when both variables are subject to errors have received little attention in the literature. In Sections 2 and 3 of this paper we estimate the parameters of equation (1); its application to pearl millet data is considered in Sections 4 and 5.

## 2 A non-linear model with errors in both variables

Consider $n$ pairs of random observations $\{(x_i, y_i), i=1, \ldots, n\}$ such that the measurements $x_i$ and $y_i$ are made independently on two variables, $X$ and $Y$ respectively, and follow the distributions represented by the probability density functions:

$$x_i \sim f_1\ (\mu_{1i}, \sigma_1^2) \tag{2.1}$$

$$y_i \sim f_2\ (\mu_{2i}, \sigma_2^2) \tag{2.2}$$

where $f_1(.)$, $f_2(.)$ are probability density functions; $\mu_{1i}$, $\mu_{2i}$ are means and $\sigma_1^2$, $\sigma_2^2$ are variances of $X$ and $Y$ variables respectively at $i$th unit of measurements. The functional relationship between $X$ and $Y$ is given by:

$$\mu_{2i} = g(\mu_{1i}),\ i=1, \ldots, n \tag{2.3}$$

When $f_1(.)$ and $f_2(.)$ are normal densities and $\mu_{1i}$, $\mu_{2i}$ are expected values of $x_i$ and $y_i$'s respectively, we can write (2.1) and (2.2) as:

$$x_i = \mu_{1i} + \varepsilon_{1i} \tag{2.4}$$

$$y_i = \mu_{2i} + \varepsilon_{2i} \tag{2.5}$$

where $\varepsilon_{1i} \sim N(0, \sigma_1^2)$, and $\varepsilon_{2i} \sim N(0, \sigma_2^2)$. The relations in (2.3) considered as linear, i.e.

$$\mu_{2i} = \alpha + \beta\mu_{1i} \tag{2.6}$$

and considered as quadratic i.e.

$$\mu_{2i} = \alpha + \beta\mu_{1i} + \gamma\mu_{1i}^2 \tag{2.7}$$

have been discussed by Kendall & Stuart (1979, Chapter 29), Causton & Venus (1981, pp. 182–209) and Wolter & Fuller (1982), among others.

We consider the following non-linear model. Let $X$ and $Y$ represent the two variables and $x_i$ and $y_i$ be values of these variables on the $i$th experimental unit $i=1 \ldots n$. In this situation, we consider the non-linear relation:

$$\mu_{2i} = \alpha(1 - e^{-\beta\mu 1i}) \tag{2.8}$$

although a more general relation would be

$$\mu_{2i} = \alpha + \gamma e^{-\beta\mu 1i} \tag{2.9}$$

We shall discuss in detail the estimation of the relation (2.8) and reparameterise it as below. With errors of measurements normally distributed, we have:

$$x_i = \xi_i + \eta_i \tag{2.10}$$

$$y_i = \zeta_i + \varepsilon_i \tag{2.11}$$

$$\zeta_i = \alpha(1 - e^{-\beta\xi_i}) \tag{2.12}$$

where $\eta_i$, $\varepsilon_i$ are independent and normally distributed with means zero and variances

$$\text{Var}\ (\eta_i) = \sigma_\eta^2 \tag{2.13}$$

$$\text{Var}\ (\varepsilon_i)=\sigma_\varepsilon^2 \xi_i^{\Theta} \tag{2.14}$$

The zero value of $\Theta$ implies the constancy of variances of errors $\varepsilon_i$, but Fig. 1 indicates that the variance of $\varepsilon_i$ is increasing with $x_i$ and hence with $\xi_i$. We have considered $\Theta$ to be known from an estimation point of view, however, the value of $\Theta$ has been chosen as the one that results in the minimum mean square for $Y$ when fitted to the model without considering errors of measurements.

## 3 Estimation of parameters

To estimate $\alpha$, $\beta$, $\xi_i$ $(i=1,\ldots, n)$, $\sigma_\varepsilon^2$, $\sigma_\eta^2$, we have considered the value of $\Theta$ to be known. Thus, these estimates are conditional to $\Theta$.

