Chapter 8
Dose-Response Modeling

Gregg E. Dinse and David M. Umbach

Abstract For any definition of additivity, evaluating whether an organism's response to a mixture is additive depends on the dose-response relationships for each of the mixture's component chemicals. Consequently, the statistical analysis of dose-response relationships is fundamental to mixture toxicology – as well as to other areas of toxicology. This chapter offers a broad overview of dose-response modeling and an introduction to some statistical issues that arise in the use of dose-response models – with an eye to evaluating additivity. It does not, however, attempt to be a handbook or guide to the use of any specific models; instead, it tries to make readers aware of issues that need attention to achieve efficient and valid inference. The chapter mentions features of study design and describes how they can influence both aspects of model fitting and the quality of results. It considers the choice of functional form used to describe how the mean response changes as dose increases as well as the evaluation of how well the chosen form fits the data at hand. The chapter also points out that proper modeling of the variability inherent in the structure of the data is crucial to efficient statistical inference. Finally, because many dose-response models require iterative numerical methods, it offers a few pointers to help overcome problems when these methods fail to converge. Dose-response modeling is an essential tool in mixture toxicology but one that demands careful application to achieve the best results.

Keywords Experimental design · Hill model · Model assessment · Nonlinear regression · Statistical model · Variance structure

G. E. Dinse (✉)
Public Health Sciences, Social & Scientific Systems, Inc, Durham, NC, USA
e-mail: dinse@niehs.nih.gov

D. M. Umbach (✉)
Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, NC, USA
e-mail: umbach@niehs.nih.gov

C. V. Rider, J. E. Simmons (eds.), *Chemical Mixtures and Combined Chemical and Nonchemical Stressors*, https://doi.org/10.1007/978-3-319-56234-6_8

8.1 Introduction

One fundamental project in mixture toxicology is the evaluation of whether a mixture obeys an additive or no-interaction null model, whereby the typical response of an organism (or an isolated biological system) to different concentrations of the mixture can be predicted based on the organism's (or system's) typical response to different concentrations of each component of the mixture individually. This project has three essential elements: (1) a suitable definition of the additive state, formulated in a way that is amenable to constructing quantitative predictions (see Chap. 9); (2) an appropriate collection of data from experiments that examined dose-response relationships both for the individual component chemicals and for the mixture; and (3) an array of statistical techniques that use these data to make inferences about the question of additivity (see Chap. 11). Shortcomings regarding any of these three elements can impinge on the value of mixture experiments for risk assessment.

Because prediction of the response of an organism to a mixture under additivity – regardless of the precise definition of additivity invoked – is rooted in the dose-response relationships for each of the mixture's component chemicals, statistical analysis of dose-response curves for individual chemicals becomes a key building block for evaluating hypotheses about departures from additivity. In fact, the statistical properties of predicted responses to a mixture under an additivity assumption depend crucially on how well the statistical models used to analyze data from the individual component chemicals represent the underlying true dose-response relationships. If dose-response relationships for the component chemicals in the mixture are estimated with bias, the prediction for the mixture would likely be estimated with bias as well. Correspondingly, if the estimated dose-response relationships for the component chemicals are highly variable, the prediction for the mixture would likely be highly variable, so that tests of the additivity null hypothesis would have low statistical power. In addition, if a statistical dose-response model represents the typical dose-response trajectory accurately but fails to properly account for multiple sources of variation in the data, the precision attributed by the data analysis to the dose-response relationships might be larger or smaller than it actually is. Such errors in modeling variability in dose-response relationships can lead to incorrect conclusions about additivity, including false positive declarations of departures from additivity or false negative declarations of no departure from additivity. Analogous kinds of statistical issues centered on adequate specification and fitting of a dose-response relationship also arise when specifying models that are intended to reflect departures from additivity and fitting those models to data from mixtures of chemicals. Sound statistical analysis is essential for valid quantification of dose-response relationships; consequently, it is important throughout toxicology – including mixture toxicology.

The purpose of this chapter is to acquaint toxicologists who are now (or intend to be) working in the area of mixtures with some statistical issues that are ubiquitous in studying dose-response relationships. Because toxicologists conduct an exceedingly broad range of experiments with various outcomes and study designs, the variety of

statistical techniques needed to analyze the range of toxicologic data is comparably broad. Any attempt to even touch on all possibilities would require a book-length treatment, not simply a chapter. Instead, this chapter highlights certain issues that are common across many statistical techniques and that can have a distinct impact on inferences about dose-response relationships. First, it defines some terminology and notation that will be used throughout the chapter. Next, it discusses statistical issues related to study design and describes how they can impinge on the choice of statistical models employed and on the quality of the results. Then, it goes on to talk about modeling strategies, including the choice of functional form for fitting the mean dose-response trajectory. The chapter also draws attention to the importance of specifying a variance structure that properly reflects sources of variation in the data. In addition, it points out techniques for evaluating the adequacy of the specified dose-response model for the data at hand. Finally, because many dose-response models useful in toxicology require iterative numerical methods for fitting, it mentions some ways to cope with failure of these methods to converge to a unique best fit to the data. The chapter closes with a summary.

8.2 A Statistical Perspective on Dose-Response Modeling

To establish some definitions and basic notation, consider as a template a simple dose-response study involving a single chemical and a single response or outcome. Assume that the study involves a total number N of experimental units, where "experimental unit" is a generic term that statisticians use to denote the entity that is assigned at random to receive a particular treatment or condition in an experiment. Thus, the experimental units could be male mice in a rodent carcinogenicity study, pregnant rats in a teratogenicity study, petri plates containing a *Salmonella* tester strain in an Ames mutagenicity assay, culture dishes growing a given cell line in in vitro studies, and so on. In a dose-response study, the treatments to which the experimental units are assigned at random are the particular dose or concentration levels of the compound being investigated. Let D denote the number of dose levels and $d_1, d_2, d_3, \cdots, d_D$ denote the D particular dose levels used in the study. Because these dose levels are under the control of the experimenter, one usually regards them as known constants and not subject to errors of measurement. For simplicity, assume that the total number of experimental units N is a multiple of the number of dose levels D, such that $N/D = n$ and that the same number, n, of units are assigned to each dose level d_i. Here the index i is one of the numbers $\{1, 2, 3, \cdots, D\}$. At each dose level, each experimental unit is labeled by a second index j that is one of the numbers $\{1, 2, 3, \cdots, n\}$. Consequently, in this study, each experimental unit is uniquely identified by two indices, i indicating dose group and j indicating experimental unit within dose group. Let Y_{ij} or y_{ij} represent the response of the j^{th} experimental unit at the i^{th} dose level; uppercase Y indicates that one is thinking of the response as a random quantity that has probabilistic properties (but lacks a known numerical value), whereas lowercase y indicates that one is referring to an

actual value that would be observed in a study. Thus, y_{ij} is the observed value of the random variable Y_{ij}. Also, one omits the subscripts when referring to a response in general or uses only subscript i to emphasize the role of dose level when the particular experimental unit is inconsequential. Typically, both subscripts are needed only when referring to the response of a particular experimental unit assigned to a given dose level. Also, though i indexes the dose levels d_i used in the experiment, the notation d is used for dose level more generically – including dose levels not included in the experiment.

Depending on the response measured, data analysis for toxicologic studies utilizes a range of statistical distributions. The response of interest in a mouse carcinogenicity study might be the presence/absence of a certain type of tumor in each mouse at death. Then, Y (viewed as 1/0 for presence/absence) might be modeled statistically as having a Bernoulli distribution, and a dose-response analysis might focus on how steeply the probability of the tumor being present increases with increasing dose level. In a teratogenicity study, if the chemical under study was known to have no effect on implantation, the response of interest might be the number of rat pups in each litter born alive and without any malformations. For each litter, the count Y (which could range from 0 to the number of zygotes implanted) might be modeled statistically as having a binomial distribution, and a dose-response analysis might focus on how steeply the probability of being born alive and without malformation decreases with increasing dose level. In an Ames assay for mutagenicity, the response of interest is the number of revertant colonies on the plate; then, for each plate, the count Y (which could range from 0 upward) might be modeled statistically as having a Poisson distribution, and a dose-response analysis might focus on the rate of increase in the expected number of revertant colonies with increasing dose level. In studies that use cell lines to assess toxic effects, the response of interest might be the level of a particular enzyme. The enzyme level Y (whose value could be any non-negative number) might be modeled as having a log-normal distribution, and a dose-response analysis might examine how the typical enzyme level changes with increasing dose level.

Although the details of an appropriate data analysis would differ for each of these examples, they share some fundamental commonalities: each asks how some characteristic of the response's statistical distribution, a characteristic whose true value is unknown, changes as the dose level changes; and each acknowledges inherent variability of the response around its true value. One can think of the characteristic of interest as the signal and the inherent variability (random error) as the noise – so the goal of statistical dose-response analysis is to uncover the signal in noisy data.

From a statistical perspective, several decisions must be made in carrying out a dose-response study. These decisions fall into two phases: the design phase and the data analysis phase. First is the design phase: one must have a plan for gathering the needed dose-response data. Having a statistically efficient design for an experiment enhances validity and cost-effectiveness. Although many aspects of study design are the purview of the toxicologist, statisticians can help address issues like choosing the number and location of dose levels or setting the number of experimental units – both overall and at each dose level – to achieve acceptable statistical performance. In

complex experiments where, for example, the experiment might need to be conducted over several days and involve multiple batches of experimental units, statistical design is indispensable for allocating experimental units from different batches to different days to ensure that the dose effects of interest can be estimated without bias and that uncertainty attributable to days, batches, and experimental units can properly be assessed. The statistical design of experiments is a broad and challenging field, beyond the scope of this chapter, but the next section (Sect. 8.3) considers a few aspects of statistical design and how they impinge on dose-response data analysis.

