A MATLAB-Based Software Tool for Changepoint Detection and Nonlinear Regression in Dose-Response Relationships

Stefan Wagenpfeil[1], Uwe Treiber[2], and Antonie Lehmer[2]

[1] Institut für Medizinische Statistik und Epidemiologie der Technischen Universität München, Klinikum rechts der Isar, Ismaninger Str. 22, D-81675 München, Germany
stefan.wagenpfeil@imse.med.tu-muenchen.de
http://www.imse.med.tu-muenchen.de/persons/wagenpfeil/index.html
[2] Urologische Klinik und Poliklinik der Technischen Universität München, Klinikum rechts der Isar, Ismaninger Str. 22,
D-81675 München, Germany

Abstract. Resulting from a clinical consulting case in urology we developed a software tool for determining nonlinear dose-response relationships. Unlike most existing statistical software packages, we directly compute and display analytical pointwise 95% confidence intervals for the prediction result. Furthermore, user-defined changepoints with 95% confidence interval can be calculated in order to estimate the dosage for a 50% response rate, for instance. This is necessary to compare the effect of different retinoids, tumor cell lines, etc. In this way we supplement the clinical software-equipment in our laboratory and encourage the evaluation of dose-response data. The numerical and computational problems arising with nonlinear regression, 4-parameter logistic as well as log-logit modelling and the respective confidence intervals are addressed in particular. Analysis of real data and an example data set demonstrate the approach. A demo version of the software tool can be downloaded from the first author's homepage.

1 Introduction

In the laboratory department of the Urologische Klinik und Poliklinik der Technischen Universität München, Klinikum rechts der Isar, the connection between doses of certain retinoids and response rates such as "cell kill" and "invasivity" of tumor cells are examined in vitro. The concentration is measured in μM, starting with small values and increasing accordingly. The response rates range from 0 (= 0% response) to 1 (= 100% response). For the target variable "invasivity" 0 equals to no invasivity of the tumor cells and 1 equals to 100% or full invasivity. For "cell kill" the response value 0 equals to no tumor cells left, i.e. 100% cell kill, whereas 1 equals to nothing happened or 100% cell survival, i.e. 0% cell kill. Fig. 1 displays an example file for a particular retinoid and tumor cell line. For low concentrations we have a limit of 100% response. Increasing concentrations

R.W. Brause and E. Hanisch (Eds.): ISMDA 2000, LNCS 1933, pp. 190–197, 2000.
© Springer-Verlag Berlin Heidelberg 2000

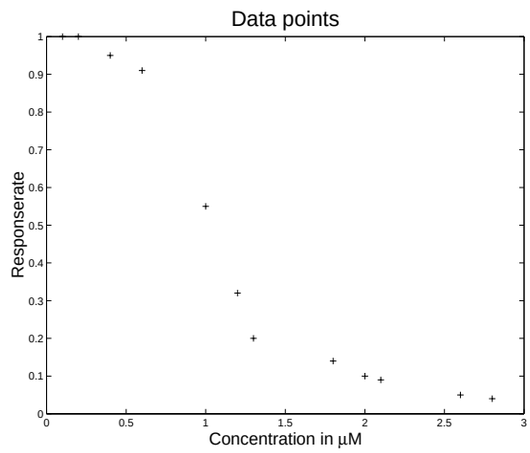

Fig. 1. Dose-response data (+) exemplifying laboratory measurements

yield decreasing response rates with a lower bound of 0% indicating a monotonic, nonlinearly decreasing relationship.

