Anthrop. Közl. *33;* 113–122. (1991)

THE METHOD OF PROBITS IN AUXOLOGY

H. Danker-Hopfe[1] and *W. Wosniok*[2]

[1]Department of Human Biology and [2]Department of Statistics, University of Bremen, Bremen, Germany

Abstract: *Since probit analysis, the classical method of analysis of dose-response relationships, in 1950 has first been applied to status quo data on menarche it has increasingly been used in auxology. This method can profitably be used whenever mean ages of occurrence of qualitative events in the development of children and youths have to be estimated from status quo data. In spite of its widespread use details concerning computational techniques with their implications on the reliability of the estimates, however, are hardly ever mentioned. The main focus of the present paper will thus be on the different approaches to estimate the parameter vector* (μ,σ^2)*, which range from graphical solution techniques to the maximum-likelihood principle. The merits and drawbacks of the different methods will be demonstrated using a set of empirical data on menarche. Finally the importance of testing the underlying assumption concerning the distribution of the time variable is stressed.*

Key words: *Status-quo Method; Probit Analysis; Menarche.*

Introduction

Probit analysis is the classical method of analysis of dose-response relationships. Since Wilson and Sutherland in 1950 first applied this method to status quo data on

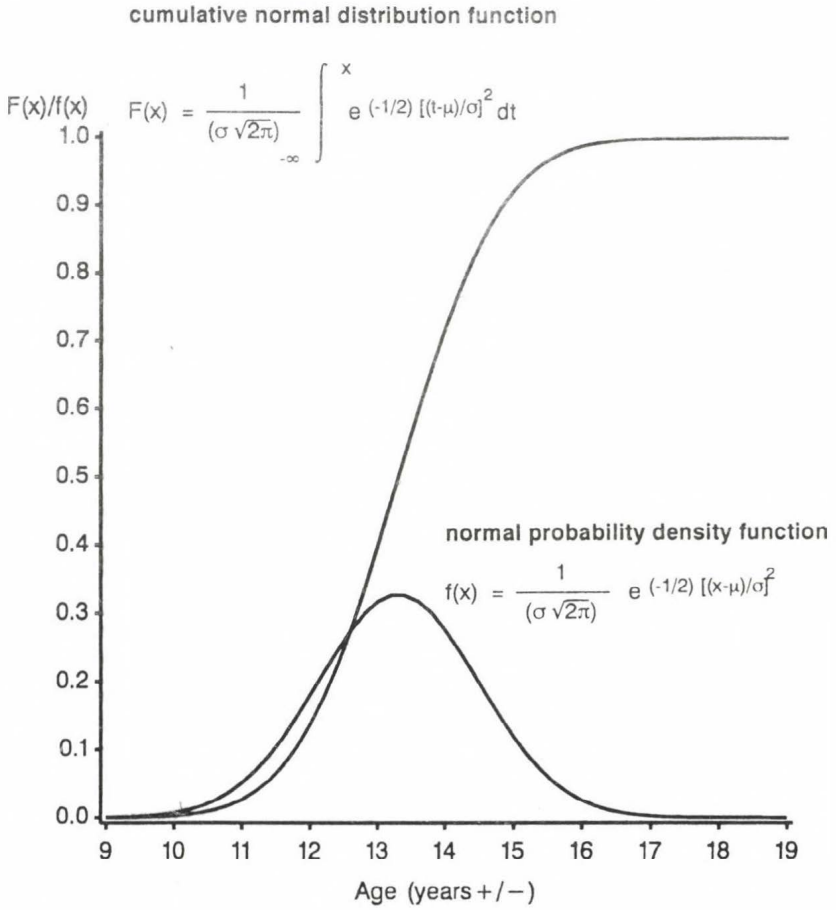


Fig. 1: Relation between age (independent variable) and probability of trait occurrence by a normal distribution

menarche it has increasingly been used in auxology. The method of probits can profitably be used whenever descriptive statistics of the distribution of qualitative events in the development of children and youths, like occurrence of menarche, occurrence of certain stages of the development of secondary sex characteristics or eruption of primary or secondary teeths have to be estimated from status quo data. In this context the dose or independent variable is represented by the age variable, the response or dependent variable is a dichotomic variable (yes or no) which indicates whether the trait under consideration has already occurred or not.