The second phase of a dose-response study is, of course, the analysis of the data. Here the decisions have to do with constructing a statistical model that describes the data and allows estimation of the dose-response relationship of interest and quantification of uncertainty in the estimate. As indicated in the examples presented earlier in this section, one must choose a statistical or probability distribution that describes the random behavior of the experimental units at each dose level. Usually, with knowledge of the type of response being measured and personal experience or guidance from the literature regarding similar responses, the analyst will quickly nominate a small set of candidate distributions that can be refined as the analysis proceeds.

A second early decision is what feature of the statistical distribution to focus on for describing the dose-response relationship. The distributional feature used most often in the dose-response context is the mean of the dose-level-specific distribution of responses. A mean is a common way to assess central tendency in a statistical distribution, not only for continuous responses but more generally. For example, for a list of presence/absence responses (coded as 1/0), their average (i.e., the number of ones divided by the number of experimental units) gives the proportion of experimental units with 1 as a response, and such a proportion can often be interpreted as an estimate of the probability of the characteristic being present. Though features other than the mean response might sometimes be useful in dose-response analysis, this chapter considers only the mean.

Another crucial aspect of the analysis phase is specifying a model or algebraic expression that describes how the mean response changes with dose level. For example, one can represent each observation using the mathematical formula:

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \tag{8.1}$$

where μ_i is the unknown true mean response at dose level d_i from the experiment and ε_{ij} is an error term that is conceptualized as a random deviation from that mean for experimental unit j at dose level d_i, one that corresponds to the particular statistical distribution under consideration. This representation embodies the common notion that an observed value is an imperfect or noisy reflection of a true but unknown "typical" value for any experimental units that experience a particular dose level. The collection of μ_i for i in the set $\{1, 2, 3, \cdots, D\}$ represents a model for the mean as a function of dose in that the μ_i values describe how the mean changes with applied

dose level. The goal of the statistical analysis might be obtaining point and confidence interval estimates for each μ_i or for differences among them.

If the investigator were only interested in the mean response at the pre-selected dose levels tested in the experiment, estimation of those particular mean responses via Eq. 8.1 would be useful. More commonly, however, dose-response studies are conducted to make inferences about dose levels in addition to those actually tested in the experiment; perhaps an investigator seeks to characterize an entire dose-response curve or to make inference about mean responses at arbitrary dose levels either between or beyond (e.g., low-dose extrapolation) those studied experimentally. In this case, one could replace Eq. 8.1 with a possibly nonlinear regression model:

$$Y_{ij} = f(d_i|\boldsymbol{\theta}) + \varepsilon_{ij}, \tag{8.2}$$

where $f(d_i|\boldsymbol{\theta})$ is a prespecified regression function that relates dose level d_i to the unknown true dose-specific mean response μ_i and that depends on a vector of unknown parameters $\boldsymbol{\theta}$, and ε_{ij} is an error term as in Eq. 8.1. For example, in the Ames assay for mutagenicity, one often specifies $f(d|\boldsymbol{\theta}) = \alpha + \beta d$ (here, $\boldsymbol{\theta} = (\alpha, \beta)$), where α reflects the background mutant yield and β represents the mutagenic potency (Bernstein et al. 1982). A second example might involve enzyme concentrations measured in such a way that the response ranges from 0 to 1 (or, equivalently, from 0% to 100%) and one might specify $f(d|\boldsymbol{\theta})$ using the Hill model (Hill 1910), namely, $f(d|\boldsymbol{\theta}) = d^{\gamma}/(d^{\gamma} + \delta^{\gamma})$ (here, $\boldsymbol{\theta} = (\delta, \gamma)$), where δ represents the ED_{50} (or median effective dose) and γ is known as the Hill coefficient. Because Eq. 8.2 allows inferences for any d, a key difference between Eqs. 8.1 and 8.2 is the latter's capacity for allowing inference to dose levels that were not observed in the experiment. Of course, there is a trade-off: proper inferences with Eq. 8.2 rely, in part, on the assumption that $f(d|\boldsymbol{\theta})$ is correctly specified for (or at least a close approximation to) the true dose-response relationship under study. Further consideration of the mean function $f(d|\boldsymbol{\theta})$ and consequences of specifying it incorrectly appear later in the chapter.

The presentation of Eqs. 8.1 and 8.2 has so far focused on modeling the relationship between dose level and mean response as embodied in the μ_i or in $f(d_i|\boldsymbol{\theta})$; but both models also involve random errors, as embodied in the set of ε_{ij}. For now, assume that the mean response model is correctly specified so that the ε_{ij}, averaged across experimental units, have mean zero at each dose level. The important remaining properties of the ε_{ij} are their variances and their covariances. Variance may be the same for every experimental unit or may change across dose levels. A covariance is zero when two experimental units are independent and non-zero when they are correlated. In fitting Eqs. 8.1 and 8.2 to data, correct specification of variances and covariances for all experimental units is critical. This specification is done via a variance-covariance matrix for the vector $\boldsymbol{\varepsilon}$ (whose N entries are the ε_{ij}). Let $\boldsymbol{\Sigma}$ denote the $N \times N$ variance-covariance matrix for $\boldsymbol{\varepsilon}$ (or, more generally, for the vector of observations $\mathbf{Y}$ given their true means); the diagonal elements of $\boldsymbol{\Sigma}$ are the variances of the ε_{ij} and the off-diagonal elements are their pairwise covariances. The matrix $\boldsymbol{\Sigma}$ can depend on one or more unknown parameters (denoted by vector $\boldsymbol{\omega}$);

one can write $\boldsymbol{\Sigma}(\boldsymbol{\omega})$ to emphasize that dependence and regard $\boldsymbol{\Sigma}(\boldsymbol{\omega})$ as representing a model for the variance-covariance matrix in terms of unknown parameters to be estimated.

For example, when the dose-level-specific response distribution is normal, a typical default assumption is that its variance is unknown but constant across dose levels and that the individual ε_{ij} are independent (their covariances are zero) both within dose level and across dose levels (in this example, the off-diagonal elements of $\boldsymbol{\Sigma}$ are all zero, and the diagonal elements are all the same unknown constant often denoted σ^2). The assumption of constant variance may not always hold, however. Some probability distributions have the property that their variance and their mean are related in a defined way. For example, for the Poisson distribution, the mean and the variance are equal; for the Bernoulli and binomial distributions, both the mean and the variance depend on the success probability; for the log-normal distribution, the variance increases as the mean increases in a prescribed way. Thus, certain forms of heteroskedasticity (nonconstant variance) are built into the data analysis via the statistical distribution chosen for the analysis. Sometimes, however, the mean-variance dependence that is built in by the chosen distribution does not adequately accommodate the observed degree of heteroskedasticity, so that the data analyst must incorporate additional parameters to accommodate extra variability (Breslow 1984; Williams 1982). For example, models that incorporate extra-binomial variation are commonly applied in studies of possible teratogens where a dam is the treated experimental unit, but a presence/absence response is assessed on each dam's individual pups (Haseman and Hogan 1975; Piegorsch and Haseman 1991; Zorrilla 1997) and summarized into a single Y_{ij} for each dam. Even for distributions like the normal that accommodate constant variance, the underlying data-generating mechanism may deliver heteroskedasticity across dose levels; and such heteroskedasticity must be properly taken into account to achieve efficient and valid statistical inference.

Another consideration when modeling the variance-covariance structure of the data is whether the observations can be modeled as independent, which implies that covariances are zero. When experimental units that are homogeneous (representing a single unstructured population of units) are assigned at random to dose levels, the homogeneity of the units together with the randomization process strongly supports that the observations would be independent. If the experimental units are not homogeneous to begin with, however, but instead represent multiple subgroups, then non-zero covariances can arise even with randomization. Say, for example, the experimental units are mice and the mice needed for the study were accumulated by taking multiple littermates from several different litters, then arguably the responses of two littermates may well be more similar to each other than the responses of two mice from different litters – leading to non-zero covariances for certain pairs of units that should be acknowledged in the data analysis. Aspects of the way an experiment is conducted can also lead to non-zero covariances between experimental units. For example, if an experiment is so large that it must be carried out in multiple runs and each run involves several experimental units and is accomplished on a different day, or if a procedure involves incubating treated plates (the experimental units) at a

controlled temperature for a certain period and the large number of plates involved necessitates the use of multiple incubators, then two units from the same run or in the same incubator could arguably be more similar in response than two units from different runs or in different incubators. In this way, details of the conduct of an experiment can have a strong bearing on the covariance structure that may exist among the observations.

Thus, an overall statistical model for the data from a dose-response study typically consists of three components: (1) a probability distribution appropriate for the response under study; (2) a mathematical model, denoted here by $f(d|\boldsymbol{\theta})$, that describes the relationship between mean response and dose; and (3) a model, denoted here by $\boldsymbol{\Sigma}(\boldsymbol{\omega})$, for the variance-covariance matrix of the data. After the data analyst has established such a statistical model for the dose-response data, the task is to estimate the unknown parameters $\boldsymbol{\theta}$ and $\boldsymbol{\omega}$ and to quantify the uncertainty in those estimates. For example, the model $f(d|\boldsymbol{\theta}) = \alpha + \beta d$ describes an entire family of possible straight lines because both α and β can take on values ranging from $-\infty$ to ∞. Estimating the unknown parameter $\boldsymbol{\theta} = (\alpha, \beta)$ amounts to choosing the specific values of α and β that produce the straight line that fits the data best.