In general, ten or twelve dose-response combinations are determined in one trial. The response rates are obtained either by ELISA-reading for "cell kill" or counted manually within a reference field for "invasivity" of tumor cells. One aim is to estimate the underlying "true" dose-response relationship, which is a classical titration problem discussed in more detail e.g. in [1] and [2]. A common approach for exploring connections of the present kind with continuous doses and scaled binomial outcomes is the log-logit model. The 4-parameter logistic regression model is a generalized version where uncertainty about lower and upper bounds of response rates can be taken into account additionally. Moreover, confidence intervals are essential in order to compare retinoids and tumor cell lines from different trials. However, the software of the ELISA-reader described in [3] as well as most existing software tools for nonlinear regression display confidence intervals in their output only for regression parameter estimates, but not, although implicitly computed, for estimated response curves or predicted values, respectively. Some of the packages like standard SPSS (Statistical Package for Social Sciences) just do it on the grid of observed doses and not on the whole range of measured concentrations. Fig. 2 gives the estimated response curve of the ELISA-reader software for our example data in Fig.1.

Therefore, the task resulting from the laboratory consulting case was to provide a software tool for displaying the estimated response curve and the respective 95% confidence intervals simultaneously. Furthermore, a changepoint concentration should be detected indicating a significant decrease in the response rate. This we call our significant changepoint. The speed of decrease and the deceleration of the response curve at the changepoint are of interest, too. According to half-life periods of radioactive elements, concentrations with y% response rate

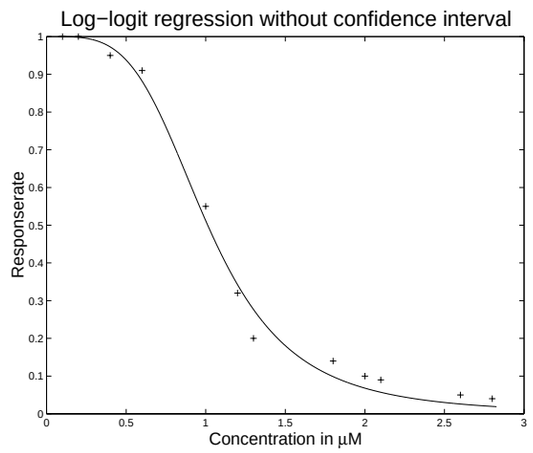

Fig. 2. Dose-response data (+) and estimated response curve (—) from log-logit model

also have to be estimated. We calculated these kind of changepoints for $y = 10$, 30, 50 and 70. The concentration with an estimated 50% response rate is called half-life concentration. Finally, we have to compute the 95% confidence intervals for the significant changepoint and the y% concentrations as measures of the accuracy of our estimates. Furthermore, the confidence intervals are essential in order to compare different tumor cell lines, retinoids, etc. for a series of trials.

The methods and routines applied are described in the next section. Results of the calculations for our example data file as well as a real data set are given in Sect. 3. Sect. 4 concludes.

2 Methods for Data Analysis

The log-logit model is a common way of analyzing monotonic, nonlinear dose-response relationships as given in Fig.1. It is a special case of the 4-parameter logistic model defined by

$$y = \frac{a-d}{1+(\frac{x}{c})^b} + d \,, \tag{1}$$

where y is the estimated, predicted response rate at concentration x. As we have a 100% response (e.g. invasivity of tumor cells) for concentration $x = 0$ and no response for sufficiently high values of x in our case, a natural choice for a and d in (1) is $a = 1$ and $d = 0$. This yields the log-logit model underlying the statistical analysis of our example data. Note that we only have to estimates two parameters in this case: b describing how rapidly the curve makes its transition from the asymptotes to the center of the curve and c, the half-life concentration.

If the lower and upper bound of the resopnse rates cannot be fixed in advance as is common in real data problems, the use of the full 4-parameter logistic regression mode (1) is a natural choice.

In both cases estimation can be carried out by the method of least squares which is equivalent to likelihood estimation for the assumption of Gaussian errors, cf. [4], subsection 6.6.1. Because the regression function in (1) is nonlinear, a solution in closed form does not exist and iterative algorithms have to be applied, cf. [5] for details. We used the MATLAB procedure *nlinfit* described in [9], where the Gauss-Newton optimization method is implemented. A general description of nonlinear regression algorithms can be found in [6]. The most important methods are implemented in the optimization toolbox of MATLAB [7].