### *3.1 Choice of* $\Theta$

To choose $\Theta$, we fitted the non-linear regression model:

$$\begin{aligned} y_i&=\alpha(1-e^{-\beta x_i})+\varepsilon_i \\ \text{Var}\ (\varepsilon_i)&=\sigma_\varepsilon^2 x_i^{\Theta} \end{aligned} \tag{3.1}$$

for several known values of $\Theta$ and estimated the corresponding values of $\sigma_\varepsilon^2$ by residual mean square (RMS), $\hat{\sigma}_\varepsilon^2$. These values of $\hat{\sigma}_\varepsilon^2$ were plotted against $\Theta$. That value of $\Theta$ that results in the minimum $\hat{\sigma}_\varepsilon^2$ would be chosen for $\Theta$. We believe that such a value of $\Theta$ could be considered to estimate $\alpha$ and $\beta$ and $\xi_i$ of the errors in variables model.

### *3.2 Estimation of* $\alpha$ *and* $\beta$

The maximum likelihood method of estimation of the parameters in linear and quadratic functional relation (Kendall & Stuart, 1979); will be considered here. We assumed that the ratio $\lambda=\sigma_\varepsilon^2/\sigma_\eta^2$ is known. The likelihood function $L$ of the parameters $\alpha$, $\beta$, $\xi_i$, $i=1,\ldots, n$, $\sigma_\eta^2$ based on observations $(x_i, y_i)$, $i=1,\ldots, n$ can be written as:

$$\begin{aligned} L=(2\pi\lambda\sigma_\eta^2)^{-n/2n}\prod_{i=1}\xi_i^{-\Theta/2}\exp\left(-(1/2)\sum_{i=1}^{n}(y_i-\alpha(1-e^{-\beta x_i}))^2/(\lambda\sigma_\eta^2\xi_i^{\Theta})\right) \\ (2\pi\,\sigma_\eta^2)^{-n/2}\exp(-(1/2)\Sigma(x_i-\xi_i)^2/(\sigma_\eta^2)) \end{aligned} \tag{3.2}$$

Writing $l=\log L$, we have:

$$\begin{aligned} l=\text{constant}\ &-n\log\,(\lambda^{1/2}\,\sigma_\eta)-\Sigma\xi_i^{-\Theta}\gamma_i^2/(2\lambda\sigma_\eta^2)-(\Theta/2)\Sigma\log(\xi_i) \\ &-n\log\,(\sigma_\eta)-\Sigma(x_i-\xi_i)^2/(2\sigma_\eta^2) \end{aligned} \tag{3.3}$$

where $\gamma_i=y_i-\alpha(1-e^{-\beta x_i})$.

The normal equations obtained by equating the derivatives of $l$ with respect to $\alpha$, $\beta$, $\xi_i$ $(i=1,\ldots, n)$, and $\sigma_\eta^2$ to zeroes, result in the following equations that can be solved iteratively, for $\alpha$, $\beta$, $\sigma_\eta^2$ and $\xi_i$, respectively, with suitable initial values of parameters:

$$\hat{\alpha}=\sum_{i=1}^{n}\xi_i^{-\Theta}\,y_i(1-e^{-\beta x_i})/\sum_{i=1}^{n}\xi_i^{-\Theta}(1-e^{-\beta x_i})^2 \tag{3.4}$$

for estimating $\alpha$. This requires the initial values of $\xi_i$ which may be taken as $x_i$;

$$\Sigma\xi_i^{-\Theta+1}\,e^{-\beta x_i}\gamma_i=0 \tag{3.5}$$

gives solution $\hat{\beta}$ for estimating $\beta$;

$$\hat{\sigma}_\eta^2=\left[\Sigma\xi_i^{-\theta}\,\gamma_i^2/\lambda+\sum_i(x_i-\xi_i)^2\right]/(2n), \tag{3.6}$$

and

$$\hat{\xi}_i=x_i-\Theta\xi_i^{-1}\,\sigma_\eta^2/2+\gamma_i\xi_i^{-\Theta}(\Theta\xi_i^{-1}\gamma_i+2\alpha\beta e^{-\beta\zeta_i})/(2\lambda) \tag{3.7}$$

for estimating $\xi_i$.