The process of fitting a model to data – that is, of estimating the particular values of the unknown parameters $\boldsymbol{\theta}$ and $\boldsymbol{\omega}$ that provide the best fit to the data at hand – is, of course, a key step in data analysis. Statisticians have devised various criteria to operationalize the concept of "best fit." For regression models for normally distributed observations, estimates of $\boldsymbol{\theta}$, denoted $\widehat{\boldsymbol{\theta}}$, are typically derived using the principle of least squares, where the best estimates are those that minimize the sum of squared differences between observed data and model predictions (Seber and Wild 1989). Variance parameters $\boldsymbol{\omega}$ for normally distributed observations can be estimated by equating observed average squared deviations to their expected values expressed as functions of $\boldsymbol{\omega}$ (Searle et al. 2006). Another widely used criterion for "best fit" is based on the principle of maximum likelihood. Generally, the likelihood is an algebraic expression of the joint probability of the observed data regarded as a function of the unknown parameters. The value of the unknown parameter declared to fit best is the one that makes this joint probability as large as possible – hence, maximum likelihood. Maximum likelihood delivers estimates of both $\boldsymbol{\theta}$ and $\boldsymbol{\omega}$ directly. For normally distributed data, the least squares and the maximum likelihood estimates of $\boldsymbol{\theta}$ coincide – a relationship that may not hold for other distributions. Maximum likelihood estimates are widely used because they have desirable statistical properties that hold for many distributions (Mood and Graybill 1963). Of course, other ways of defining "best fit" could also be used: some are versions of maximum likelihood such as restricted maximum likelihood or penalized maximum likelihood; others are developed in a Bayesian framework (Box and Tiao 1992; Carlin and Louis 2000). Maximum likelihood receives the most attention in this chapter.

8.3 Design Considerations

Proper experimental design can have a substantial impact on the precision and power of statistical inferences from a dose-response study. A study's design can also influence the range of models that can be successfully fitted to the eventual data. Designing an experiment always entails trade-offs. In a study with multiple goals, a good design for addressing one goal may be less useful for addressing another; in a study with multiple endpoints, an ideal design for one endpoint may be suboptimal for another. Consider a study where the investigator anticipates that the appropriate dose-response function is a straight line between the lowest and highest dose levels of interest. Assume further that the responses will be normally distributed and have constant variance across all dose levels. Then, for the goal of having the most precise estimate of the slope of the line, the optimal design is to allocate half the experimental units to the lowest dose of interest and the other half to the highest (Seber 1977). If the investigator entertained any doubts that a straight line was the correct dose-response model, this design could be catastrophic, as it is completely unable to detect any curvature in dose-response trajectory. Allocating experiment units to intermediate dose levels would allow one to check whether, in fact, the straight line was an appropriate dose-response model, though at the cost of some loss of precision for slope estimation, and would enable the investigator to fit a more appropriate model if necessary. When selecting dose levels for an Ames mutagenicity assay, the investigator is faced with just this kind of trade-off; at higher dose levels, toxicity begins to dominate mutagenicity, and an increasing dose-response curve tends to bend downward. Possible goals for a dose-response study for an individual endpoint include choosing a model that aptly describes the dose-response trajectory, estimating the unknown parameters of the selected model, assessing risk through quantities such as the median effective dose or some other benchmark dose, and predicting expected response at any desired dose level – and seeking the most accurate and precise estimation possible for these last three. A chosen design will often have to compromise among such competing goals – while simultaneously honoring any constraints imposed by budgets, facilities, and available time.

Consider first a simple setting, known to statisticians as a completely randomized design. In that setting all the experimental units are viewed as homogeneous, and all are allocated at random to the chosen dose levels; such a design incorporates no structure arising from different batches of experimental units or from different technicians or different runs or other restrictions on randomization. In that setting, the basic quantities contributing to the statistical design are the number of dose levels in the study, the actual dose levels employed (think of their placement – location and spacing – on the dose axis of a graph), and the number of experimental units assigned to each dose level. The choice of each of these quantities is informed both by the nature of the dose-response models to be fitted to the data, namely, the set of functions $f(d|\boldsymbol{\theta})$ under consideration, and by the number of experimental units that the investigator can afford to obtain and carry through the study. Generally speaking, more flexible models can be fitted when more dose levels are studied and,

for a fixed number of dose levels, estimates become more precise as the number of replicate experimental units allocated to each dose level increases – but costs and other practicalities typically limit these numbers.

A fixed number N of experimental units could be allocated to particular dose levels in various ways, ranging from placing all of them at a single dose to placing a single observation at N distinct dose levels. Statisticians often think in terms of an optimal experimental design – one where the placement of dose levels and the proportion of the total number of experimental units allocated to each dose level is guaranteed to meet some desirable property (often to minimize the variances of the parameter estimates) for any total number of experimental units. An optimal design for minimizing the variance of an estimated slope for a straight line was mentioned earlier. Such optimal designs are a prominent area of theoretical statistics, but most of the work and the results are applicable when the model used to analyze the data is linear in the unknown parameters, for example, a straight line or many models used in analysis of variance. In these settings, often there will be a unique optimal design regardless of the true values of the unknown parameters. A majority of dose-response models used in toxicology, however, are nonlinear in the unknown parameters (e.g., the Hill model mentioned earlier). Though optimal designs may still be found, finding them is much more difficult for nonlinear models; and, when found, they often have the unfortunate property that the placement of the dose levels for the optimal design changes depending on the true values of the unknown parameters (Seber and Wild 1989) – which means that the investigator has to know a good deal about the underlying truth before a dose-response experiment can be designed optimally. Consequently, focusing on optimal designs is not practical in the present context, and instead the focus will be on heuristics for good design.

In general, the number of dose levels (including dose zero) in the study must, at minimum, equal the number of individual parameters in the vector $\boldsymbol{\theta}$. Under ideal circumstances, that minimal number of dose levels ensures all individual parameters can be estimated uniquely (if several different models are being considered, then use the $\boldsymbol{\theta}$ with the most elements to determine the minimum). For a simplistic example illustrating the problem with having too few dose levels, consider fitting a straight line (where $\boldsymbol{\theta}$ has two individual parameters) to a design where all the replicates were allocated to a single dose level. Although one could certainly estimate a mean response at the single dose level, one could not estimate an intercept and slope uniquely: there is no best choice among the infinitely many lines that can be fitted through a single point. To estimate both parameters, a design must employ at least two dose levels. Analogously, at least two dose levels would be required to estimate the parameters γ and δ in the Hill model mentioned earlier. Similarly, if $\boldsymbol{\theta}$ contained four individual parameters, then a study would need at least four distinct dose levels to have any chance of obtaining unique estimates of the four parameters.

While the number of parameters in the selected model provides a minimum for the number of dose levels required, usually it would be unwise to implement a design with such a limited number of dose levels – for several reasons. One has already been mentioned in connection with the optimal design for a straight line: if an investigator wants to check whether a more complex model, one with more parameters to allow a

richer variety of shapes, might fit better than the one used to determine the minimal number of dose levels, the design needs to have additional dose levels to accommodate the more complex alternative model. A second reason has to do with the placement of the dose levels.

How the placement of dose levels can influence the ability of a design to estimate unknown parameters can be illustrated by an example using a sigmoid dose-response function, a shape often relevant in toxicology. Consider an extended version of the Hill model mentioned previously, namely,

$$f(d|\boldsymbol{\theta}) = L + (U - L) \times [d^{\gamma}/(d^{\gamma} + \delta^{\gamma})] \tag{8.3}$$

(here, $\boldsymbol{\theta} = (L, U, \delta, \gamma)$), where L is the lower response limit, U is the upper response limit, δ represents the ED_{50} (or median effective dose), and γ is known as the Hill coefficient and is related to how sharply the curve rises (or falls). Whereas in the original version of the Hill model (the portion enclosed in brackets in Eq. 8.3), the possible response values ranged from 0 to 1 (assuming an increasing function, i.e., $\gamma > 0$); in this version, response values range from L to U, and both of these limits are to be estimated. If $f(d|\boldsymbol{\theta})$ is plotted against the logarithm of dose, the resulting graph is sigmoid, with a long shallow rise from an asymptote at L with small dose levels, eventually rising more quickly, reaching its steepest slope at the ED_{50}, then gradually rising more slowly as it increases toward an upper asymptote at U with large dose levels.

For the four-parameter model of Eq. 8.3, a design must incorporate at least four distinct dose levels. Suppose an experimental design assigned four non-zero dose levels to be equally spaced on a logarithmic scale (often a convenient spacing for a laboratory because of the ease of serial dilutions). If the four dose levels selected all corresponded to the early part of the curve when the increase was shallow (say, all were well below the ED_{50}), then one gets a lot of information about response levels close to L but little information about other response levels (Fig. 8.1a). Thus, it is likely that only the parameter L would be estimated satisfactorily; the data from such a design would not serve to estimate the other parameters well at all. Similarly, if the four dose levels selected for the design were all above the ED_{50}, then data resulting from the design might estimate U well but not the other three parameters (Fig. 8.1b). Alternatively, if the four doses were widely spaced on the log scale but the true curve rose rapidly over a narrow intermediate dose range, it is possible that the two lower doses would correspond to the early relatively flat part of the curve and the two higher doses would correspond to the late relatively flat part of the curve, with no data collected in the region where the dose-response function changed rapidly (Fig. 8.1c). In that situation, data arising from the design might estimate L and U well but not the ED_{50} or Hill coefficient. Of course, estimating all the model parameters well simultaneously is important if one seeks a reliable estimate of the entire dose-response curve, which requires appropriate placement of the four dose levels (Fig. 8.1d). Even the placement of doses d_1–d_4 in Fig. 8.1d, though better than the placements in the other panels, may not sufficiently capture information about the flat parts of the curve. The use of two additional dose levels (d_0, d_5) to extend the