Special attention has to be drawn to badly-scaled data yielding singular matrices to be inverted within the nonlinear programming procedure. An indication of that problem are extremely high condition numbers as discussed in [8], subsection 4.5. Remedial measures can be taken by strictly monotonic increasing transformations of the concentration values. The log-function is a commonly used transformation in order to try come to a well-scaled problem. On the whole, the success of the nonlinear regression approach depends heavily on the scaling of the data. This is true for all available nonlinear regression tools.

To compute analytical 95% confidence intervals for the estimated response rates there are two ways, in principle. One is to impute the lower and upper confidence limit of the parameter estimates supplied by most of the usual software packages into formula (1). However, this is a heuristic approach and yields unsymmetric confidence intervals, in general. Moreover, a possible correlation between the likelihood estimates of b and c is not considered. Therefore we applied the δ-method described for example in [4], chapter 2, in order to get symmetric, pointwise confidence intervals. The necessary inverse of Fisher's information matrix is calculated from the score function stored in the Jacobian of the MATLAB function *nlinfit*, cf. [9]. A numerical stable implementation can be cribbed from the MATLAB procedures *nlparci* or *nlpredci* contained in the statistical toolbox cf. [9]. The relevant first partial derivatives of the log-logit function in (1) with respect to the regression parameters b and c can easily be obtained by symbolic differentiation in the symbolic math toolbox of MATLAB, cf. [10]. The toolbox is based on a MAPLE V kernel. An analytic formulation of the derivatives is given in the demo version of our software tool, procedures *geschw* and *beschl*. In the same way the first partial derivatives of (1) with respect to a and d can be calculated for the 4-parameter logistic model.

Exact confidence intervals in nonlinear regression as developed in [11] are not in common use. Bootstrap confidence intervals established in [12] would be a further choice if the analytical approach does not converge. However, in the setting of nonlinear regression they are extremely demanding with respect to computation time and power.

For computation of changepoints we have to devide between the y% concentrations and the significant changepoint. The y% concentration is easily obtained via the inverse formula of (1) given by

$$x = c\left(\frac{a-y}{y-d}\right)^{1/b} . \tag{2}$$

Again we have $a = 1$ and $d = 0$ for log-logit regression.

The significant changepoint is defined as the concentration value x where the upper limit of the regression curve confidence interval is instantly below a response rate of 1. This is the starting point for a significant reduction of the response rate. The respective confidence intervals for both types of changepoints are calculated by a grid search on the confidence limits of the response curve ensuring at least two valid figures after the decimal point. The speed of reduction of the response rate at the significant changepoint x_{sig} is defined as the first derivative of (1) with respect to x, evaluated at x_{sig}, and the deceleration is the respective second derivative. Again the symbolic formulation of the derivatives can easily be obtained from [10].

The previously discussed features together with the log-logit model (1) are implemented in our software tool "titration".

3 Results

Applying the software tool to the example data from Fig. 1 we get the regression curve estimate with 95% confidence interval within a graphic window as displayed in Fig. 3. Most of the data points lie inside the confidence region. As the 0 μM concentration is forced to have response rate 1 according to $a = 1$ in formula (1), the length of the confidence interval tends to zero in this area.

