RRB 89-15

# APPLICATION OF NEGATIVE BINOMIAL REGRESSION MODELS TO THE ANALYSIS OF QUANTAL BIOASSAYS DATA

by

A. Maul[1], A.H. El-Shaarawi[2] and J.F. Ferard[3]

[1]Departement Informatique
Université de Nancy II
2 bd Charlemagne
54000 NANCY, France

[2]Rivers Research Branch
National Water Research Institute
Canada Centre for Inland Waters
Burlington, Ontario, L7R 4A6

[3]Centre des Sciences de
l'Environnement
Université de Metz
1 rue des Récollets
57000 METZ, France

February 1989
NWRI Contribution #89-176

## ABSTRACT

Negative binomial and mixed Poisson regression analyses are presented for modelling the association between a quantal response, which is assumed to follow either a negative binomial distribution with a given dispersion parameter, or its limited form (i.e. the Poisson distribution), and a set of explanatory variables. The procedure used for estimating the unknown parameters of the model and performing various statistical tests is given. The method is illustrated by an example about the determination of thresholds in quantal toxicity experiments.

# RÉSUMÉ

Cet ouvrage présente des analyses binomiales négatives et des analyses mixtes de régression de Poisson pour la modélisation de l'association entre une réponse binaire, qui, croît-on, suit une distribution binomiale négative avec un paramètre de dispersion donné, ou sa forme limitée (c.-à-d. la distribution de Poisson), et un ensemble de variables explicatives. On y présente le procédé utilisé pour évaluer les paramètres inconnus du modèle et effectuer divers essais statistiques. La méthode est illustrée à l'aide d'un exemple portant sur l'établissement des seuils dans le cadre d'expériences de toxicité binaire.

## MANAGEMENT PERSPECTIVE

The problem of developing an approach for modelling the response of an organism to chronic toxicity is discussed in this paper and illustrated by studying the toxic effect of NaBr on the reproduction process of a population of Daphnia Magna. A general model is given which includes both the negative binomial and poisson distributions as special cases depending on the values of a single parameter. The steps involved in estimating the parameters of the model and testing the goodness of fit are presented. In particular the iterative solution of the estimating equations are described in detail along with the problem of setting confidence limits for model parameters. This approach is very useful in the analysis of quantal bioassays data.

## PERSPECTIVE - GESTION

Le problème de la mise au point d'une méthode de modélisation de la réaction d'un organisme à une toxicité chronique est traité dans cet ouvrage et illustré par l'étude de l'effet toxique du NaBr sur le processus de reproduction d'une population de <u>Daphnia</u> <u>magna</u>. On y présente un modèle général qui comprend à la fois les distributions binomiales négatives et la distribution de Poisson étant donné que certains cas spéciaux dépendent des valeurs d'un seul paramètre. On y décrit les étapes à suivre pour évaluer les paramètres du modèle et vérifier la validité de l'ajustement. On y fait en particulier la description détaillée de la solution itérative des équations estimatives, et on y présente le problème de l'établissement des limites de confiance des paramètres du modèle. Cette approche est très utile dans l'analyse des données binaires des bio-essais.

## INTRODUCTION

The negative binomial and the Poisson distributions are frequently used to model count data in many areas of biostatistics (Anscombe 1949; Bliss and Fisher 1953; El-Shaarawi et al. 1981; Maul et al. 1985). Regression analysis of count data following such distributions can be performed after transforming the crude data in order to approximately satisfy the requirements for the application of the standard regression methods. Several transformations which are based on the variance-mean relationship, have thus been suggested to achieve normality and homogeneity of the variances (Anscombe 1948). However, the previous approach for performing regression analysis is not always desirable (El-Shaarawi et al. 1987) since there is no evidence that a single transformation is sufficient to achieve all of the conditions needed for using regression . Therefore, it is preferable to perform regression analysis directly on the basis of the exact assumed probability distribution of the crude data. Consequently, Poisson and negative binomial regression models have been used by various authors (El-Shaarawi et al. 1987; Frome et al. 1973; Frome 1983; Engel 1984; Lawless 1987) for the analysis of count data.

The objectives of the paper are two-fold. The first is to give a clear and easily applicable description of the negative binomial and mixed Poisson regression analyses. Special attention is given to the discrimination between both models with emphasis on the fact that the latter is a special case of the former. Moreover, the maximum likelihood procedure is developed for estimating the unknown regression parameters, using a nonlinear (i.e. the log-linear) regression model. The second objective is to suggest the use of negative binomial and mixed Poisson regression models for analyzing quantal bioassays data, namely for determining a threshold dose level in chronic toxicity studies. The example used to illustrate the method is concerned with studying the toxic effect of NaBr on the reproduction process of a population of _Daphnia magna_. In particular, the aim of the assay is to estimate the dose level capable of inducing a prespecified relative inhibition on the reproduction of _Daphnia magna_ when the animals are exposed to increasing concentrations of the substance considered under controlled experimental conditions.