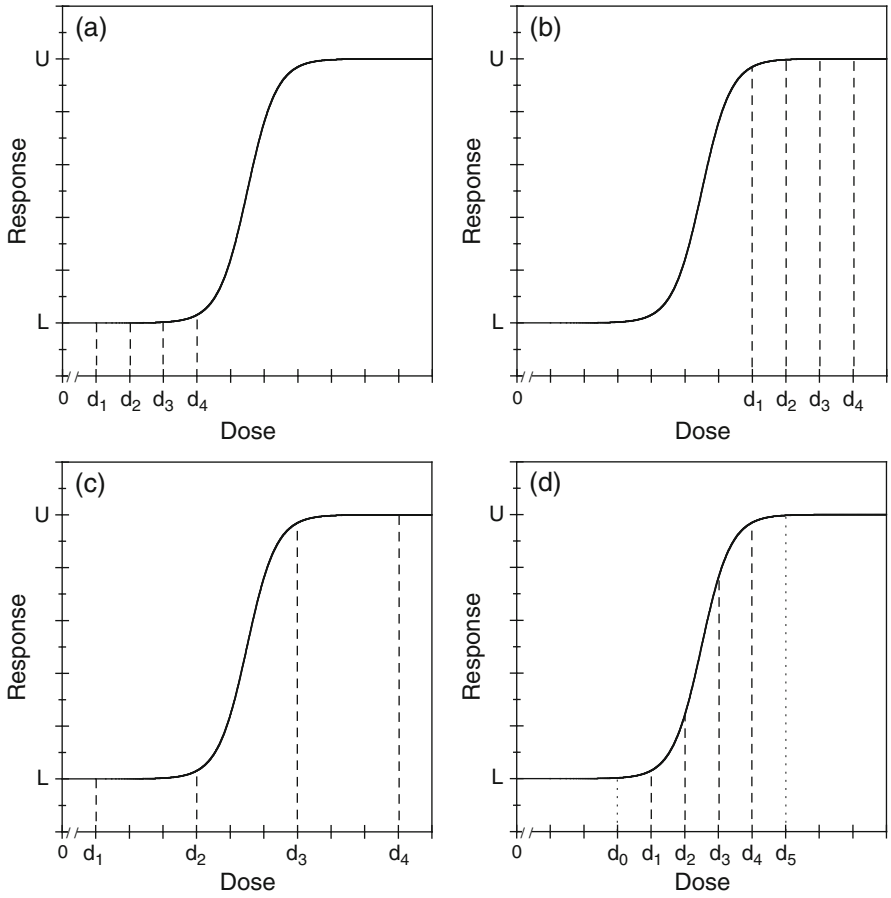

Fig. 8.1 Illustration of how dose-level placement may influence estimation of dose-response relationships. Each panel shows a dose-response function generated from the four-parameter Hill model in Eq. 8.3 and four evenly spaced dose levels (d_1, d_2, d_3, d_4). Panel (**a**): all dose levels on the lower flat portion of the curve provide information mostly about parameter L. Panel (**b**): all dose levels on the upper flat portion of the curve provide information mostly about parameter U. Panel (**c**): failure to include dose levels in the region where the curve rises provides information mostly about parameters L and U. Panel (**d**): dose-level placement that includes the region where the curve rises as well as the shoulders where the curve is nearly level provides information about all four parameters; placing additional dose levels at d_0 and d_5 (alternately, replacing dose levels d_1 and d_4 with d_0 and d_5) would provide better information about the flat parts of the curve and enhance estimation of all parameters

range of dose levels would likely improve estimation of all parameters, pointing to the value of using more dose levels than the minimum required number.

Another way to think about why the dose level placements mentioned in the previous paragraph are problematic is by relating the dose levels in the design to the expected response levels. The set of dose levels included in a design should

correspond to a set of true unknown response levels that are representative of the entire range of possible underlying response values. Thus, a design whose dose levels only capture the low end but not the middle or high end of the response range – like the design mentioned earlier with all dose levels below the ED_{50} – is unlikely to be a good design for a sigmoid curve. Similarly, a design whose dose levels capture only the highest and lowest response levels but none in the middle – because the dose spacing is too wide to have doses corresponding to intermediate response levels – is unlikely to be a good design. The placement of dose levels is important so that the design captures a range of responses at the collection of chosen doses.

This discussion serves to illustrate a point made earlier – that an optimal design for nonlinear models depends on the unknown true values of the parameters. One must have some idea about the shape and location of the dose-response curve one is trying to estimate if one hopes to design an experiment to estimate it well. Although equally spaced dose levels, either on an additive or logarithmic scale, are common default choices, irregular spacing of dose levels – farther apart where the dose-response curve is expected to be flat and closer together where the dose-response is expected to change rapidly – can be a useful strategy, but one that demands more extensive prior knowledge of the dose-response curve. If an investigator does have a good sense of an appropriate range of dose levels to represent the full range of response levels, then choosing a number of dose levels nearer the minimum required by the number of model parameters seems reasonable – though including several additional dose levels that will allow flexibility for fitting more complex models and will accommodate any lingering uncertainty as to the appropriate dose range would be a prudent strategy. On the other hand, if the investigator is very uncertain of the nature of the dose-response curve, other strategies are needed. Most important would be to employ a pilot dose-finding study to help home in on an appropriate dose range. In addition, when the dose-response shape is uncertain, designs using a relatively large number of dose levels (and necessarily fewer experimental units at each dose) would increase the chances for avoiding some of the aforementioned problems with poor dose placement.

Usually designs assign an equal number of observations at each dose level – this strategy is just simple to execute. If the investigator expects that variability in response is greater at some dose levels than others – say, variability is greater at higher dose levels – allocating more experimental units at those dose levels expected to be most variable can be an efficient strategy but one that is not widely used.

The design considerations to this point have focused on a completely randomized design; but such simple designs are not the best approach in every study. Many dose-response experiments must accommodate distinct batches of experimental units or involve procedures that must be carried out on different days or using several instruments of the same type to accommodate the throughput. In such experiments, statistical efficiency demands that one account for the batches or days or instruments at the data analysis stage – because they contribute batch-to-batch or day-to-day or instrument-to-instrument variability. The statistical design goal is to be able to remove, when possible, such unavoidable but attributable variability from the variances of dose-level comparisons and the variances of parameter estimates. One

statistically useful but relatively simple design in this context is a randomized complete block design, where "block" is a generic statistical term for a set of experimental units that have some feature in common. That feature might be a common source (e.g., mice from different litters could constitute separate blocks, or mice from different strains could constitute separate blocks), or that feature might be some commonality in the way units are handled during the conduct of the experiment (e.g., if multiple incubators are used in an experiment, the set of experimental units assigned to each incubator would be considered a block). A randomized complete block design assigns the experimental units within each block at random to the set of dose levels under study. Thus, each block is a dose-response study with one experimental unit allocated to each dose level; in a sense, each block is a separate mini-dose-response study. This design is useful when the number of experimental units in a block (the block size) is large enough to handle the entire set of dose levels under study. If the block size is smaller than the number of distinct dose levels, then more complicated designs known as randomized incomplete block designs may be appropriate. Statisticians have devised many sorts of statistical designs to handle various sorts of restrictions on randomization imposed by the way experiments must be conducted. The intent here is to make readers aware of these issues and to point out that consultation with a statistician at the design stage can help an investigator make the best use of resources in settings where complex blocking may be needed.

8.4 The Model Describing How Mean Response Depends on Dose: $f(d|\theta)$

Toxicologists typically consider dose-response curves that are continuous (i.e., without jumps) and where the mean response either increases across the entire range of doses or decreases across that entire range. Dose-response curves that never change directions are called "monotone." A monotone dose-response model can have one or more flat regions; however, if a monotone model has no flat regions, it is called "strictly monotone." A model is called "non-monotone" when it exhibits any change in direction (i.e., the response increases over some dose ranges but decreases over others).

Non-monotone dose-response models are not widely used in toxicology, but they have some applications. For example, in the Ames assay, dose-response curves may turn downward at the highest dose levels when cell toxicity dominates mutagenicity (Margolin et al. 1981); others have allowed the possibility of non-monotonicity when fitting flexible curves for relative potency estimation (Guardabasso et al. 1987, 1988). Consideration of hormesis also leads to non-monotonicity in dose-response models (Hunt and Bowman 2004; Kim et al. 2016). As with the toxicity-mutagenicity competition just mentioned, models constructed to reflect an increase in response attributable to one process (say, mutagenicity) with a decrease in

response due to another (say, toxicity) have been applied to non-monotone dose-response data in other settings, such as high-throughput screening experiments (EPA 2016; Shockley 2016).

Monotone models with flat regions are used more often, however. A typical example is a threshold model where the response is initially flat from dose zero up to some critical dose where the monotone increase or decrease in response begins (e.g., Casey et al. 2004). The initial flatness of a threshold model is interpreted as a continuation of the response in the absence of chemical exposure until the dose becomes sufficiently high to elicit a measureable response. Though such models are useful, using noisy data to distinguish a flat region (where the slope is zero) from a region with a very shallow positive or negative slope is difficult; consequently, inference about the critical dose (join point) is difficult in the sense that estimates of the join point often have large standard errors. Also, from a curve-fitting perspective, a flexible and carefully chosen strictly monotone model can often closely mimic and be difficult to distinguish statistically from a threshold model. Consequently, this chapter will focus primarily on strictly monotone dose-response models.

Another reason to favor strictly monotone dose-response models in the context of mixtures is that methods for constructing the dose-response curve for a mixture under an assumption of additivity (e.g., Berenbaum's definition of dose additivity; Berenbaum 1985) become problematic without strict monotonicity. A dose-response model maps a given dose to the corresponding expected response level; but use of Berenbaum's definition requires mapping an observed response level to the dose of each component chemical that separately would induce that response level. What is required is a mapping from response to dose – the inverse of the dose-response function. If a dose-response curve has a flat section between two dose levels, the response level of the flat portion does not correspond to a unique dose – but to any dose between those two dose levels. A strictly monotone dose-response curve, however, has a unique response corresponding to each dose and can be inverted to provide a unique dose corresponding to each response. Thus, strict monotonicity of the dose-response model is desirable in connection with mixtures.