An important basic assumption in probit analysis is that the trait under consideration represents a normally distributed random variable – an assumption which, as far as the distribution of qualitative events in the development of children is concerned, might not always be the best one. The normal probability density function can be represented by the bell-shaped graph shown in *Figure 1*.

The underlying function depends on the mean μ which determines the location of the distribution and the standard deviation σ, which determines its shape. The function f (x) – which depends on age – represents the probability with which the trait occurs at a given age. However, when we are dealing with status quo data on the occurrence of a specific trait, we do not have informations concerning the probability that the trait occurs *at a given age* but, we do have informations about the probability that the trait has already occurred *prior to or at this age;* this distribution is represented by the cumulative normal distribution function (see *Figure 1*). That is we do have observed relative frequencies of individuals showing a response by age of the form shown in *Figure 2*.

The problem to be solved is to estimate the parameter vector (μ,σ^2) which produces the best fit to the data. There are various statistical approaches to estimate this parameter vector which are all summarized under the term probit analysis. In the following a brief summary of those methods most commonly subsumed under the term probit analysis is given. Instead of using a lot of statistical formulae, an empirical data set is chosen to illustrate the different approaches which range from simple graphical solutions over regression techniques to the maximum likelihood principle.

The data set used for illustration consists of status quo data on menarche sampled from 2796 girls aged between 10.01 and 18.32 years during the *First Bremerhaven Growth Survey* carried out in 1979/80 (see also Osterschlt and Danker-Hopfe in this volume). The distribution of pre- and postmenarcheal girls by half year age groups – with the age displayed being the midpoint of the underlying interval – is shown in *Figure 3*. The youngest postmenarcheal girl in the sample was 10.77 years old, the oldest premenarcheal on the other hand was 17.24 years which is a quite exceptional event (see Ostersehlt and Danker-Hopfe 1991), the next oldest premenarcheal girl was 16.22 years old. We will refer to this exceptional observation again a little later. Since sample size is comparatively large further results are based on a subdivision of the age variable by 0.2 years.

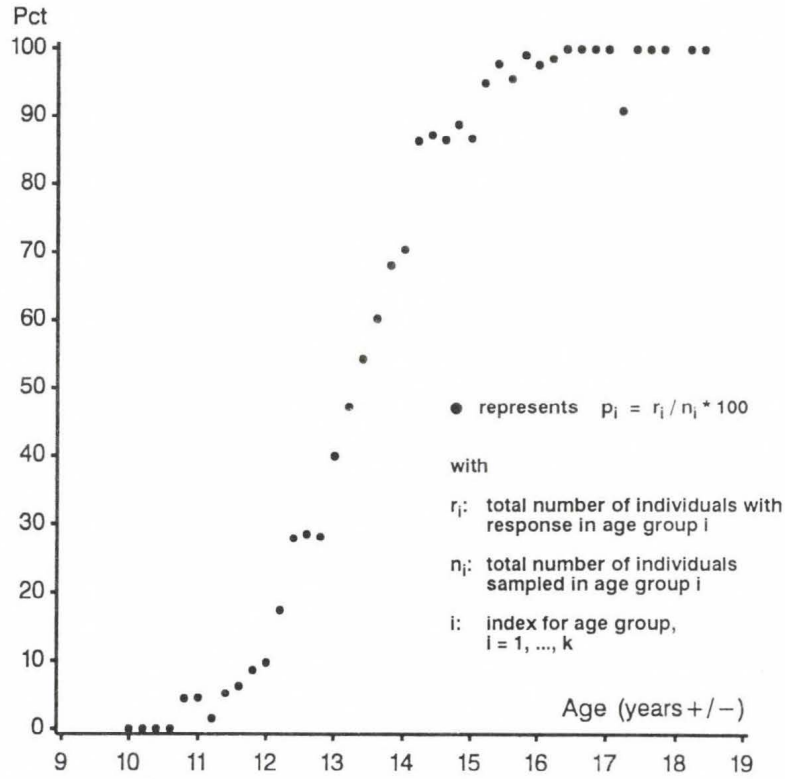


Fig. 2: Empirical distribution of postmenarcheal girls – age grouped in 0.2 year intervals

