Automatic Best of Fit Estimation of Dose Response Curves

Dr. Denise Rey
INFORMATION WORKS
Unternehmensberatung & Informationssysteme GmbH
Rolshover Str. 45
51105 Köln
d.rey@information-works.de

Abstract

Real data dose-response measurements stored in huge databases do not always fulfil the requirements for a successful typical sigmoidal curve fit. The problems appear when measuring an improper dose range, when not having enough measurements or when the response is measured by visual assessment. Even though, indicators of the behaviour of the measured item are still needed, therefore the statistical algorithm used for the calculus has to adapt to the given circumstances. The automatic best-of-fit estimation procedure presented in this paper addresses this challenge by using a proper package of models and a filter based on the Akaike information criterion. The automatic aspect of the procedure contributes to the usability of the implemented routine in the IT landscape of the client.

Keywords: Dose-response curves, effective dose, sigmoidal, estimation, automatic.

1 Introduction

The denomination “dose-response" analysis indicates that responses of some examinations are measured in dependence of certain dose administrations. Here, “dose” could represent the amount of an active ingredient in a pharmaceutical product used in the healing process of a patient or the concentration of a plant protection chemical product. The “response” is usually the probability of occurrence of an event of interest like the healing of the patient, the non-appearance of secondary effect like toxicity or the mortality of fungi and insects in the case of plant protection products. In the preclinical studies of the pharmaceutical research and the leadfinding and greenhouse studies of the plant protection research, trials for the analysis of dose-responses are planned, with the scope of the formulation of the case adequate doses. In the early stages of the research of life sciences disciplines, the number of the substances in the screening process is huge, leading to databases of dose-response measurements in the magnitude of terabytes. Therefore, automatic statistical routines need to be implemented, to deliver at a push of a button, useful indicators for the analysis of a single dose-response curve like effective doses at x% response, slopes or confidence intervals. In the case several substances need to be compared, parameters of interest are for example activity ratios or test statistics for the log-parallelism. Combinations of substances imply other parameters of interest like the indicator for synergism or antagonism. We will restrict in this paper to a statistical algorithm designed for the automatic calculus for the analysis of a single dose-response curve.

2 User requirements for the statistical algorithm

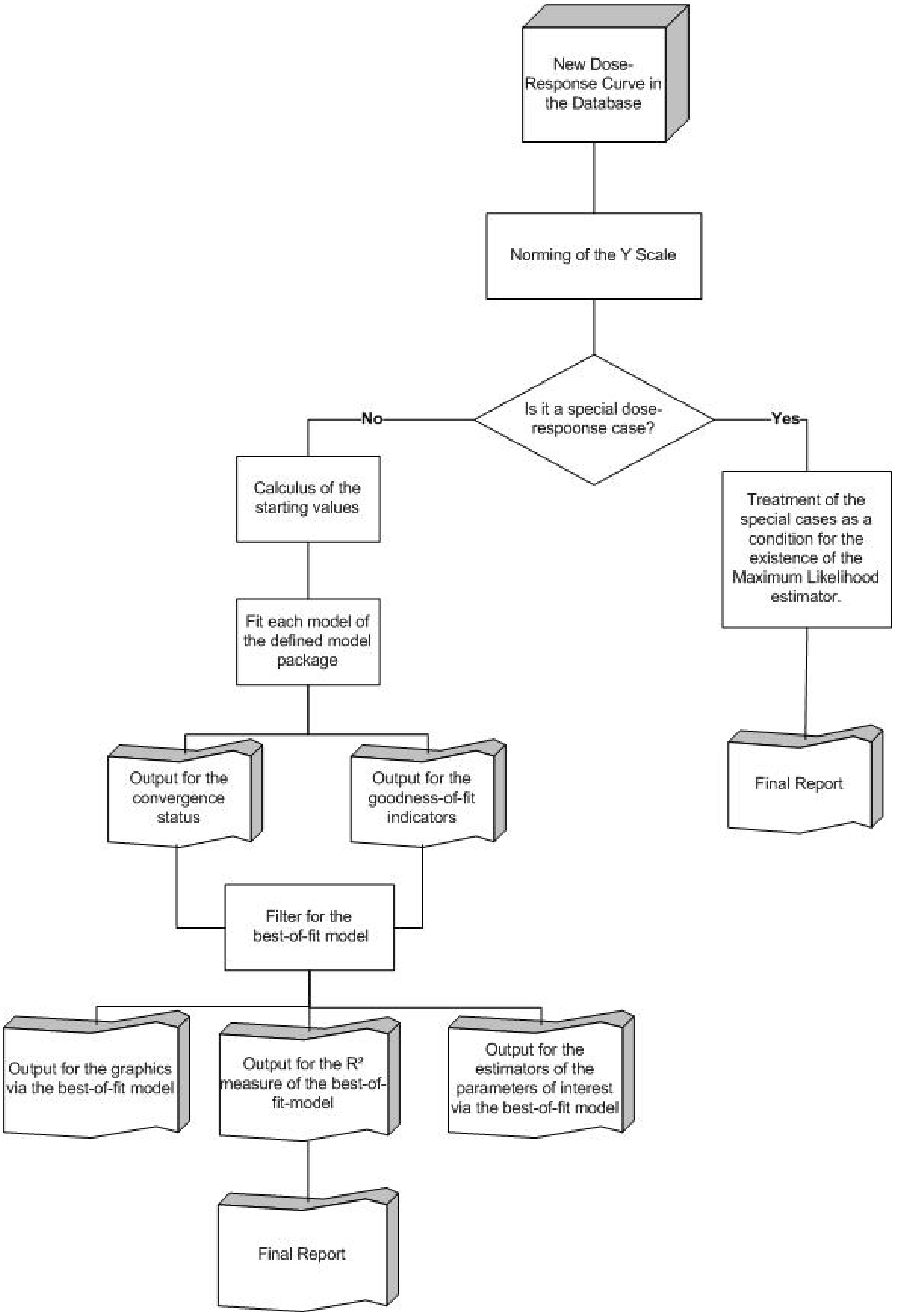
At the entry of a new dose-response curve of a chemical compound in the database, the automatic statistical algorithm calculates the parameters of interest and delivers the results back into the database. The user can access the report via the desired front-end.
Assume that the general functional requirements for the statistical algorithm for the analysis of a single dose-response curve are the following:

- The statistical algorithm should generally model S-shaped or sigmoidal dose-response curves. Even though, in case of different systematic errors, the sigmoidality of the curve can get lost. The algorithm should be able to capture valuable information also out of these cases.
- The statistical algorithm should deliver valuable reports for compounds starting with a minimum of two measured doses and one replication per dose.
- The statistical algorithm should deal with responses measured on different scales, like continuous, percentage or binary. Moreover it should also capture special cases of the scaling like, negative percentages (this could appear for example in the plant protection case of reproduction measurements).
- The statistical algorithm should capture all particular cases of dose-response curves arising usually in the case of visual assessment (like equal responses for all treated doses) and treat them accordingly.
- The developed program should deliver following estimated parameters: The estimated effective dose values at 50% response (ED50) which is defined as the point of inflection of the sigmoidal dose-response curve. Moreover, the corresponding confidence intervals, the calculated slope in ED50, the dose points at certain values of the response and the corresponding confidence intervals. Additionally, pseudo R^2 values should be recorded for each fitted dose-response curve.
- The statistical algorithm should be implemented via the SAS System. The enclosure of the statistical algorithm as a database automatic application should be included in the SAS development.

One can immediately remark, that the user functional requirements are intended to cover a very bright range of dose-response measurements. Therefore, a suitable package of modelling functions to cover all requirements will be defined. For each entry in the database the best-of-fit model will be searched by a well-defined procedure and the corresponding outputs are created.

3 The automatic process of the dose-response calculus

We will restrict this paper to the description of the statistical algorithm and not go to the development of the enclosure into the database landscape applications. Before we go into the details of the adopted statistical algorithm, we present the final process view of the implemented SAS program. A simplified picture of the solution has the form:


New Dose-Response Curve in the Database
Norming of the Y Scale
Is it a special dose-respoponse case?
No
Yes
Calculus of the starting values
Fit each model of the defined model package
Output for the convergence status
Output for the goodness-of-fit indicators
Filter for the best-of-fit model
Output for the graphics via the best-of-fit model
Output for the R² measure of the best-of-fit-model
Output for the estimators of the parameters of interest via the best-of-fit model
Final Report
Treatment of the special cases as a condition for the existence of the Maximum Likelihood estimator.
Final Report

Picture 1: Implementation process of the statistical algorithm

4 The data input and preparation step

As in the requirements mentioned, the response can be of several types: It can be a binary outcome (i.e. success or failure), a percentage (i.e. percentage of mortality or a percentage visual assessment) or a continuous outcome on the real line. Therefore, suitable transformations of the responses are applied to restrict the dose-response modelling curves to the following two functional forms:

$$f : D \rightarrow [0,1], D \subset [0,\infty) \text{ or } f : D \rightarrow \Re, D \subset [0,\infty),$$

where D is the domain of the dose points. The transformation could be done either by using indicators for the type of the measurement or by suitable separately created algorithms.