Luckily, the class of strictly monotone mathematical functions to use as dose-response models is large and can accommodate a wide variety of curve shapes. Many dose-response models that toxicologists use routinely are nonlinear, often sigmoid. In Eq. 8.3, an extended version of the Hill model was introduced; this model is strictly monotone so long as $\gamma \neq 0$, and it is increasing or decreasing depending on the sign of γ. Consider replacing the Hill function, $d^{\gamma}/(d^{\gamma} + \delta^{\gamma})$, in Eq. 8.3 by a general strictly monotone function $g(d|\boldsymbol{\theta}^*)$ that is governed by a vector of parameters $\boldsymbol{\theta}^*$ and that takes values between 0 and 1 as d increases from 0 to ∞ (or, equivalently, as the logarithm of d increases from $-\infty$ to ∞). One can write the resulting more general dose-response model as:

$$f(d|\boldsymbol{\theta}) = L + (U - L) \times g(d|\boldsymbol{\theta}^*) \tag{8.4}$$

(here, $\boldsymbol{\theta} = (L, U, \boldsymbol{\theta}^*)$). This formulation allows $f(d|\boldsymbol{\theta})$ to increase from L to U or to decrease from U to L, depending on whether $g(d|\boldsymbol{\theta}^*)$ is monotone increasing or

decreasing in d. For concreteness, the mean response in this chapter will be assumed to increase monotonically with dose, unless otherwise stated. Therefore, the lower response limit L is the value of $f(d|\boldsymbol{\theta})$ when $d = 0$ and the upper response limit U is the value of $f(d|\boldsymbol{\theta})$ as dose d gets arbitrarily large.

Generally speaking, the response limits in Eq. 8.4 can take any values such that L is smaller than U. Some experiments may involve natural boundaries for the mean response, in which case L and U might be assigned fixed values a priori. For example, if the response is a percentage of experimental units responding, one might specify $L = 0$ and $U = 100$ or in fact any pair of intermediate values that satisfy $0 \leq L < U \leq 100$. Alternatively, one or both of the response limits can be treated as unknown and estimated from data in the current experiment (Dinse and Umbach 2011) or from data on negative and positive controls in historical studies. Even if the mean response is a proportion, the lower limit could exceed 0% if there were a background rate, which could occur if a fraction of the population showed an effect even without chemical exposure. Similarly, the upper limit could be less than 100% if a fraction of the population did not show an effect regardless of how great the dose became.

The family of curves described by the dose-response model of Eq. 8.4 changes for different specifications of the function $g(d|\boldsymbol{\theta}^*)$. Of course, the Hill model of Eq. 8.3 is one family where $g(d|\boldsymbol{\theta}^*) = d^\gamma/(d^\gamma + \delta^\gamma)$ with $\boldsymbol{\theta}^* = (\delta, \gamma)$**.** For statisticians, one convenient way to specify a monotone increasing curve for $g(d|\boldsymbol{\theta}^*)$, which has a lower bound of 0 and an upper bound of 1, is to use a cumulative distribution function for a known statistical distribution, which by definition increases monotonically from 0 to 1. (If a monotone decreasing $g(d|\boldsymbol{\theta}^*)$ is desired, subtract the cumulative distribution function from 1 and use that difference as $g(d|\boldsymbol{\theta}^*)$.) Some common statistical distributions used in toxicology include the logistic (Reeve and Turner 2013); the normal, which leads to the well-known probit model (Finney 1971); and the Weibull (Christensen and Nyholm 1984). Many other choices are possible, of course; but usually choices are restricted to familiar statistical distributions.

Statistical cumulative distribution functions have rigidly defined shapes. Investigators looking for more flexibility in the shape of a dose-response model to better fit the data at hand have several options. One is to create a new dose-response model by replacing dose d in an existing model with a transformed version of dose d; a second is to use other kinds of models that allow more flexible dose-response shapes, such as regression splines or smoothing splines.

Perhaps the most frequently used transformation of dose is the natural logarithm of dose, in symbols, $t(d) = ln\,(d)$. Here, $t(\cdot)$ is notation for a generic transformation, and $ln(\cdot)$ is the natural (base e) logarithm function. If one starts with a function $f(d|\boldsymbol{\theta})$ and substitutes $t(d)$ for d, one gets a new dose-response function $f^*(d|\boldsymbol{\theta})$. Consider the linear dose-response model $f(d|\boldsymbol{\theta}) = \alpha + \beta d$. If one replaces d with $\ln(d)$, the new dose-response model becomes $f^*(d|\boldsymbol{\theta}) = \alpha + \beta \ln(d)$. The latter model describes a different family of curves than the original model. The same kind of manipulation is possible starting from any model $f(d|\boldsymbol{\theta})$; of course, one has no guarantee that the resulting model $f^*(d|\boldsymbol{\theta})$ will be better suited than the original for the data at hand.

The distinction between creating a new dose-response model and simply reparameterizing the original model is important and is often a point of confusion. Often a model expressed as a function of d is reexpressed as a function of $ln(d)$ for mathematical convenience or computational stability. This reexpression takes advantage of the mathematical identity: $d \equiv \exp(\ln(d))$. For example, the popular Hill model (limited here to the response range 0–1) is frequently written as a function of dose d:

$$g(d|\boldsymbol{\theta}) = d^{\gamma}/(d^{\gamma} + \delta^{\gamma}). \tag{8.5}$$

Alternatively, Eq. 8.5 can be algebraically rearranged and reparameterized to give the following logistic model, which is expressed as a function of log dose:

$$g(d|\boldsymbol{\theta}) = 1/[1 + \exp(-\alpha - \beta \ln(d))], \tag{8.6}$$

where parameters α and β have different interpretations from parameters δ and γ. Often α is referred to as an intercept and β as a slope because Eq. 8.6 can be rewritten as a linear function of $\ln(d)$, namely, $\ln\{g(d|\boldsymbol{\theta})/[1 - g(d|\boldsymbol{\theta})]\} = \alpha + \beta \ln(d)$. If one sets $\alpha = -\gamma \ln(\delta)$ and $\beta = \gamma$, then Eqs. 8.5 and 8.6 coincide and both specify exactly the same mean response for any particular dose. Thus, simply because two dose-response models look distinct algebraically does not mean that they must specify two different families of dose-response relationships – sometimes two models are exactly the same even though the functional forms appear to be different.

For some models, one parameterization may offer computational or interpretational advantages over another. For example, when summarizing results in terms of the ED_{50}, one might prefer a model that incorporates the ED_{50} directly as a parameter, such as parameter δ in Eq. 8.5, rather than calculating it indirectly from other parameters. Alternatively, a model expressed in terms of $\ln(d)$, such as Eq. 8.6, has properties that make it less susceptible to numerical problems with model fitting: it minimizes curvature and thus reduces bias by more closely mimicking a linear model (Bates and Watts 1988; Reeve and Turner 2013).

Rather than selecting a fixed dose transformation in advance, one can write the transformation function $t(\cdot)$ as a function of unknown parameters and build those additional parameters into an original dose-response model, in essence estimating a dose transformation that enhances model fit. Perhaps the most common transformation function with an adjustable parameter is the Box-Cox transformation (Box and Cox 1964): $t(d) = (d^{\lambda} - 1)/\lambda$, where the parameter λ governs the shape of the dose metric, with a continuum of dose transformations for the range of λ values between $-\infty$ and ∞. This family of transformations includes (after changing multiplicative and additive constants to 1 and 0, respectively) $t(d) = \sqrt{d}$ if $\lambda = \frac{1}{2}$, $t(d) = 1/d$ if $\lambda = -1$, and $t(d) = ln(d)$ in the limit as $\lambda \to 0$. For example, to obtain a linear dose-response model with an arbitrary dose metric, one could substitute $(d^{\lambda} - 1)/\lambda$ for d in $f(d|\boldsymbol{\theta}) = \alpha + \beta d$ to create a new model $f^{*}(d|\boldsymbol{\theta}) = \alpha + \beta[(d^{\lambda} - 1)/\lambda]$ and then estimate λ together with α and β using software for nonlinear regression. The same general procedure could be applied to almost any dose-response model. For instance,

Altenburger et al. (2000) considered several models that included a Box-Cox transformation of dose.

One feature to be aware of when applying the Box-Cox approach is that the mean response at dose 0 can differ from the response limit expected under the original model. For example, consider a dose-response model $f(d|\boldsymbol{\theta})$ as in Eq. 8.4, with $g(d|\boldsymbol{\theta}^*)$ having the form shown in Eq. 8.6, except that $(d^\lambda - 1)/\lambda$ is substituted for $\ln(d)$. An estimate of λ close to 1 would suggest using $d - 1$ (or, equivalently, d) as the dose metric. By definition, at dose 0 the values of $g(d|\boldsymbol{\theta}^*)$ and $f(d|\boldsymbol{\theta})$ should be 0 and L, respectively. For an increasing dose-response curve ($\beta > 0$), however, substituting d for $\ln(d)$ in Eq. 8.6 gives $g(0|\boldsymbol{\theta}^*) > 0$, and thus $f(0|\boldsymbol{\theta}) > L$. Therefore, use of the Box-Cox approach can force a non-zero background rate even if $L = 0$. A similar problem occurs for a decreasing dose-response curve ($\beta < 0$), where the mean response at dose 0 would remain below the upper limit, i.e., $f(0|\boldsymbol{\theta}) < U$, thereby precluding 100% response even if $U = 100$. In other words, employing the Box-Cox transformation approach can imply a redefinition of parameters L and U in models having the basic structure of Eq. 8.4.