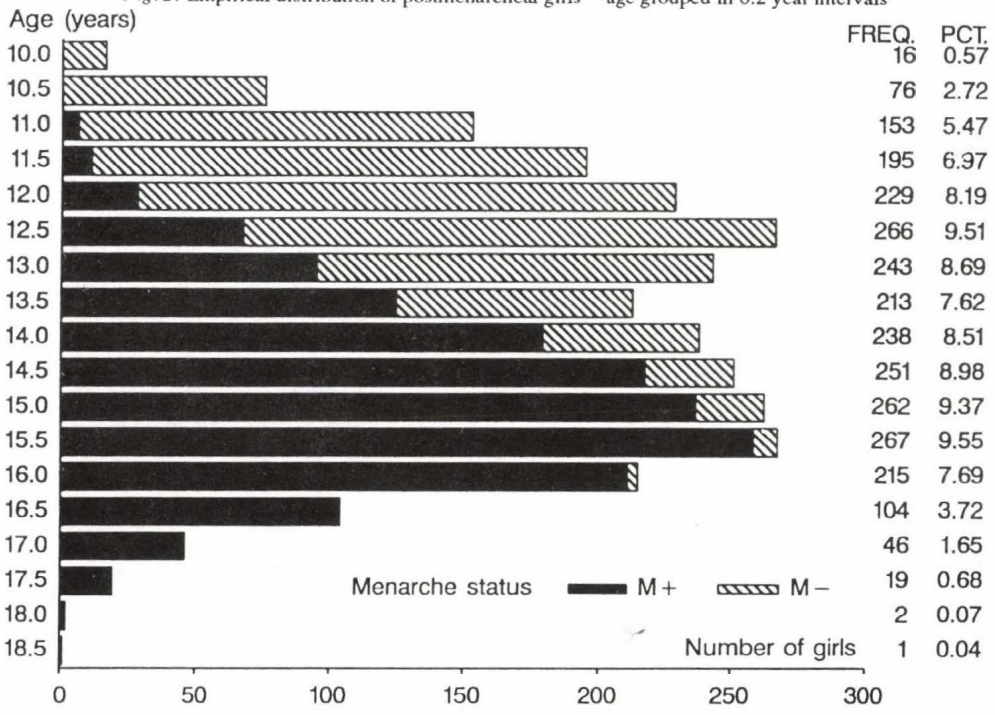


Fig. 3: Age structure of the sample

Transformation

The basic step when estimating the parameter vector (μ,σ^2) is the probit transformation. Probit transformation transforms the sigmoid curve (1) into a linear one (2), where the vertical axis represents the transformed probability of trait occurrence and where the horizontal axis – as before – represents the age variable. The transformed dependent variable is the so-called *probit* (probability unit). For a detailed discussion of the single steps of axis transformation as well as of probit analysis in general see e.g. Finney (1971), Weber (1980), Unkelbach and Wolf (1985) and McCullagh and Nelder (1989).

$$F(x) = \frac{1}{(\sigma\sqrt{2\pi})} \int_{-\infty}^{x} e^{(-1/2)\,[(t-\mu)/\sigma]^2}\, dt \qquad (1)$$

$$y = a + bx \qquad (2)$$

The functional relationships between the parameter vector (μ,σ^2) and the parameter vector (a, b) of the linear equation defining the probits are demonstrated by equations (3) and (4), where the '5' is a historical relict, introduced with the intention to avoid computations with negative numbers.