The next important issue in the preparations of the data for a dose-response analysis is the consideration of the special cases. Examples of some special cases of dose-response curves are:

Special Case 1:
The response is constant for every dose (i.e. in case of no response $f(x) = 0, \forall x \in D$).

Special Case 2:
The response takes only two values (i.e. $f(x) = 0, \forall x \in D_1$ and $f(x) = 1, \forall x \in D_2$, $D_1 \cup D_2 = D$).

These special cases arise mostly in the case of subjective visual assessment responses. For example, instead of a more detailed differentiation between two unsuccessful doses, the assessor confers them a constant response.

The generalized form of the special cases arises from a theorem for the existence of the Maximum-Likelihood estimation (estimation method used for a part of the models). Summarized, the theorem asserts that the Maximum-Likelihood estimator exists and is unique if in the case of one replication per dose, the array of responses contains a minimum of two jumps. These special cases are treated separately and they are not subject for the actual modelling. The outcome of the special cases is directly stored in the final report. For a theoretical background of the special cases, consult (Unkelbach and Wolf (1985)).

Regardless of the special cases, the dose-response curves are modelled by using a predefined set of models via the Maximum Likelihood estimation (i.e. proc probit models) and a set of models via the Least Squares estimation (i.e. proc nlin models). In view of the modelling with proc nlin, suitable starting values have to be calculated for the dose-response curves, this step belonging also to the data preparations step- More about this follows in the next chapter.

5 The modelling step

A package of models suitable for the accomplishment of the requirements is defined:

The four parametric logistic model (SAS procedure: proc nlin)

$$f(x) = c + \frac{d-c}{1+\exp\{b*(\ln(x)-\ln(ED50))\}}$$

The three parametric logistic model (with fixed minimum) (SAS procedure: proc nlin)

$$f(x) = \frac{d}{1+\exp\{b*\left(\ln(x)-\ln(ED50)\right)\}}$$

The three parametric logistic model (with fixed maximum) (SAS procedure: proc nlin)

$$f(x) = c + \frac{1-c}{1+\exp\{b*\left(\ln(x)-\ln(ED50)\right)\}}$$

The five parametric logistic model (additional location parameter)
(SAS procedure: proc nlin)

$$f(x) = c + \frac{e-c+f*x}{1+\left(1+2*\dfrac{f*ED50}{e-c}\right)*\exp\{b*\left(\ln(x)-\ln(ED50)\right)\}}$$

The two parametric generalized linear model (GLM) with logistic link
(SAS procedure: proc probit)

$$f(x) = \frac{1}{1+\exp\{-\left(\alpha*\ln(x)+\beta\right)\}}$$

The two parametric GLM with normal link
(SAS procedure: proc probit)

$$f(x) = \Phi(\alpha*\ln(x)+\beta)$$

The two parametric GLM with Gompertz link
(SAS procedure: proc probit)

$$f(x) = 1-e^{-e^{\left(\alpha*\ln(x)+\beta\right)}}$$

The reasons for the composition of the model package arise from the functional requirements of the statistical algorithm:

The GLM models implemented with proc probit are designed to model the probability of success of an event, therefore suitable for the first functional form mentioned in §4. The rest of the models, implemented with proc nlin, address the second functional form, for continuous responses on the real line. The proc nlin models also succeed to resolve a special case of dose-response curves: The case of percentage responses which could be negative or larger than 100%. This case appears for example in the case of reproduction of the insects: The mortality can be larger relative to the precount before the treatment application.

Sometimes you want to fix a parameter at a certain value, rather than estimating it, this is the case for the models with three parameters. This is useful if you do not have enough data to estimate all parameters precisely and/or you know the value a parameter should take on. In our case, theory tells us that d, the upper asymptote of the response, should be 1, since the response is expressed relative to a control value from which it should monotonically decrease, or c, the lower asymptote of the response should be 0. These would be the cases of the three parametric logistic regressions.