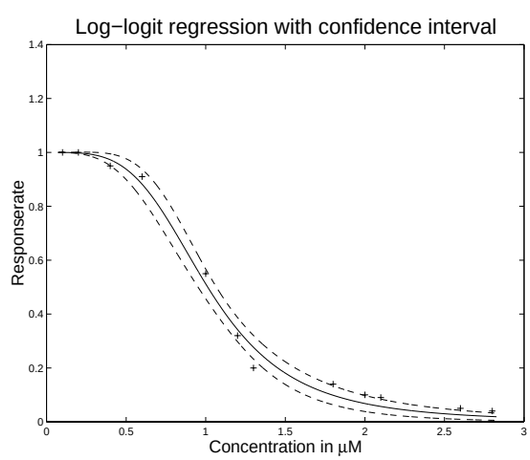

Fig. 3. Dose-response data (+) and estimated response curve (—) with pointwise 95% confidence interval (- -) from log-logit model

The respective regression parameter estimates for the model in (1), y% concentrations for $y = 10$, 30, 70, and the significant changepoint with its speed and response deceleration are obtained as text output summarized in Table 1. In

particular, the confidence intervals are necessary to compare different retinoids, tumor cell lines, etc. for a series of trials.

Table 1. Estimates of regression parameters and changepoints with 95% confidence intervals from log-logit modelling of example data in Fig.1. Sig. refers to significant.

Parameter or changepoint	Estimate	95% confidence interval
Parameter b	3.8454	[3.0821 ... 4.6087]
Parameter c = half-life concentration in μM	1.0124	[0.9554 ... 1.0694]
70% concentration in μM	0.8122	[0.7412 ... 0.8762]
30% concentration in μM	1.2620	[1.1962 ... 1.3353]
10% concentration in μM	1.7928	[1.6242 ... 1.9846]
Significant changepoint in μM	0.1533	[0.1176 ... 0.3322]
Speed of decrease at sig. changepoint in %/μM	-0.0176	
Deceleration at sig. changepoint in %/(μM)^2	-0.3266	

Due to uncertainties in laboratory measurements, upper and lower bounds of response rates cannot be fixed in advance when evaluating real data sets. Therefore we apply the full 4-parameter logistic model in the following.

The real data set indicates the rate of cell kill for ten different concentrations of the retinoid 4-HPR for prostate tumor cell line DU145. Fig. 4 displays the estimated regression curve with 95% confidence interval on a semilog(x) scale. The measurement uncertainty is reflected by the width of the confidence interval at initial concentrations as well as the three upmost concentrations. Changepoints as indicated in the figure and regression parameter estimates of the 4-parameter logistic model are summarized in Table 2. The resulting point estimates in combination with their estimated 95% confidence intervals can be used for reasons of comparison with our example date, other prostate tumor cell lines like LnCaP, PC3, or different kinds of retinoids like 13cRA. For example, the half-life concentration of 4-HPR in Table 2 is significantly lower than the half-life concentration in our example data as the point estimate of c in Table 2 is below the lower bound of the 95% confidence interval of c in Table 1.

Further comparisons have to be left to a more extensive evaluation. At this stage, we want to demonstrate the usefulness and the necessity of comparisons of this kind for our application. Furthermore, we now have a basis for an automatic evaluation with our software tool “titration”.

4 Conclusion

Initiated by a consulting case in urology, we developed a software tool for displaying estimated dose-response rates and their 95% confidence intervals simultaneously. Furthermore, a series of user-defined changepoints can be calculated together with respective confidence intervals. In this way, the effect of different

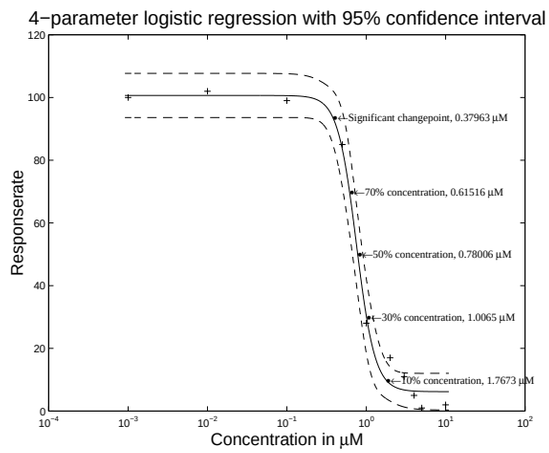

Fig. 4. Dose-response data (+) and estimated response curve (—) with pointwise 95% confidence interval (- -) from 4-parameter logistic model. Data: Response rate = cell survival = 1 - cell kill of DU145 tumor cells, concentration refers to 4-HPR on a semilog scale.