## STATISTICAL METHODS

### Negative Binomial and Mixed Poisson Regression Models

Let $R_1, R_2, \ldots, R_n$ be a set of n independent random variables where $R_i$ follows a negative binomial distribution with mean $m_i$ and dispersion parameter k, denoted by $R_i \sim NB(m_i, k)$.

We have

$$P\ (R_i = r_i) = \frac{(k + r_i - 1)!\ k^k\ m_i^{\ r_i}}{r_i!\ (k-1)!\ (k + m_i)^{k + r_i}} \qquad (1a)$$

for i = 1,2,...,n.

As k approaches infinity, the negative binomial converges in distribution to the Poisson which is:

$$P\ (R_i = r_i) = e^{-m_i}\ \frac{m_i^{\ r_i}}{r_i!} \qquad (1b)$$

Consider the model,

$$\ln m_i = \beta 1\ x_{i1} + \beta 2\ x_{i2} + \dots + \beta_p\ x_{ip} \qquad (2)$$

where $x_{i1},\dots,x_{ip}$ are the values of p explanatory variables $x_1,\dots,x_p$ which are associated with the random variable $R_i$ and $\beta_1,\dots,\beta_p$ are p unknown parameters.

## Estimation of the Parameters of the Model

If the response variable is a negative binomial then the likelihood function for $\beta_1$, $\beta_2,\dots,\beta_p$ is:

$$L(\beta 1,\dots,\beta_p,k) = \prod_{i=1}^{n} P(R_i = r_i)$$

$$= \prod_{i=1}^{n} \frac{(k+r_i-1)!}{r_i!\ (k-1)!}\ \frac{k^k\ e^{r_i(\beta_1 x_{i1}+\dots+\beta_p x_{ip})}}{(k + e^{\beta_1 x_{i1}+\dots+\beta_p x_{ip}})^{k+r_i}} \qquad (3)$$

The maximum likelihood estimates $\hat{\beta}_1,\ldots,\hat{\beta}_p$, $\hat{k}$ of $\beta_1,\ldots,\beta_p$ and k satisfy the equations.

$$\sum_{i=1}^{n} (k + r_i) \frac{x_{ij}m_i}{k + m_i} = \sum_{i=1}^{n} r_i x_{ij} \qquad (j = 1,\ldots,p) \qquad (4a)$$

$$\sum_{i=1}^{n} \sum_{t=1}^{r_i} \frac{1}{k-1+t} = \sum_{i=1}^{n} \ln\left(1 + \frac{m_i}{k}\right) + \sum_{i=1}^{n} \frac{r_i - m_i}{k + m_i} \qquad (5)$$

It should be noted that this system reduces to

$$\sum_{i=1}^{n} x_{ij}m_i = \sum_{i=1}^{n} r_i x_{ij} \qquad (j = 1,\ldots,p) \qquad (4b)$$

when the response variable $R_i \sim P\ (m_i)$ $(i = 1,\ldots,n)$.

Equations (4a and 5) or (4b) are nonlinear and their solution can be obtained by iteration. The simplest way to obtain $(\hat{\beta}_1,\ldots,\hat{\beta}_p,\hat{k})$ is to maximize the likelihood function with respect to the βj's (j = 1,...,p) for selected values of k (Lawless 1987).

This approach is similar to that suggested by Breslow (1984) wherein moment estimation is used for k instead of maximum likelihood. Lawless (1987) studied both the efficiency and the robustness of moment estimation for k relative to maximum likelihood; the former being more robust but less efficient than the latter when model (1a) is correct.

Equations (4a) or (4b) can be solved by using the Newton-Raphson method which requires the evaluation of the observed information matrix I. The p x p portion of I corresponding to $\beta_i \ldots, \beta_p$ has

$$i_{ls} = \sum_{i=1}^{n} \frac{k(k+r_i)x_{il}x_{is}m_i}{(k + m_i)^2} \qquad (6a)$$

respectively, as the element in the l<u>th</u> (r=1,2,...,p) row and sth (s=1,2,...,p) column when the negative binomial is assumed. This reduces to

$$i_{ls} = \sum_{i=1}^{n} x_{il}x_{is}m_i \qquad (6b)$$

when model (1b) is used.