$$\sigma = 1/b \qquad (3)$$

$$\mu = (5-a)/b \qquad (4)$$

Estimation Techniques and Results

The most simple approaches to estimate the parameters a and b, and hence μ and σ^2, are graphical ones. Graphical approximations can be carried out on so-called probability paper. As shown in *Figure 4*, the observed frequencies of postmenarcheal girls by age may simply be plotted and a "regression" line may be fitted to these data by eye. A rough estimate of the mean can be obtained by raising a horizontal line from 50% until it intersects with the eye fitted "regression" line and then reading the age on the linear age scale. Since it is known that for normally distributed traits approximately 68% of the observations lie within the interval $\mu \pm \sigma$ one might get an estimate of σ by raising horizontal lines from approximately 84% and 16% until they intersect with the eye fitted "regression" line and then reading the corresponding ages for $\mu + \sigma$ and $\mu - \sigma$ at the age scale. One will get estimates of σ by simple subtraction. It should be noted, however, that the results obtained are highly subjective since they entirely depend on the "regression" line fitted by eye. If probability paper is not available one might directly transform frequencies into probits, using a table which is to be found in every good

statistical textbook or in a collection of statistical tables (Fisher and Yates, 1963, Finney, 1971). Then a graphical estimation can be performed using ordinary millimeterpaper.

Since it is expected that the plot of probits of postmenarcheal girls by age can be approximated by a straight line the next step of complexity or towards more accuracy would be to apply regression techniques (see *Figure 5*). There are 0.5n!/[2(2–n)!] straight lines which can be fitted to the scatter of probits. Every two points define uniquely a regression line. Every pair of coordinates, that is ages and their corresponding probits, can be chosen to get estimates of the parameters a and b. But again the problem is that the results entirely depend on the subjectivity of selecting the points.