### *3.3 Asymptotic variance-covariance matrix*

The asymptotic variance-covariance matrix of various parameter estimates is obtained by computing the information matrix $I$. The element $I(\phi, \phi')$ of information matrix $I$ associated with parameters pair $(\phi, \phi')$ is expressed in terms of log likelihood function $l$ as below.

$$I(\phi, \phi')=E(-\partial^2 l/\partial\phi\partial\phi')=E((\partial l/\partial\phi)(\partial l/\partial\phi'))$$

From the expression for $l$ in the present case we obtain

$$\begin{aligned}
&I(\alpha, \alpha)=(\lambda\sigma_\eta^2\,\alpha^2)^{-1}\,\Sigma\,\zeta_i^2\xi_i^{-\Theta}\\
&I(\alpha, \beta)=-(\alpha\lambda\sigma_\eta^2)^{-1}\,\Sigma\,\xi_i^{-\Theta+1}\zeta_i(\zeta_i-\alpha)\\
&I(\alpha, \xi_1)=-\beta(\alpha\lambda\alpha_\eta^2)^{-1}\,\xi_i^{-\Theta}\zeta_i(\zeta_i-\alpha)\\
&I(\alpha, \sigma_\eta^2)=0\\
&I(\beta, \beta)=(\lambda\sigma_\eta^2)^{-1}\,\Sigma\,\xi_i^{-\Theta+2}(\zeta_i-\alpha)^2\\
&I(\beta, \xi_i)=\beta(\lambda\sigma_\eta^2)^{-1}\,\xi_i^{-\Theta+1}(\zeta_i-\alpha)^2\\
&I(\beta, \sigma_\eta^2)=0\\
&I(\xi_i, \xi_i)=(4\lambda\sigma_\eta^2)^{-1}\,[2\Theta^2\lambda\xi_i^{-2}\sigma_\eta^2+\lambda+4\beta^2\xi_i^{-\Theta}(\zeta_i-\alpha)^2]\\
&I(\xi_i, \xi_i')=0 \qquad i\neq i'=1\ldots n\\
&I(\xi_i, \sigma_\eta^2)=\Theta\xi_i^{-1}\,\sigma_\eta^{-1} \qquad i=1,\ldots n\\
&I(\sigma_\eta^2, \sigma_\eta^2)=4n\sigma_\eta^{-2}
\end{aligned}$$

The variance covariance matrix $V$ is given by $V=I^{-1}$. One can simplify the inversion using the following result (Rao, 1973, p. 33)

$$\begin{pmatrix}A & B\\ B' & D\end{pmatrix}^{-1}=\begin{pmatrix}A^{-1}+F\,E^{-1}F' & -F\,E^{-1}\\ -E^{-1}F' & E^{-1}\end{pmatrix}$$

where $E=D-B'A^{-1}B$ and $F=A^{-1}B$.

### *3.4 Asymptotic confidence interval*

At a fixed point $\xi$, the $\zeta$ could be estimated by

$$\hat{\zeta}=\hat{\alpha}(1-e^{-\hat{\beta}\xi})$$

with the asymptotic variance of $\hat{\zeta}$ as

$$\begin{aligned}
\operatorname{avar}(\hat{\zeta})&=(\partial\zeta/\partial\alpha)^2\,\operatorname{Var}(\hat{\alpha})+(\partial\zeta/\partial\beta)^2\,\operatorname{Var}(\hat{\beta})+2(\partial\zeta/\partial\alpha)(\partial\zeta/\partial\beta)\,\operatorname{Cov}(\hat{\alpha}, \hat{\beta})\\
&=(1-e^{-\beta\xi})^2\,\operatorname{Var}(\hat{\alpha})+(\alpha\xi\,e^{-\beta\xi})^2\,\operatorname{Var}(\hat{\beta})\\
&\quad+2\,\alpha\,\xi\,(1-e^{-\beta\xi})e^{-\beta\xi}\,\operatorname{Cov}\,(\hat{\alpha}, \hat{\beta})
\end{aligned}$$

Estimate of $\operatorname{avar}(\hat{\zeta})$ is obtained by substituting the estimate of $\alpha$, $\beta$, their variances, and covariances.

Some conditions for the asymptotic normality of the maximum likelihood estimates of $\alpha$ and $\beta$ parameters have been recently established by Amemiya & Fuller (1986) for a more general form of the non-linear relationships.

## 4 An illustration

We present here computations for fitting the model (2.10-2.14) to pearl millet data (Table 1). The light interception values ($Y$) are plotted against leaf area index ($X$) for a set of 42 points during the growth period of a millet cultivar (Fig. 1).