Inferences about the $\beta_i$'s (i=1,2,...,p) can be made by noting that $\underline{\hat{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)$ has approximately for large n, a multivariate normal distribution with mean $\underline{\beta}$ and variance-covariance matrix $I^{-1}$. In some applications, confidence limits are required for the marginal mean m, of the response variable, given a set of values of the explanatory variables $x_1, \ldots, x_p$. It is easy to show that the confidence limits for m are:

$$\hat{m} \underset{+}{x} e^{u_{1-\alpha/2} \sqrt{\text{var } [\ln \hat{m}]}},$$

where $u_{1-\alpha/2}$ is an appropriate normal quantile and

$$\text{var}[\ln \hat{m}] = \sum_{j=1}^{p} x_j^2 \; I^{-1}(j,j) + 2 \sum_{i<j} X_i X_j \; I^{-1}(i,j)$$

where $I^{-1}(i,j)$ is the i,j element of the inverse of the matrix I.

A special case of interest arises when model (2) is expressed as ln m = f(x) as a polynomial of a single variable x, and it is required to estimate the value $x_q$ which may induce a prespecified relative variation (say a decrease by a rate 1-q) in the mean of the response variable, $m_s$ of x.

An estimate of $x_q$ is easily obtained as:

$$x_q = \hat{f}^{-1} \; (\ln q\hat{m}_s) \tag{7}$$

Further, confidence limits for $x_q$ at level 1-α are given as:

$$\hat{f}^{-1} \; (\ln q \; \hat{m}_s \pm u_{(1-\alpha/2)} \sqrt{\text{Var } [\ln \hat{m}_s]}) \tag{8}$$

where $\hat{m}_s$ is the estimate of the marginal mean given $x_s$.

## Significance Test for a Poisson Assumption

When presenting the iterative procedure, emphasis was laid on the necessity to have a test which will enable us to discriminate between model (1a) and model (1b). One convenient method for doing this is to use the standardized dispersion statistic developed by Dean and Lawless (1987), which is given as:

$$S = \frac{\sum_{i=1}^{n} [(R_i - \hat{m}_i)^2 - \bar{R}]}{(2 \sum_{i=1}^{n} \hat{m}_i^2)^{\frac{1}{2}}} \qquad (9)$$

Under the hypothesis that the $R_i$'s are independent Poisson random variables (i.e. $R_i \sim P(m_i)$), S has an asymptotic standard normal distribution. A departure from a Poisson regression model (the null hypothesis) can be shown by large positive values of S and will, subsequently, lead to a negative binomial assumption. In addition to this test, other possible ways for testing the Poisson hypothesis are given by Lawless (1987). However, these tests, (e.g. the likelihood ratio statistic for testing $k = +\infty$) (Chernoff 1954), appear more like a posteriori ways of testing the adequacy of the Poisson model since they require the maximum likelihood estimate, $\hat{k}$, for k.

## EXAMPLE AND DISCUSSION

The negative binomial regression method is illustrated hereafter by a numerical example referring to a chronic aquatic toxicity test.

The aim of the study is to determine the concentration of NaBr which will inhibit the reproductive capacity of a population of _Daphnia magna_ by 25 percent. The response variable $R_i$ is the number of young animals produced per adult over a period of 23 days, and the explanatory variable is the dose level of NaBr to which the organisms were exposed. Ten independent observations were made at each of five (i.e. 0, 3.0, 7.5, 19.0, and 47.0 mg/L) concentrations of NaBr. Figure 2 shows the data for the control and the test solutions.

A log-linear regression model,

$$\ln m_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 \tag{10}$$

was employed to describe the dependence of $\ln m_i$, the mean number of young on the concentration x of NaBr. Model (10) has been fitted to the data under each of the following three assumptions:

a) $\ln R_i \sim N(\ln m_i, \sigma_{ln})$

b) $R_i \sim P(m_i)$

c) $R_i \sim NB(m_i, k)$

where both of the parameters $\sigma_{ln}$ and k are assumed to be unknown.

The situations corresponding to a), b) and c) will be referred to as: standard log-linear, Poisson and negative binomial regression models, respectively.

Table 1 presents the maximum likelihood estimates for the regression parameters as given in model (10), and their standard errors which have been obtained under the three distributional assumptions a), b) and c). Estimates and confidence limits for the concentration corresponding to a 25% reduction in reproductivity are also given in the table. The three assumptions resulted in similar estimates for the regression parameters and the $x_{0.75}$ threshold. On the other hand, substantial discrepancies appear between the standard errors of the estimates for the regression parameters: the lowest (resp. highest) values corresponding to the Poisson, and the highest corresponding to the standard log-linear model.