The five parametric logistic regression is used to detect a special case in the plant protection case which is called *hormesis*. In this case, it can happen that the responses for some small doses are larger than the control response, therefore a fifth parameter for a linear shift in the sigmoidal shape is introduced. For more details about this functional form, see (Schabenberger & other (1999)).

Moreover, the constitution of model package addresses the problematic of diverse trial designs (models with fewer parameters resolve the designs with fewer observations while models with a larger number of parameters protect against underfit).

The starting values for the proc nlin models have to be calculated in an a-priori step and temporary stored. The quality of the starting values is an essential criterion in the modelling step. One reason is that the quality of the starting values flows directly into the quality of the final estimators. Another reason is the speed of convergence of the proc nlin procedures. The starting values are calculated via an a-priori transformation and regression step of the data (Ritz and Streibig (2005)). A general theory for the nonlinear regression with proc nlin can be also find at the Web Site of UCLA Academic Technology Services (see Literature).

6 The filter step

For each dose-response entry, we fit the seven models defined in the preceding to the entry data. The fitted models will pass through a goodness-of-fit filter in order to define the best-of-fit model which will be finally used for the report of the requirements results. The filter is constructed in two steps: First, the convergence status of each fitted model is checked. If the convergence status is without errors (convergence status=0), then the model goes to the next filter step. In the next filter step, the Akaike information criterion (and the Akaike information criterion corrected for small samples (Hurvich

and Tsai (1989)) is calculated for each model. Moreover, the size of the confidence interval for the ED50 estimator is calculated. In the last step, the Akaike values are compared through the models and the model with the smalles Akaike value is picked out. If there are two such models, then the one with the smallest confidence interval becomes the best-of-fit model.
The Akaike information criterion (AIC), defined by Hirotogu Akaike in 1973, is minus twice a penalized log-likelihood value, maximized using the maximum-likelihood estimator of the unknown parameter. The AIC selects the best approximating model to the unknown true data, amongst the set of models under consideration. Even if our models are based not only on the maximum likelihood estimation but also on the least squares estimation, we have involved the AIC measure into the model selection step relying on the fact that the AIC criterion is constructed via an estimated expected value of the Kullback-Leibler (KL) distance from the unknown true data generating mechanism and the parametric model under consideration. The KL distance between two models is independent of the estimation procedure. The used AIC formula is

$$AIC_{\mathrm{mod}el} = N * \log\left(\frac{SS_{\mathrm{Re}s}}{N}\right) + 2p$$

where N is the total sample size , $SS_{\mathrm{Re}s}$ the residual sum of squares and p the number of parameters involved in the model. For more subject relevant information about the used AIC criterion, see (Motulsky and Christopoulos (2003)).

There are certainly other measures for the model selection step. We have restricted the model selection to the AIC (respectively the corrected AIC) since the AIC, compared to the Bayesian information criterion (BIC) is efficient. The efficiency leads to a better prediction power of the AIC criterion. The statistical algorithm can be extended to the use of other information criteria.

6.1 The filter for the proc nlin models

Two following output tables from the proc nlin procedures are needed for the filter of the models: The Anova table output and the ConvergenceStatus table output, both to be called via the ods line. As the name says, the ConvergenceStatus table is needed for the convergence status of the grid search algorithm. The Anova table is needed for the calculus of the AIC criterion and moreover, for the calculus of the pseudo R^2 measure.

6.2 The filter for the proc probit models

The following ouput tables from the proc probit models are needed for the filter of the models: The output table with the predicted values for the calculus of the AIC and R^2 criterion and the ConvergenceStatus table for the indicator of the convergence status for the maximum likelihood estimation.

7 The output to be reported

The model succeeding through the defined filter is called the best-of-fit model and it is used for the further reporting of the estimated parameters of interest. The parameters of interest, called in the requirements are:

- ***The pseudo R^2 measure of the model***

$$R^2 = 1 - \frac{SS_{\mathrm{Re}s}}{SS_{CorrTotal}}$$

where we have denoted by $SS_{\mathrm{Re}s}$ the residual sum of squares and by $SS_{CorrTotal}$ the corrected total sum of squares.