TABLE 1. Values of leaf area index ($X$ in m² of leaf/m² of land) and percent light interception ($Y$) measured for 42 plots of a millet cultivar

| Serial no. | $X$ | $Y$ | Serial no. | $X$ | $Y$ | Serial no. | $X$ | $Y$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.15 | 3 | 16 | 1.99 | 45 | 31 | 1.16 | 39 |
| 2 | 0.13 | 5 | 17 | 2.61 | 51 | 32 | 1.40 | 39 |
| 3 | 0.12 | 10 | 18 | 1.96 | 42 | 33 | 1.35 | 51 |
| 4 | 0.41 | 16 | 19 | 1.67 | 49 | 34 | 1.20 | 42 |
| 5 | 0.49 | 23 | 20 | 1.14 | 36 | 35 | 1.55 | 53 |
| 6 | 0.63 | 21 | 21 | 1.56 | 38 | 36 | 1.67 | 55 |
| 7 | 1.01 | 33 | 22 | 1.09 | 46 | 37 | 1.06 | 29 |
| 8 | 1.39 | 37 | 23 | 2.56 | 60 | 38 | 1.40 | 48 |
| 9 | 1.35 | 39 | 24 | 2.30 | 46 | 39 | 1.72 | 63 |
| 10 | 1.34 | 49 | 25 | 1.40 | 25 | 40 | 1.44 | 50 |
| 11 | 1.27 | 46 | 26 | 1.34 | 44 | 41 | 1.39 | 62 |
| 12 | 1.89 | 48 | 27 | 1.56 | 41 | 42 | 1.36 | 47 |
| 13 | 1.50 | 33 | 28 | 1.12 | 42 | | | |
| 14 | 2.20 | 54 | 29 | 1.53 | 57 | | | |
| 15 | 1.64 | 50 | 30 | 1.31 | 58 | | | |

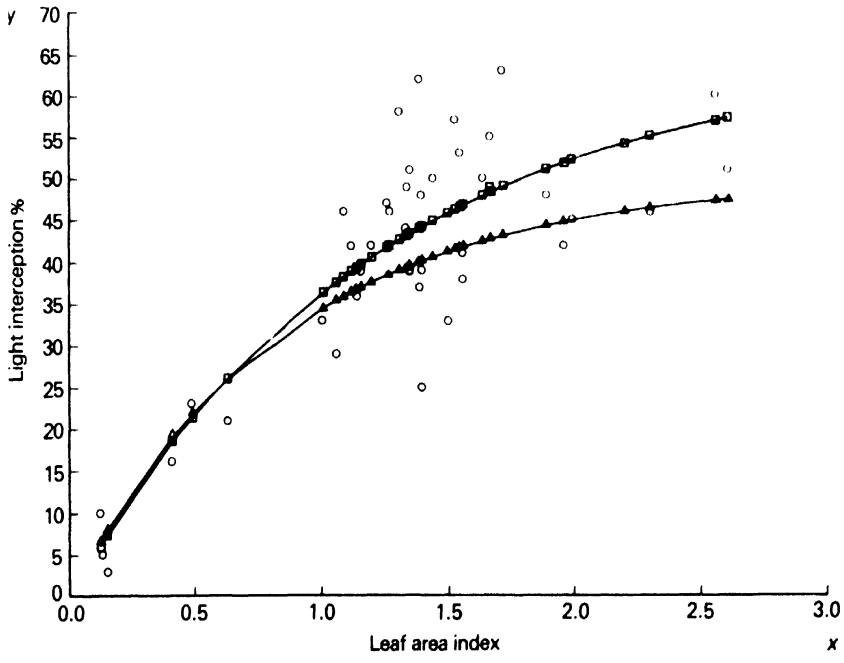


FIG. 1. Percent light interception and leaf area index; Observed value (○) fitted model (2.10-2.14) with $\lambda=1$ (△-△) and fitted model (3.1) (□-□).

△-△: $y=49.81\ (1-e^{-1.179x})$

□-□: $y=65.13\ (1-e^{-0.813x})$

The OPTIMIZE directive in the GENSTAT statistical package can provide the least square estimates of the parameters in the non-linear regression model with normal errors. Similar procedures are also available in the SAS package. The model (3.1) was fitted (using the OPTIMIZE directive in GENSTAT) for a range of $\Theta$ values $\{\Theta = -1.50, -1.00, -0.75, -0.50(0.25)2.00\}$ and estimates of parameters $\alpha$ and $\beta$ and residual mean square (RMS) $\hat{\sigma}_\varepsilon^2$ are given in Table 2. The plotted curve of RMS with $\Theta$ is presented in Fig. 2. The optimum value of $\Theta$ judged by the minimum RMS was one.