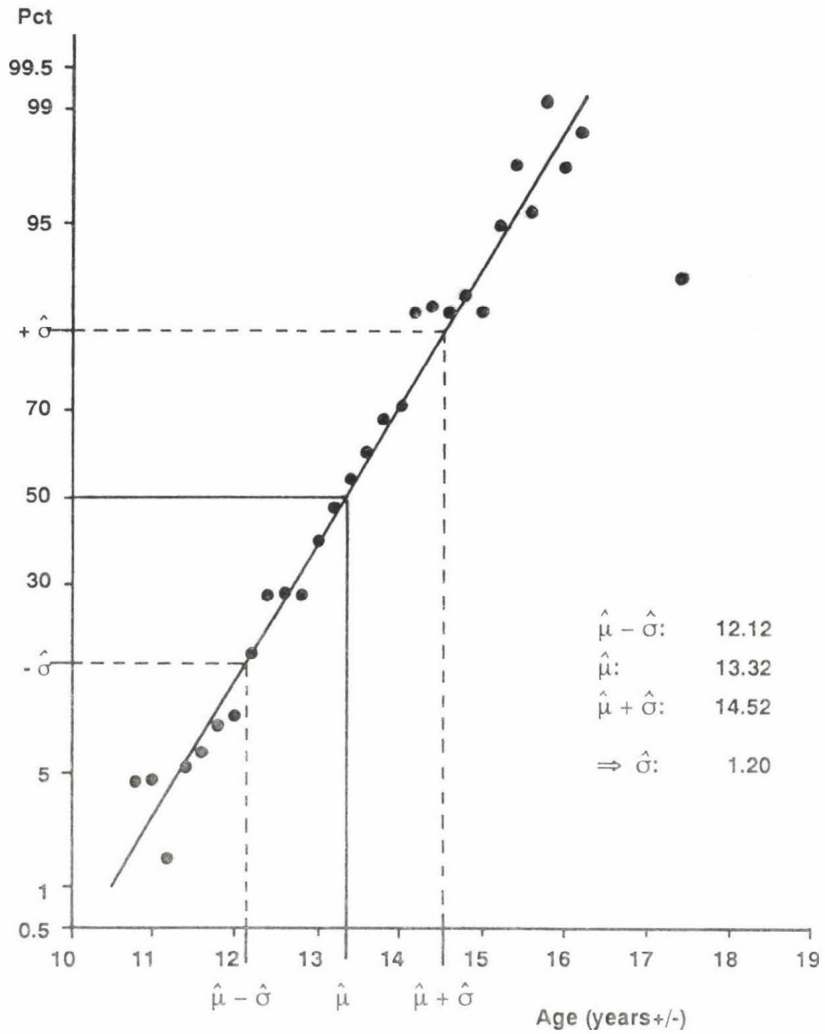
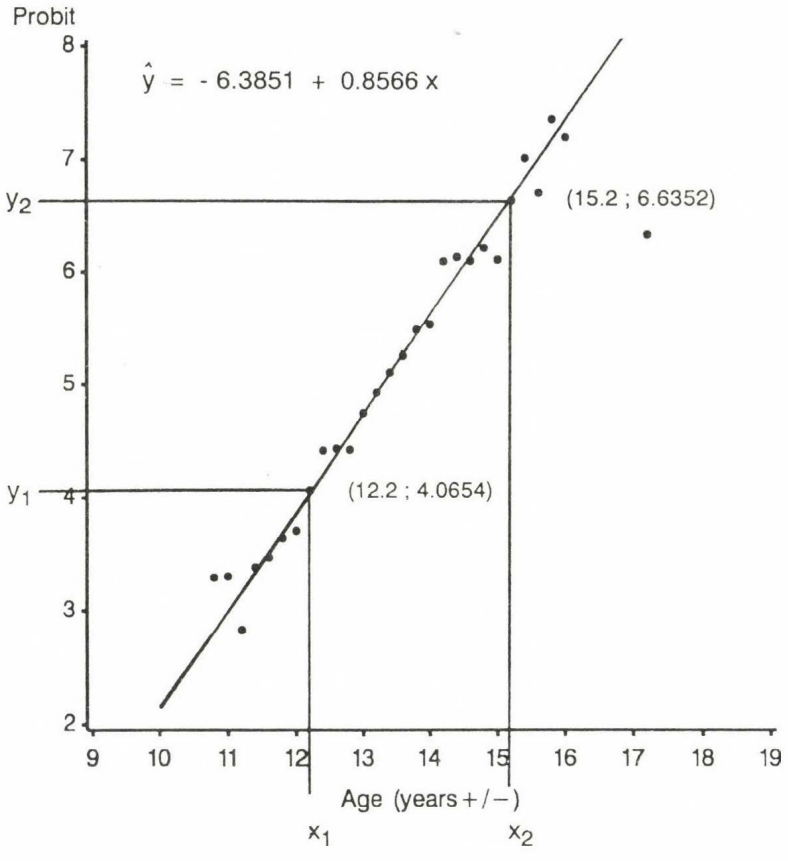


Fig. 4: Estimation of μ and σ by fitting a straight line by eye



estimation of the slope: $b = (y_2 - y_1) / (x_2 - x_1)$

$\Rightarrow$ estimate of σ: $\hat{\sigma} = 1 / b = 1.17$

estimation of the intercept: $a = y_1 - bx_1$ or $a = y_2 - bx_2$

$\Rightarrow$ estimate of μ: $\hat{\mu} = (5 - a) / b = 13.29$

Fig. 5: Estimation of μ and σ by fitting a straight line through two selected points

This approach – estimation by selecting two points – produces a lot of numerically different estimates of a and b, but how to find the ones which fit the empirical data best? The usual criterion for the best fit is that the sum of squares of the differences between observed and expected probits is minimal, which leads to the *least-squares method* of estimation where a and b are estimated according to equations (5) and (6):

$$b = \frac{\Sigma x_i y_i - \bar{y} \Sigma x_i}{\Sigma x_i^2 - \bar{x} \Sigma x_i} \qquad \begin{matrix} (\bar{y} = \Sigma y_i/k) \\ (\bar{x} = \Sigma x_i/k) \end{matrix} \tag{5}$$

$$a = \bar{y} - b\bar{x}. \tag{6}$$

One of the main drawbacks using simple least-squares regression techniques in this context, however, is that the method is based on the assumption that the variances of the dependent variable (y_i) are equal for all values of the independent variable (x_i), an assumption which is not met by the underlying data. To avoid the error introduced by unequal variances a *weighted regression analysis* should be performed with the reciprocal variances being the weights. This leads to a problem of nonlinear optimization, which can only be solved by iterative estimation techniques.