Furthermore, Bartlett's test was used for testing the assumption of homogeneity of variances among the concentration levels. The observed variances among the five concentration levels were statistically significantly different ($P < 0.01$). This indicates that analysis a) is inappropriate in making inferences about the parameters of the model. A similar outcome was obtained for analysis b) since the observed value for S, as given by formula (9), was 4.79 ($P < 0.01$), which also provides strong evidence against the Poisson assumption. In the opposite case, the likelihood ratio test, which was performed to test the equality of k's yielded a non-significant value ($P > 0.10$).

Consequently, inferences about the parameters of the model or the mean can easily be performed under the less stringent assumption of analysis c). The estimated variance-covariance matrix of $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ is,

$$\begin{bmatrix} 76888.2\ 10^{-8} & -8367.91\ 10^{-8} & 146.709\ 10^{-8} \\ & 1771.25\ 10^{-8} & -35.7684\ 10^{-8} \\ & & 0.792986\ 10^{-8} \end{bmatrix}$$

which could be used for making inferences about the parameters. For instance, testing the hypothesis $H_0:\beta_1 > 0$ is of special interest, with $\beta_1$ representing a non-monotonic concentration trend in the concentration-response relationship. In certain cases, the test substance may slightly stimulate reproduction at very low concentrations before causing toxicity at higher levels (Capizzi _et al_., 1985). In the assay in question, such a phenomenon could not be discarded ($P > 0.01$). Further, ln m is plotted in Figure 1 as a function of the concentration x. The estimated threshold, $\hat{x}_{0.75}$, which was in fact the major objective of the example examined, and a confidence interval at level 0.95 for $x_{0.75}$ are also indicated in Figure 1.

## REFERENCES

Anscombe, F.J. 1948. The transformation of Poisson, binomial and negative binomial data. Biometrika, 38:246-254.

Anscombe, F.J. 1949. The analysis of insect counts based on the negative binomial distribution. Biometrics, 5:165-173.

Bliss, C.I. and R.A. Fisher. 1953. Fitting the negative binomial distribution to biological data. Biometrics, 9:176-200.

Breslow, N. 1984. Extra-Poisson variation in log-linear models. Appl. Statist., 33:38-44.

Capizzi, T., Oppenheimer, L., Mehta, H., Naimie, H., and J.L. Fair. 1985. Statistical considerations in the evaluation of chronic aquatic toxicity studies. Environ. Sci. Technol. 19:35-43.

Chernoff, H. 1954. On the distribution of the likelihood ratio. Ann. Math. Statist., 25:573-578.

Dean, C. and Lawless, J.F. 1987. Testing for overdispersion in Poisson regression models, unpublished, mentioned by Lawless (1987).

El-Shaarawi, A.H., Esterby, S.R. and Dutka, B.J. 1981. Bacterial density in water determined by Poisson or negative binomial distributions. Appl. Environ. Microbiol., 41:107-116.

El-Shaarawi, A.H., Maul, A. and Block, J.C. 1987. Application of Poisson regression to the analysis of bacteriological data. Water Poll. Res. J. Canada, 22:298-307.

Engel, J. 1984. Models for response data showing extra-Poisson variation. Statist. Neerlandica, 38:159-167.

Frome, E.L. 1983. The analysis of rates using Poisson regression models. Biometrics, 39:665-674.

Frome, E.L., Kutner, M.H. and Beauchamp, J.J. 1973. Regression analysis of Poisson-distributed data. J. Am. Statist. Ass., 68:935-940.

Lawless, J.F. 1987. Negative binomial and mixed Poisson regression. Can. J. Statist., 15:209-225.

Maul, A., El-Shaarawi, A.H. and Block, J.C. 1985. Heterotrophic bacteria in water distribution systems. I. Spatial and temporal variation. II. Sampling design for monitoring. Sci. Total Environ., 44:201-224.

Williams, D.A. 1972. The comparison of several dose levels with a zero dose control. Biometrics, 28:519-531.