TABLE 2. Estimates of parameters $\alpha$ and $\beta$ based on model (3.1) (non-linear regression with varying $\Theta$)

| $\Theta$ | $\hat{\alpha}$ | SE($\hat{\alpha}$) | $\hat{\beta}$ | SE($\hat{\beta}$) | RMS |
|---|---|---|---|---|---|
| −1.50 | 56.40 | 4.677 | 1.136 | 0.2729 | 104.63 |
| −1.00 | 56.96 | 5.155 | 1.103 | 0.2698 | 85.27 |
| −0.75 | 57.38 | 5.440 | 1.081 | 0.2662 | 77.35 |
| −0.50 | 57.92 | 5.761 | 1.054 | 0.2611 | 70.42 |
| −0.25 | 58.61 | 6.125 | 1.022 | 0.2544 | 64.41 |
| 0.00 | 59.48 | 6.540 | 0.986 | 0.2459 | 59.24 |
| 0.25 | 60.56 | 7.013 | 0.946 | 0.2359 | 54.92 |
| 0.50 | 61.87 | 7.554 | 0.902 | 0.2248 | 51.53 |
| 0.75 | 63.41 | 8.167 | 0.857 | 0.2130 | 49.24 |
| 1.00 | 65.13 | 8.855 | 0.813 | 0.2017 | 48.42 |
| 1.25 | 66.89 | 9.612 | 0.773 | 0.1919 | 49.73 |
| 1.50 | 68.47 | 10.436 | 0.741 | 0.1853 | 54.36 |
| 1.75 | 69.56 | 11.338 | 0.722 | 0.1832 | 64.29 |
| 2.00 | 69.94 | 12.384 | 0.716 | 0.1873 | 82.87 |

SE: Estimated asymptotic standard error.

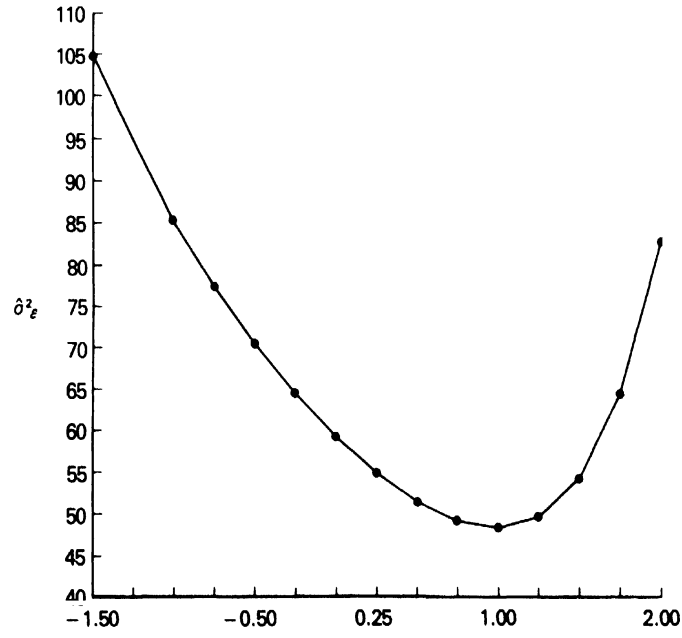


FIG. 2. Residual mean square ($\hat{\sigma}_\varepsilon^2$) and power ($\Theta$) in (2.14).

Thus, the value of $\Theta$ was taken as unity for fitting the model equations (2.10–2.14). The initial value of $\xi_i$ was taken as $x_i$ $(i=1 \ldots n)$, with initial value of $\alpha$ and $\beta$ as the estimates of $\alpha$ and $\beta$ in Table 2 for $\Theta=1$. The OPTIMIZE directive was used again with the weight option. The weight variate $WT(w_i)$ was proportional to the inverse of variance, i.e. $w_i=\xi_i^{-1}$. The new values of $\alpha$ and $\beta$ were used to compute $\sigma_\eta^2$ from (3.6) and $\xi_i$ from (3.7). The estimates of $\xi_i$'s were restricted to the feasible range of the positive leaf area index values. With these new values of $\xi_i$'s, the above process of fitting was repeated until convergence of the values of $\alpha$, $\beta$, $\xi_i (i=1, \ldots, n)$ and $\sigma_\eta^2$ occurred. In the process of computation, a set of five points lead to high residuals resulting in integer overflow. Accordingly the analysis was restricted to the set of remaining 37 points.

This OPTIMIZATION procedure was carried out for several values of $\lambda$. Table 3 gives the estimates, their asymptotic standard errors, and RMS for a few values of $\lambda$ for which convergence was achieved. The plot for the fitted models (3.1) and (2.10–2.14) with $\lambda=1$ along with observed values are displayed in Fig. 1.