Up to this point, the presentation has focused on parametric dose-response models. Each of these models has a family of curves associated with it, and each expresses the mean response as a smooth and strictly monotone function of dose. These models often have parameters with useful interpretations. The shapes of the curves within each family are rigid in certain ways, however. Despite one's ability to adjust a model's parameters to achieve a best fit within that model or family of curves, one may not always be able to find a model that adequately fits the data at hand. Certain desired inferences, such as extrapolation below the lowest doses in the experiment, rely heavily on the dose-response shape determined by the parametric model.

As an alternative, one might seek approaches that also produce a smooth and strictly monotone curve for the dose-response function but alleviate some of the rigidity in shape of particular parametric models. To achieve greater flexibility in curve shape, one generally has to sacrifice some interpretability of model parameters. Still, for mixture applications, the goal is often fitting a smooth monotone dose-response curve. From such a curve, one can estimate any quantities of interest such as the ED_{50} or other benchmark doses even when a single parameter identified with the quantity of interest is not part of the model. On the other hand, low-dose extrapolation is even more uncertain with such models because their flexibility precludes using the rigid parametric model structure to help make inferences beyond the range of the available data. Various kinds of flexible smoothing models could be applied for dose-response modeling with choices dictated to some extent by the nature of the available data. Splines, which are piecewise polynomial models, could be useful in many dose-response settings (Harrell 2001; Ramsay 1988). The analyst chooses the number and location of knots (the dose levels where the polynomial pieces join), the degree of polynomial to employ, and sometimes a constraint on the function beyond the lowest and highest knots. For dose-response modeling, constraining the fitted spline to be monotone is an important consideration. For example, with respect to evaluating mixture data for departures from additivity,

Kelly and Rice (1990) used a monotone hybrid of smoothing and least-squares splines, where monotonicity is achieved by constraining coefficients and smoothing is controlled by a penalty parameter and by the number of knots. For data with a binary response, Dette et al. (2005) proposed a nonparametric method for obtaining monotone estimates of effective dose without requiring constrained optimization or function inversion. In a relative potency setting, Guardabasso et al. (1987) assumed that multiple chemicals share a common but arbitrarily shaped dose-response curve, which can be shifted or stretched along the log-dose scale to give chemical-specific curves; they model the common curve by a cubic spline (though without requiring monotonicity) and then apply chemical-specific shift and scale parameters. Nottingham and Birch (2000) combined parametric and nonparametric estimates of a dose-response function, with a mixing parameter that adjusts the relative weight given to each component based on how well it individually fits the data.

Another modeling strategy, one that is capable of generating quite flexible predictive models but one that has largely been unexplored for dose-response modeling, is model averaging. The idea is that one postulates a list of K possible dose-response model families, say $f_1(d|\boldsymbol{\theta_1})$, $f_2(d|\boldsymbol{\theta_2})$, $f_3(d|\boldsymbol{\theta_3})$, $\cdots$, $f_K(d|\boldsymbol{\theta_K})$, and fits each of the K models to the available data. The final predictive model is a weighted average of the best-fitting models, one from each family. Typically, model fitting is accomplished using a Bayesian paradigm where the weights are also estimated. Although we have not yet seen model averaging used in the context of mixture models, model averaging has been proposed by several authors as a way to estimate a benchmark dose that is not tied to any single parametric model family (Fang et al. 2015; Simmons et al. 2015; Wheeler and Bailer 2007, 2008, 2009) and to detect hormesis (Kim et al. 2016).

When examining whether a mixture is obeying some definition of additivity, there is a premium on accurate and precise estimation of the dose-response curves of the component chemicals because the predictions from those curves are combined to calculate the expected dose-response curve for the mixture under additivity. Hertzberg et al. (2013) proposed an approach based on guidance from the U.S. Environmental Protection Agency that assumes that all component chemicals are toxicologically similar and that the specific dose-response curve for each chemical comes from a common family of models; that is, the component-specific curves differ only in their specific parameter values, but not the form of $f(d|\boldsymbol{\theta})$. Those authors use model selection criteria (see Sect. 8.5) to select the simplest model family that still provides an adequate fit to the data for each component chemical. This approach is straightforward and likely sound if all component chemicals are indeed toxicologically similar. There is, however, no guarantee that a single model family will provide the best fit to the data on each component chemical. The Hertzberg et al. approach could suffer if the dose-response relationships of certain component chemicals were poorly fit by the common model family. In a less restrictive approach, Altenburger et al. (2000) specified a list of model families (such as described above for model averaging) and selected a best-fitting model for each component chemical across the list of model families. Thus, each component chemical could have a dose-response curve from a different model family, a strategy

which improves overall fit at the expense of greater complexity. A model averaging approach would, in principle, offer better response prediction for each component chemical than could any single model – but that potential advantage would come at the expense of a vastly more computationally intensive fitting procedure.

8.5 Analysis Considerations

As mentioned previously, a dose-response model is specified via three components: a probability distribution, a model $f(d|\boldsymbol{\theta})$ that describes the relationship between mean response and dose, and a model $\boldsymbol{\Sigma}(\boldsymbol{\omega})$ for the variance-covariance matrix of the data. The process of fitting a dose-response model to data is the process of estimating the values for the unknown parameters $\boldsymbol{\theta}$ in the mean model and $\boldsymbol{\omega}$ in the variance-covariance model that best fit the data. Denote the estimated values by $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\omega}}$, respectively**.** Although the details of the calculations differ depending on the assumed probability distribution for the data and the criterion employed in defining "best fit," statistical software will provide the estimates $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\omega}}$ together with estimates of the variances and covariances of those parameter estimates – as well as related quantities such as confidence intervals and test statistics. After fitting a particular model $f(d|\boldsymbol{\theta})$ to data from an experiment and having the estimate $\widehat{\boldsymbol{\theta}}$ available, one can estimate the true mean response under the model at any specified dose level d^* by calculating $f(d^*|\widehat{\boldsymbol{\theta}})$, that is, by substituting d^* and $\widehat{\boldsymbol{\theta}}$ into the relevant equation. The value $f(d^*|\widehat{\boldsymbol{\theta}})$ is called the predicted response (or predicted value) at dose d^*. The variance of these predicted responses can be estimated using $\widehat{\boldsymbol{\omega}}$**,** but details of the calculation differ depending on the particular dose-response model assumed. Again, many statistical software tools will provide these predicted values and their variances (or standard errors). Thus, fitting a dose-response model allows one to plot the estimated mean response trajectory as a function of dose and to construct confidence bands for and test hypotheses about that trajectory.

A well-known aphorism attributed to statistician GEP Box is "All models are wrong; but some are useful." With any regression models, and dose-response models are no exception, one seeks a model where the estimated dose-response trajectory closely mimics the trajectory of the observed data, and the assumed variance model and probability distribution are faithfully reflected by the observed data. If the assumed model does not satisfactorily reflect the data, then inferences based on that model are questionable and conclusions are potentially misleading. The model must be a sufficiently good approximation to be useful. Thus, a critical part of any dose-response analysis involves assessing the aptness of the model for the data at hand. Essentially, this process is one of model criticism – what are the good and bad aspects of the model in terms of being in accord with the data.

Inspecting residual plots is often a sensible first step in model criticism. Procedures for examining model assumptions through residual plots are commonplace for models that are linear in the parameters; for such linear models, the residuals have

two desirable features: (1) the variability in the residuals is a straightforward reflection of the variability in the data; and (2) the residuals and the predicted values are uncorrelated. A so-called raw residual, $\hat{\varepsilon}_{ij}$, is the difference between an observed response and its predicted response under the fitted model; thus, $\hat{\varepsilon}_{ij} = y_{ij} - f(d_i|\widehat{\boldsymbol{\theta}})$; sometimes these raw residuals are rescaled to have variance 1 by dividing each by some measure of variability. The raw residuals, and their various rescaled counterparts, are commonly used with linear models. Residuals with names such as Pearson residuals or deviance residuals are often used with models involving binomial or Poisson distributional assumptions (Agresti 2013), but the fundamental idea remains the same – a residual assesses how far an observed response is from the value expected under the fitted model.

What is less often recognized is that, for the nonlinear models common in dose-response modeling, raw residuals do not necessarily retain the features that make them useful for checking aptness of linear models via plotting. (Problems arise when a nonlinear model has high intrinsic curvature, a concept beyond the scope of the current presentation.) Instead, another concept of residuals, called projected residuals (Cook and Tsai 1985), can be used with nonlinear models to recover the desirable properties needed for using simple plots of residuals as diagnostic tools. Statistical software packages such as SAS provide projected residuals for plotting.

Consider modeling a continuous response, like enzyme activity, with $n > 1$ experimental units per dose level under the assumption that the response at each dose level is normally distributed about its mean with constant variance across dose levels. Aspects of model aptness can be examined by plotting the residuals appropriate for the type of model fitted against dose level d_i or against predicted response $f(d_i|\widehat{\boldsymbol{\theta}})$. If the assumed model for the mean fits well, the n residuals at each dose level should be centered near zero at every dose (Fig. 8.2a); departures where the residuals are centered above zero for some dose levels and below zero for others suggest that a different mean function may provide a better fit (Fig. 8.2b). If, in addition, the variance is constant as assumed, residuals at every dose level will exhibit approximately equal spreads (Fig. 8.2a); patterns where the spread in the residuals grows larger or smaller with increasing dose level provide evidence for heteroskedasticity (Fig. 8.2c) and may indicate that a transformation of the response or a more complex model that incorporates heteroskedasticity is needed. Moreover, a histogram of the residuals should reveal a symmetric distribution if the data are distributed normally.