- ***The effective dose at 50% response (ED50) and the corresponding confidence interval***

There are several definitions of the ED50 values in literature. Here, the ED50 is the point of inflection of the fitted curve. A necessary condition for x to be an inflection point, is that

$$f''(x) = 0.$$

Moreover, a sufficient condition for x to be an inflection point requires that

$$f''(x+\varepsilon) \quad \text{and} \quad f''(x-\varepsilon)$$

to have opposite signs.

- ***The $x \in D$ value such that f(x)=0.5 (denoted by x@y=0.5) and the corresponding confidence interval***

Per definition, this is

$$x = f^{-1}(0.5)$$

The existence of this value depends of the response scale. If the value is not well-defined, the calculus will lead to an empty cell of the report.

- ***The calculated slope in the inflection point (denoted by slope)***

Per definition, this is

$$slope = f'(x)\Big|_{x=ED50}$$

7.1 The output for the proc nlin models

Assume that the best-of-fit model arises from the estimation via the proc nlin procedure. The R^2 measure is calculated from the Anova ods table of the proc nlin. Since the ED50 is a parameter in the proc nlin models, the estimated ED50 and the corresponding confidence interval are extracted from the ParameterEstimates ods table. The inverse and the derivative of the proc nlin models need to be calculated and programmed in a separate step and then applied to the estimated ED50 from the ParameterEstimates ods table to obtain the x@y=0.5 and the slope in ED50 values. For the calculus of the confidence intervals for the x@y=0.5 value, proc nlin needs to be run again and the predicted value will lead to the desired confidence interval.

7.2 The output for the proc probit models

Assume that the best-of-fit model arises from the estimation via the proc probit procedure. As we have mentioned in the paragraph before, the R^2 measure is calculated directly from the predicted values. In case of the proc probit models, the point of inflection is exactly the inverse in y=0.5, i.e. f(ED50)=0.5. The ED50 and the corresponding confidence interval are extracted from the ods table ProbitAnalysis. The slope needs to be calculated separately via the derivative of the implied model function.

All the required estimated parameters of interest are gathered together in one output table which is then reported for each dose-response item in the database. For each dose-response item, a graphic containing the measured samples, the estimated best-of-fit curve and the confidence area can be attached to the report. The graphical output is not subject of this proceeding paper. Nevertheless, as a remark, the graphics for the proc probit models can be reported automatically via the *gout* and *predplot* options of proc probit.

8 Summary

The established model package covers all the assumed user functional requirements. For every dose-response item for which an estimation with valid confidence intervals is possible, an output is created. Only if none of the models in the package converged to a solution (for the proc nlin models respectively proc probit models) there is no output for the dose-response item created. The case of non-convergence of every model corresponds to the case where a valid estimation of the ED50 is really not possible. Special cases are treated separately. The filter for the best-of-fit model based on the Akaike information criterion assures not only the estimation of the parameters of interest via the model with smallest residual sum of squares but also a good prediction power of the chosen model. One can certainly test the performance of the statistical algorithm by using other information criteria. The automatic development of the statistical algorithm assures a user-friendly push at a button report for every dose-response item to be analyzed. The automatic statistical algorithm can be imbedded into several standard SAS front-ends: As a stored process via SAS Enterprise Guide, via Microsoft Word or Microsoft Excel with the SAS Office Add-In or via a customized Web Front-End.

Literatur

[1] Hurvich CM, Tsai CL: „Regression and time series model selection in small samples.” *Biometrika*, **76**, 297-397 (1989)

[2] Ritz C, Streibig JC: „Bioassay Analysis using R.“ *Journal of Statistical Software* **12**, Issue 5 (2005)

[3] Motulsky H, Christopoulos A: Fitting Models to Biological Data using Linear and Nonlinear Regression, GraphPad Software Inc. San Diego CA (2003)

[4] Schabenberger O, Tharp BE, Kells JJ, Penner D: „Statistical Tests for Hormesis and Effective Dosages in Herbicide Dose Response.” *Agronomy Journal,* **91,** 713-721 (1999)

[5] UCLA Academic Technology Services „Nonlinear Regression in SAS” *http://www.ats.ucla.edu/stat/sas/library/SASNLin_os.htm*

[6] Unkelbach HD, Wolf T: *Qualitative Dosis-Wirkungs-Analysen*. Gustav Fischer Verlag Stuttgart (1985)