A method which asymptotically leads to the same estimates is the maximum-likelihood principle. *Maximum-Likelihood estimation* is a universal technique to estimate unknown parameters from empirical (observed) data if – and this is essential – the model of distribution is known. To understand the principle of this estimation technique some notions and basic ideas will briefly be repeated. Let

$$p_i = r_i / n_i \tag{7}$$

be the frequencies with which a specific response occurs at age x_i. The number r_i of responses given n_i trials follows a binomial distribution, that means the probability that responses have already occurred in r_i of n_i individuals is defined by equation (8).

$$p_i = \binom{n_i}{r_i} p_i^{ri} (1 - p_i)^{n_i - r_i} \tag{8}$$

Since it can be assumed that the observations in the different age groups are independent and identically distributed the probability of the result of the whole experiment is the product of the single probabilities:

$$L = \prod_{i=1}^{k} p_i = \prod_{i=1}^{k} \binom{n_i}{r_i} p_i^{ri} (1 - p_i)^{n_i - r_i} \tag{9}$$

This equation is called the likelihood function, where L is considered to be a function of p_i given r_i and n_i.

Assuming that the observed result is most likely, the unknown parameters will be estimated according to those values which make the observed or empirical data most likely, again leading to a problem of optimization. We have to look for the maximum of the function, which can be determined by estimating the zeros of the first partial derivatives. To estimate p in this context p and individual age have to be related, where p_i is a function of age x_i. From the literature it is known that most of the qualitative events dealt with in auxology are normally distributed traits, so it is reasonable to

assume that p_i can be obtained from the cumulative normal distribution function mentioned earlier. But, this is the place where different distribution models (e.g. logistic distribution, Gompertz distribution etc.) can be used.

Estimation of μ and σ in this relationship which means to maximize L by a choice of μ and σ, again requires iterative techniques and the use of computers is recommended. There are several commercial statistical software packages which include probit analysis by the maximum-likelihood principle. We used SAS and got the maximum likelihood estimates shown in *Table 1* together with the results obtained by application of the other methods.

Table 1. **Results of probit analysis by different estimation techniques and results of the likelihood ratio test of goodness-of-fit**

Estimation technique	μ [years]	σ [years]	χ^2 (40)	p
Graphical approach on probability paper	13.32	1.20	34.50	0.7155
Regression analysis: regression line defined by two coordinates	13.29	1.17	34.80	0.7030
Regression analysis unweighted least-squares estimates	13.35	1.33	47.14	0.2037
Regression analysis weighted least-squares estimates	13.32	1.18	34.68	0.7080
Maximum-likelihood estimates	13.30	1.19	34.24	0.7264

Except for the estimates obtained by the unweighted least-squares regression method the means exhibit comparatively little variation. This is primarily due to the large sample size and detailed subdivision of the age scale. If sample sizes are small and age classification is rough the results will certainly be more diverging. Differences between the estimates of the standard deviations obtained by the different methods seem to be more pronounced.