Another common graphical diagnostic approach applied with linear regression is to look for influential observations – those that have an unusually large influence on parameter estimates or on predicted values when they are deleted from the data set – by plotting various statistics known collectively as influence diagnostics against a variable that identifies each observation. Again, as with residuals, inherent characteristics of nonlinear regression models that differ from those of linear models imply that some concepts used for influence diagnostics must be reinterpreted for use with nonlinear models (St. Laurent and Cook 1992, 1993).

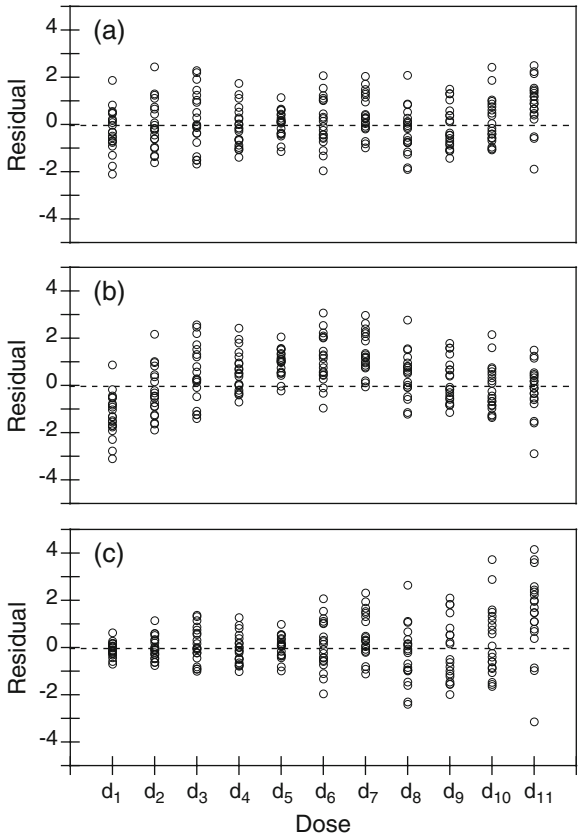

Fig. 8.2 Characteristic residual plots for a dose-response model whose errors ε_{ij} have a normal distribution. Underlying data have responses measured on 20 experimental units at each of 11 dose levels (d_1, d_2, d_3,. . ., d_{11}). Panel (**a**): data-generating model has homoskedastic errors and the fitted $f(d\,|\,\boldsymbol{\theta})$ was correctly specified; residuals have similar spreads and are vertically centered near zero at all dose levels. Panel (**b**): data-generating model has homoskedastic errors, but the fitted $f(d\,|\,\boldsymbol{\theta})$ was incorrectly specified; residuals have similar spreads at all dose levels, but their centers exhibit a non-horizontal trajectory (above zero at intermediate dose levels and below zero at lower or higher dose levels). Panel (**c**): data-generating model has heteroskedastic errors, but the fitted $f(d\,|\,\boldsymbol{\theta})$ was correctly specified; residuals are vertically centered near zero at all dose levels but have spreads that increase with dose level

In addition to informal visual inspections of residual plots, formal statistical tests can be applied as well. Perhaps the most basic test for aptness of the regression model $f(d|\boldsymbol{\theta})$ arises from comparing the fit of Eqs. 8.1 and 8.2. When each of the D distinct dose levels in a dose-response experiment has $n > 1$ experimental units assigned, the estimates of the set of μ_i in Eq. 8.1, denoted $\hat{\mu}_i$, are the average values of the observations at each dose level. Those values should be unbiased estimates of the true unknown responses at those dose levels (most accurate estimates available with the data at hand); the larger n, the more precise the estimates. The use of regression

model $f(d|\boldsymbol{\theta})$ to provide inference about responses at untested doses proceeds under the belief that it also provides unbiased estimates of the true unknown response at each tested dose level. On the other hand, if the mean responses at tested dose levels based on fitting $f(d|\boldsymbol{\theta})$ appear biased, then it may not be a useful model. Thus, one regards the regression model $f(d|\boldsymbol{\theta})$ as fitting the data well when its predicted responses at tested doses closely match the corresponding dose-specific means, i.e., when $f(d_i|\widehat{\boldsymbol{\theta}}) \approx \hat{\mu}_i$. This idea is the basis of a test of model fit that can be applied whenever the vector $\boldsymbol{\theta}$ contains fewer than D parameters. The general procedure is to fit both Eqs. 8.1 and 8.2 via maximum likelihood and construct a likelihood ratio test (Seber and Wild 1989) comparing the fit of Eq. 8.2 to that of Eq. 8.1. Rejection of the null hypothesis that both models fit equally well implies that the regression model $f(d|\boldsymbol{\theta})$ was unsuccessful in estimating the observed dose-specific mean responses and should be replaced with a different model. Of course, details of constructing the likelihood ratio test depend on the statistical distribution assumed for the responses.

For example, consider a study involving D distinct dose levels with n experimental units assigned to each dose and assume that the response has a normal distribution at each dose level. Then, the fitting of Eq. 8.1 amounts to conducting a one-way analysis of variance with the dose levels as the treatment groups. Taking $f(d|\boldsymbol{\theta})$ to be the Hill model of Eq. 8.3 where $\boldsymbol{\theta}$ contains four parameters, a least-squares-based test that is essentially equivalent to the likelihood ratio test comparing Eq. 8.2 to 8.1 would involve an F-statistic with $D - 4$ degrees of freedom in the numerator and $D \times (n - 1)$ in the denominator (Seber and Wild 1989).

The underlying principle used in the goodness-of-fit test described above for comparing Eqs. 8.1 and 8.2 can be applied to any pair of nested models to decide whether the smaller model fits as well as the larger. Consider two regression models, with $f_2(\cdot|\boldsymbol{\theta_2})$ being a nested sub-model of $f_1(\cdot|\boldsymbol{\theta_1})$. One way to think of a nested sub-model is that the parameter vector $\boldsymbol{\theta_2}$ of the sub-model contains the same parameters as $\boldsymbol{\theta_1}$ except that, in the sub-model, some of the parameters are fixed at specified values and do not need to be estimated. For example, consider the Hill model of Eq. 8.3 with parameter vector $\boldsymbol{\theta_1} = (L, U, \delta, \gamma)$. Another Hill model with the lower and upper asymptotes fixed at 0 and 1, respectively, would have parameter vector $\boldsymbol{\theta_2} = (0, 1, \delta, \gamma)$ and be nested within the first because the parameters to be estimated in $\boldsymbol{\theta_2}$ are a subset of those to be estimated in $\boldsymbol{\theta_1}$. When one model is a special case of another, the one with more parameters is more flexible and will necessarily fit at least as well as the one with fewer parameters; however, if the difference in fit is negligible, the simpler model with fewer parameters would typically be preferred based on parsimony. Formal statistical tests such as likelihood ratio tests can be used to decide whether or not fixing a subset of the parameters at specified values degrades model fit (Seber and Wild 1989).

Comparing two regression models that are not nested requires a different strategy. Formal tests to compare non-nested models are rarely used in toxicology; instead, one chooses the "better" model by using a model selection criterion. Because models

with more parameters might be expected to fit better than models with fewer parameters, model selection criteria typically make adjustments for the number of parameters in the model. The general procedure is to choose a model selection criterion and calculate its value for each candidate model under consideration. Then select as the best model the one with the largest value (or sometimes smallest, depending on the particular criterion used) of the criterion. A great variety of such criteria are in use. The coefficient of determination (R^2), or a version of it adjusted for the number of parameters in the model, selects according to the proportion of variation in the data accounted for by the fitted model. Other commonly used criteria, such as the Akaike information criterion (AIC) (Akaike 1974) or the Bayesian information criterion (BIC) (Schwarz 1978; Montgomery et al. 2012), evaluate goodness of fit through the likelihood of the observed data but impose a penalty that increases with the number of model parameters. Thus, if two models produced the same likelihood, these criteria would favor the one with fewer parameters. In experiments that involve very few observations, one might prefer the AICc (Burnham and Anderson 2002), a version of the AIC with a correction for small sample sizes. Model selection criteria such as these are useful adjuncts to strategies for modeling mixture components, such as those of Hertzberg et al. (2013) and Altenburger et al. (2000), which involve selecting best-fitting models. On the other hand, even the best-fitting among a list of candidate regression models may exhibit important lack of fit when compared to fitting the dose-specific mean responses via Eq. 8.1.

When the model $f(d|\boldsymbol{\theta})$ for the mean response shows evident lack of fit, how should it be remediated? An obvious answer is to choose a different model with superior fit – but that may be easier said than done. Experience with the particular response or assay may suggest alternative models to try; similarly, a literature search or reaching out to colleagues might turn up alternatives. A plot with the fitted dose-response model overlaying the observed data sometimes reveals the main discrepancies between model and data, thereby suggesting modifications to improve the model – including perhaps changing the dose metric. Such a plot or residual plots may instead reveal "unusual" or "influential" data values that adversely affect model fit, initiating careful scrutiny of the validity or correctness of those data points. If efforts to find a better-fitting parametric model fail, one could consider more flexible modeling approaches such as splines or model averaging, as mentioned earlier. Another consideration is whether the model is useful for its intended purpose despite some lack of fit. For example, suppose model fit suffers mainly at high doses but is satisfactory at lower doses. If inferences at lower doses are the primary use for the model, perhaps the formally ill-fitting model will yield useful information – interpretation should be cautious, however: lack of fit in the mean model can distort variance estimates so that, for example, confidence interval coverage may suffer even at the low doses where the mean model fits well. Although the use of a model that exhibits substantial lack of fit is undesirable, it is occasionally unavoidable; in those unavoidable cases, one should clearly acknowledge evident lack of fit in reporting the results of the data analysis.

Although the primary focus is often on the mean response, a full dose-response model must also properly model the variability of the data around the mean. This variability is partly described by the nature of the probability distribution employed and partly by the model $\boldsymbol{\Sigma}(\boldsymbol{\omega})$ for the variance-covariance matrix.