TABLE 3. Estimates of $\alpha$, $\beta$ and their asymptotic standard errors based on model in (2.10–2.14) residual mean squares for a range of values of $\lambda$ (the ratio of variances in $X$ and $Y$), $\Theta$ was taken as 1

| $\lambda$ | $\hat{\alpha}$ | SE($\hat{\alpha}$) | $\hat{\beta}$ | SE($\hat{\beta}$) | $\hat{\sigma}_\eta^2$ | RMS $\hat{\sigma}_\eta^2 (=\lambda^{-1}\hat{\sigma}_\varepsilon^2)$ |
|---|---|---|---|---|---|---|
| 0.10 | 62.01 | 3.167 | 0.684 | 0.0581 | 4.49 | 40.90 |
| 0.25 | 59.68 | 3.039 | 0.728 | 0.0645 | 4.92 | 19.68 |
| 0.50 | 57.69 | 2.618 | 0.770 | 0.0628 | 5.44 | 10.88 |
| 1.00 | 49.81 | 5.156 | 1.179 | 0.2606 | 49.49 | 49.49 |
| 1.50 | 53.97 | 3.210 | 0.911 | 0.1056 | 15.74 | 10.49 |
| 2.00 | 54.07 | 2.621 | 0.885 | 0.0842 | 11.38 | 5.69 |
| 4.00 | 53.29 | 2.117 | 0.888 | 0.0724 | 9 46 | 2.37 |

## 5 Results and discussion

We obtained estimates of the parameters in a non-linear model when the independent variable is measured with error. Though it is not the purpose of this paper to cause any effect on the conclusions derived from the example data on the subject matter, we found, however, that the error variation in percent light interception varied in proportion to the leaf area index (Fig. 2). In the set of values 0.10, 0.25, . . ., 2.0, 4.0 for $\lambda$, the estimates and their standard errors for parameters $\alpha$ and $\beta$ showed a tapering off trend around $\lambda=1$ (Table 3). In the close neighbourhood of $\lambda=1$, the error does not increase and a closer fit was obtained at $\lambda=1$ than for any other value of $\lambda$. Thus, under errors in both variables model, with $\lambda=1$, in the example, the estimates $\hat{\alpha}=49.81$ $(\pm 5.156)$ and $\hat{\beta}=1.179$ $(\pm 0.2606)$ can be taken for $\alpha$ and $\beta$ respectively.

The OPTIMIZE directive of GENSTAT showed convergence for all the values of $\Theta$ in its range when considered for fitting the model without errors in variable ($X$). However, in the case of error in $X$, remarkable closeness was found over iteration in the values of $\xi_i$ in the range of its admissible values. The OPTIMIZE showed convergence for estimation of $\alpha$ and $\beta$ for each set of $\xi_i$ obtained iteratively.

We think that the method considered in Section 3 could be used for other non-linear forms worth examining in practice.

**Acknowledgements**

Authors are grateful to Dr G. Alagarswamy, Millet Physiologist for providing the data for illustration. Thanks are also due to Mr Bruce Gilliver, Principal Statistician, Drs A. K. S. Huda, Agroclimatologist, V. M. Ramraj, Groundnut Physiologist and F. R. Bidinger, Principal Millet Physiologist for their helpful comments and Mr C. P. Jaiswal for Secretarial assistance.

*Correspondence:* Dr Murari Singh, Statistics Unit, ICRISAT, Patancheru 502 324, A.P., India.

## REFERENCES

AMEMIYA, Y. & FULLER, W.A. (1986) *Estimation for non-linear functional relationship*, Preprint Number 86-18 (Iowa State University, Ames, Department of Statistics).

CAUSTON, D.R. & VENUS, J.C. (1981) *The Biometry of Plant Growth* (London, Edward Arnold)

KENDALL, M.G. & STUART, A. (1979) *The Advanced Theory of Statistics*, Vol II (London, Charles Griffin)

MADANSKY, A. (1959) The fitting of straight lines when both variables are subject to error, *Journal of American Statistical Association*, 54, pp. 173-205

RAO, C.R. (1973) *Linear Statistical Inference and Its Applications* (New Delhi, Wiley Eastern)

WOLTER, K.M. & FULLER, W.A. (1982) Estimation of the quadratic errors-in-variables model, *Biometrika*, 69, pp. 175-182.