Table 2. Estimates of regression parameters and changepoints with 95% confidence intervals from 4-parameter logistic modelling of real data in Fig.4. Inf refers to infinity and sig. to significant.

Parameter or changepoint	Estimate	95% confidence interval
Parameter a	100.6533	[93.6089 ... 107.6976]
Parameter b	3.6986	[2.0429 ... 5.3543]
Parameter c = half-life concentration in μM	0.7500	[0.6265 ... 0.8735]
Parameter d	6.1924	[0.3387 ... 12.0461]
70% concentration in μM	0.6152	[0.5105 ... 0.7211]
30% concentration in μM	1.0065	[0.8511 ... 1.2078]
10% concentration in μM	1.7673	[1.2078 ... Inf]
Significant changepoint in μM	0.3796	[0.0010 ... 0.5224]
Speed of decrease at sig. changepoint in %/μM	-63.513	
Deceleration at sig. changepoint in %/(μM)^2	-359.1965	

tumor cell lines and retinoids can be compared for a number of trials. Special attention has to be drawn to numerical issues as badly-scaled problems may occure. The success of the underlying nonlinear regression procedure depends heavily on the well-scaling of the data.

A demo version of the software tool "titration" can be downloaded as ZIP-file from the first author's homepage at `http://www.imse.med.tu-muenchen.de/persons/wagenpfeil/index.html`. Please consider the readme_first.txt file. In order to run the program, MATLAB 5.0 Release 10 or lower, has to be installed.

References

1. Ortiz-Fernandez, M.C., Herrero-Gutierres, A.: Regression by least median squares, a methodological contribution to titration analysis. Chemometrics and Intelligent Laboratory Systems **27** (1995) 231–243
2. Chuang-Stein, C., Agresti, A.: Tutorial in Biostatistics: A Review of Tests for Detecting a Monotone Dose-Response Relationship with Ordinal Response Data. Statistics in Medicine **16** (1997) 2599–2618
3. Softmax Pro 3.0 User's Guide. Molecular Devices GmbH, Munich, 85737 Ismaning (1999)
4. Mardia, K.V., Kent, J.T., Bibby, J.M.: Multivariate Analysis. 8th printing, Academic Press, London, San Diego (1992)
5. Fahrmeir, L., Tutz, G.: Multivatiate Statistical Modelling Based on Generalizel Linear Models. 2nd edn. Springer-Verlag, Berlin Heidelberg New York (1996)
6. Dennis, J.E., Schnabel, R.B.: Numerical Methods for Unconstrained Optimization and Nonlinear Equations. Prentice Hall, Englewood Cliffs, New York (1983)
7. Coleman, Th., Branch, M.A., Grace, A.: Optimization Toolbox for Use with MATLAB, User's Guide, Version 2. 3rd printing, The Math Works Inc., Natick, MA 01760-1500, USA (1999)
8. Stoer, J., Bulirsch, R.: Introduction to Numerical Analysis. Texts in Applied Mathematics Vol. 12. Springer-Verlag, Berlin Heidelberg New York (1996)
9. Jones, B.: Statistics Toolbox User's Guide. 3rd printing, The Math Works Inc., Natick, MA 01760-1500, USA (1997)
10. Moler, C., Costa, P.J.: MATLAB Symbolic Math Toolbox, User's Guide, Version 2.0. 3rd printing, The Math Works Inc., Natick, MA 01760-1500, USA (1997)
11. Hartley, H.O.: Exact confidence regions for the parameters in non-linear regression laws. Biometrika **51** (1964) 347–353
12. Efron, B.: Bootstrap Methods: Another Look at the Jackknife. Annals of Statistics **7** (1979) 1–26