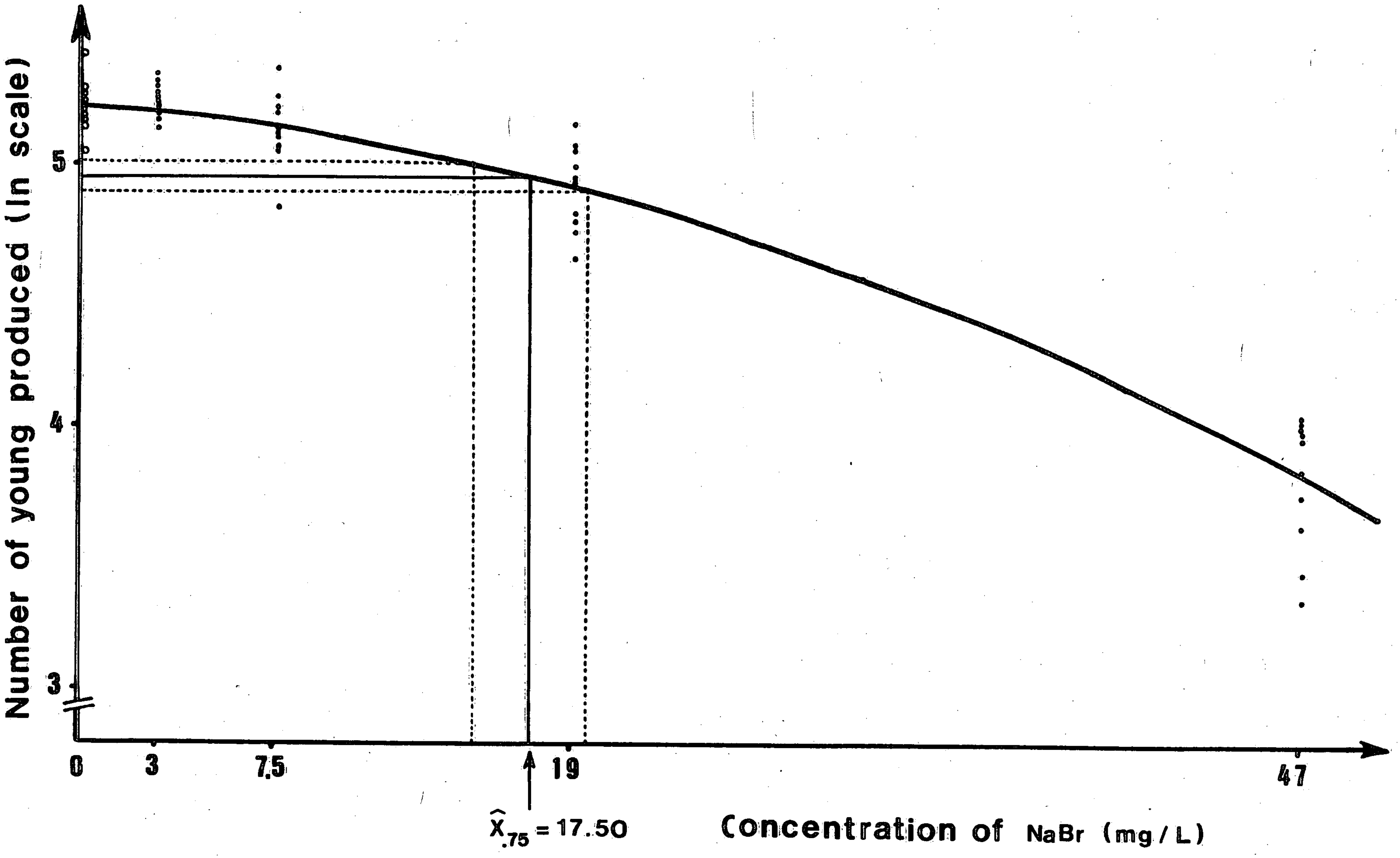
Table 1. Estimates of the regression parameters and the 75% threshold dose level in the concentration-response model.

| | Approach | | |
|---|---|---|---|
| Dispersion parameter | $\hat{\sigma}_{ln} = 0.1503$* | $k = \infty$ | $\hat{k} = 130.05$ |
| | Regression parameters<br>Estimates (standard errors) | | |
| $\beta_0$ | 5.249706 (0.037489) | 5.255168 (0.017816) | 5.255337 (0.027729) |
| $\beta_1$ | -0.007982 (0.005438) | -0.008050 (0.002852) | -0.008106 (0.004209) |
| $\beta_2$ | -0.000489 (0.000109) | -0.000477 (0.000064) | -0.000476 (0.000089) |
| $\hat{x}_{0.75}$ | 17.43 | 17.53 | 17.50 |
| Confidence interval at level 0.95 | (14.30; 20.21) | (16.08; 18.90) | (15.21; 19.61) |

*The value corresponds to a coefficient of variation of 15%.

## FIGURES

Figure 1. Observed (points) and fitted (line) ln mean number of young produced. Estimated value, $\hat{x}_{0.75}$, and confidence interval at level 0.95 for $x_{0.75}$.


Number of young produced (ln scale)
5
4
3
0
3
7.5
19
47
$\hat{X}_{.75} = 17.50$
Concentration of NaBr (mg / L)