Discussion

The quality of the different estimates must be discussed. There are at least two criteria which can be used for this purpose. The first assessment of the estimates is based on their standard errors. Since these, however, are not available for the first two methods the second criterion, which can be used with all the methods of estimation mentioned, will be used in this context. This is a test of goodness-of-fit of the empirical data to the distribution defined by the estimated parameters. There are two main possibilities: the first is to use Pearson's ordinary χ^2-test (10):

$$\chi^2 = \sum_{i=1}^{k} (N_{oi} - N_{ei})^2 / N_{ei} \qquad (10)$$

with:
N_{oi} : observed number of postmenarcheal girls in the ith age group
N_{ei} : expected number of postmenarcheal girls in the ith age group, $N_{ei} = n_i * \hat{p}_i$;
k : number of age groups.

and the second is use of the likelihood-ratio test (11):

$$\chi^2 = -2 \log 1 \sum_{i=1}^{k} n_i \left(p_i \log \frac{p_i}{p_i} + (1 - p_i) \log \frac{1 - p_i}{1 - p_i}\right) \qquad (11)$$

with:
k : number of age groups
n_i : sample size in the ith age group
p_i : observed relative frequency of postmenarcheal girls in the ith age group
p_i : expected relative frequency of postmenarcheal girls in the ith age group – derived from the model.

Application of both tests to the data yielded the following results:

Pearson's χ^2 $\chi^2_{(40)} = 189.53$; $p < 0.001$;
likelihood ratio test: $\chi^2_{(40)} = 34.24$; $p = 0.7264$.

Focussing on the probability of error for rejection of the hypothesis – that the data was generated by a normal distribution – it is seen that according to Pearson's χ^2-test the fit seems to be very bad – leading to rejection of the model – while the likelihood ratio test reveals a comparatively good fit. How can these contrasting results be explained? Both tests are based on the same empirical data and the same ML parameter estimates. Here the original data of pre- and postmenarcheal girls have to be recalled. In our sample there was one girl aged 17.24 years who had not yet experienced menarche. This single girl leads to the extremely high Pearson's χ^2-statistic. Due to an expectation which is close to zero, age group 17.2 contributes a value of 164.15 to the total χ^2, that is more than 86% (86.61%), while the contribution of this age group to the likelihood ratio χ^2 is only 8.33, which amounts to 24.33%.

The data thus demonstrate that as far as rare events are included in a data set, ordinary Pearson's χ^2-test might be quite misleading. To avoid biases by rare events the likelihood ratio test of goodness-of-fit is to be preferred. The results of the likelihood ratio test are also summarized in *Table 1*. The error probabilities indicate that simple least squares regression leads to estimates which do not fit the empirical data very well. On the other hand ML-estimation of the parameters leads to the best fit as indicated by the highest error probability which is almost 73%.

Summarizing the results, ML estimation or weighted regression analysis is recommended because they are defined on adequate model assumptions and allow assessment of precision. Furthermore they are free of subjective effects which is not true for graphical and related procedures. And finally a goodness-of-fit test should be performed, preferably the likelihood ratio test because of its robustness against effects of rare events, to get an idea about the fit of the emprical data to the model assumptions.

*

Paper presented at the Fifth International Symposium of Human Biology, Keszthely, Hungary, June 1991; Received 17 July, 1991.

References

Finney DJ 3rd ed. (1971) *Probit Analysis.* — Cambridge University Press, Cambridge.
Fisher RA, Yates F 6th ed. (1963) *Statistical Tables.* — Longman, Edinburgh.
McCullagh P, Nelder JA 2nd ed. (1989) *Generalized Linear Models.* — Chapman & Hall, London.
Ostersehlt D, Danker-Hopfe H (1991) Changes in age at menarche in Germany: Evidence for a continuing decline. — *American Journal of Human Biology, 3;* 647—654.
Ostersehlt D, Danker-Hopfe H (1992) Preliminary results of a study on changes in growth of girls from Bremerhaven. — *Anthrop. Közl., 33;* 147—154.
Unkelbach HD, Wolf T (1985) *Qualitative Dosis-Wirkungs-Analysen.* — Gustav Fischer Verlag, Stuttgart.
Weber E 8th ed. (1980) *Grundriß der biologischen Statistik.* — Gustav Fischer Verlag, Stuttgart.

Mailing address: Dr Heidi Danker-Hopfe
Department of Human Biology
University of Bremen
W–2800 Bremen 33
Germany