As mentioned previously, the nature of the response (binary, count, or continuous) and experience often is sufficient for properly specifying an appropriate probability distribution. Nevertheless checking distributional assumptions is always useful. Histograms or *Q*-*Q* plots (Wilk and Gnanadesikan 1968) of residuals can reveal deviations from an assumed distributional shape, particularly for continuous responses. Formal procedures for testing distributional assumptions are available. Some, such as the Kolmogorov-Smirnov test or the Anderson-Darling test, are general purpose (Stephens 1974); others, such as the Shapiro-Wilks normality test, are directed toward particular distributions (Stephens 1974).

Building models for the variance is often a complex undertaking and well beyond the scope of this chapter – so the remarks here merely scratch the surface. As mentioned earlier, for distributions like the binomial or Poisson, the variance is a function of the mean, so the variance model is partly preset. Additional variance parameters are introduced only when the preset model proves inadequate for the data. When the normal distribution is the relevant probability model, often the default assumption is that the variance is constant and governed by a single parameter. Nonconstant variance is possible; it can sometimes be stabilized by a transformation of the response variable to achieve constant variance (Bates and Watts 1988; Bickel and Doksum 1977) or, alternatively, be modeled as a parametric function of dose. For complex experimental designs that involve more structure than a completely randomized design – such as multiple batches of experimental units or an experiment that is carried out in blocks over several days – modeling the variance often requires the use of several variance parameters. For example, the model might need a parameter for variance in response among units in a single batch and one for variance among batches. If the overall variability can be partitioned into such components, statistical analysis implements a so-called mixed model approach that allows simultaneous estimation of mean parameters and multiple variance parameters (Searle et al. 2006).

Another kind of departure from assumptions arises if observations are correlated instead of being independent as many models assume. Such correlations typically arise from block structure in the experimental design or from certain pre-analysis data manipulations. One common practice is to rescale a continuous response so that its limits are 0 and 1 on a probability scale or 0 and 100 on a percent scale. If the largest responses occur at dose zero, one might rescale by dividing all responses by the mean response among the negative controls so that on average the responses have an upper limit of 100% (Crofton et al. 2005; Hertzberg et al. 2013). Although such rescaling has intuitive appeal, in principle, dividing several responses by the same random quantity (average of negative control responses) induces correlations among them that contradict independence – because the rescaled responses all depend on the same mean response among controls. Particularly when the rescaling is done separately for different runs that are part of the same experiment (using

run-specific control means), the analysis should arguably account for this dependence. Again, mixed models can be constructed to account for correlations. Alternatively, with likelihood-based estimation methods, statisticians have developed alternative estimators for the standard errors of parameter estimates that adjust for such correlations. These estimators, known as "sandwich estimators" (Kauermann and Carroll 2001; Freedman 2006), are also useful in settings with heteroskedasticity. In situations where appropriate variance estimators are difficult to derive theoretically, bootstrap methods or other resampling methods can be used to estimate standard errors appropriately (Efron and Tibshirani 1993).

In evaluating model fit, it is important to recognize that all three components of the dose-response model play off one another. If the probability model assumes a symmetric distribution but the actual data distribution is skewed, the variance model may be (or appear to be) misspecified. If the mean function fails to fit the data well, that could also impinge on the diagnostics for evaluating the variance model or the distribution. In evaluating the aptness of a model, one must keep in mind that all three components work together.

8.6 Computational Issues

Many dose-response models used in toxicology require iterative computational methods for model fitting. Whether the criterion for best fit is least squares or maximum likelihood, estimation for models that are nonlinear in the parameters uses computational algorithms that approach the best fit incrementally through successive cycles of computation. The iterations stop when an additional cycle fails to improve the fitting criterion by a preset amount – in which case the algorithm is said to have converged. For maximum likelihood, one can think of the process as analogous to finding the highest point in a landscape (for least squares, the lowest); only the dimension of the "landscape" – which depends on the number of parameters in the model – is often higher than three. Also, the algorithms cannot just look around and see the highest point and head toward it; they must use clues available at the current location, such as steepness and uphill direction, to determine which way and how far to go for the next iteration. When the topology is complicated, the algorithms have trouble finding the maximum. A long and nearly flat ridge makes finding the maximum difficult. Sometimes an algorithm "falls off a cliff" and has difficulty climbing back. Multiple nearby peaks of different heights also make finding the unique maximum difficult. All these issues may lead to failure of an algorithm to converge.

One common cause for failure to converge is a design that is not well suited to the model at hand, as was more fully discussed earlier. Thus, an appropriate choice of the number and spacing of dose levels goes a long way toward avoiding convergence problems in model fitting. Even with a sound design, however, nonlinear models can be quirky to fit. Sometimes a model can be reparameterized so that the numerical properties of the fitting algorithms are improved while the trajectory of predicted

responses remains unchanged. Although the best way to parameterize a given model is not always obvious, some choices work better than others. A reparameterization of the Hill model to improve numerical performance was described earlier (see Sect. 8.4). In addition to the mean response, the overall model may involve one or more variance parameters, and reparameterizing the variance may also help in some situations. For example, rather than directly estimating the variance or even the standard deviation, the logarithm of the standard deviation may work better, possibly due to its scale being more similar to that of the mean parameters. Sometimes centering dose levels can enhance convergence and lessen variability of parameter estimates by reducing multicollinearity (Reeve and Turner 2013). All iterative algorithms must start at some initial set of guesses at the parameter values and proceed iteratively to improve those estimates – but convergence can be highly dependent on the choice of starting values. Ill-chosen values can lead to non-convergence. Moreover, in situations where the likelihood surface has multiple peaks, different starting values can lead to seemingly successful convergence to estimates that represent different local maxima – but the goal is to find the global maximum. Even if convergence appears to have been achieved, it is good practice to try multiple distinct starting values and confirm that all produce the same ultimate estimates. When fitting closely related dose-response curves simultaneously to multiple chemicals that differ widely in potency, one simple but effective step that frequently helps with convergence issues is to rescale the doses of each chemical, so corresponding parameter estimates will have similar magnitudes across chemicals. For example, in simultaneously fitting Hill models to two chemicals, the first with an ED_{50} near 0.005 mg/kg and the second with an ED_{50} near 5.0 mg/kg, convergence might be improved if the dose levels of the first chemical were rescaled to μg/kg, so both ED_{50} values were near 5.0 in their respective units. Of course, the resulting estimates and confidence limits could be reexpressed in any common units desired. Finally, there are usually several different computational algorithms that can be used to estimate the parameters of a given model. For example, the NLIN procedure in SAS offers four choices: steepest descent (or gradient), Newton, modified Gauss-Newton, and Marquardt. Some methods may work better than others for a given set of data, so if one algorithm fails to converge, try another. Also, most algorithms involve preset constants that control aspects of the algorithm; sometimes adjusting these "tuning parameters" helps with computational issues.

8.7 Summary

Dose-response modeling is an important data analysis tool throughout toxicology, particularly so in evaluating chemical mixtures. This chapter provides a broad introductory overview to statistical issues that arise in studying dose-response relationships. Statistical dose-response models consist of a probability distribution for the response, a function $f(d|\boldsymbol{\theta})$ to describe the relationship between the mean response and dose, and a model $\boldsymbol{\Sigma}(\boldsymbol{\omega})$ for the variance-covariance matrix of the data.

The chapter discusses the role of statistical study design, including the roles of dose placement and of properly accounting for multiple sources of variation or correlations among observations, in achieving accurate and precise parameter estimation and efficient hypothesis testing. It discusses the choice of a functional form for $f(d|\boldsymbol{\theta})$ and describes strategies for examining the adequacy of the proposed dose-response model. Finally, the chapter considers some ways to cope with the failure of iterative model-fitting algorithms to converge to a unique solution. Careful attention to study design and the use of statistical models that are appropriate for the data at hand are critical for achieving the best possible results from dose-response studies.

Acknowledgments This work was supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences.

References

Agresti, A. 2013. *Categorical data analysis*. 3rd ed. Hoboken: Wiley.

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19 (6): 716–723.

Altenburger, R., T. Backhaus, W. Boedeker, M. Faust, M. Scholze, and L.H. Grimme. 2000. Predictability of the toxicity of multiple chemical mixtures to vibrio fischeri: Mixtures composed of similarly acting chemicals. *Environmental Toxicology and Chemistry* 19 (9): 2341–2347.

Bates, D.M., and D.G. Watts. 1988. *Nonlinear regression analysis and its applications*. New York: Wiley.

Berenbaum, M.C. 1985. The expected effect of a combination of agents: The general solution. *Journal of Theoretical Biology* 114: 413–431.

Bernstein, L., J. Kaldor, J. McCann, and M.C. Pike. 1982. An empirical approach to the statistical analysis of mutagenesis data from the Salmonella test. *Mutation Research* 97: 267–281.

Bickel, P.J., and K.A. Doksum. 1977. *Mathematical statistics: Basic ideas and selected topics*. San Francisco: Holden-Day.

Box, G.E.P., and D.R. Cox. 1964. An analysis of transformations. *Journal of the Royal Statistical Society, Series B* 26 (2): 211–252.

Box, G.E.P., and G.C. Tiao. 1992. *Bayesian inference in statistical analysis*. New York: Wiley.

Breslow, N.E. 1984. Extra-poisson variation in log-linear models. *Applied Statistics* 33 (1): 38–44.

Burnham, K.P., and D.R. Anderson. 2002. *Model selection and multimodel inference: A practical information-theoretic approach*. 2nd ed. New York: Springer.

Carlin, B.P., and T.A. Louis. 2000. *Bayes and empirical Bayes methods for data analysis*. 2nd ed. Boca Raton: Chapman & Hall.

Casey, M., C. Gennings, W.H. Carter, V.C. Moser, and J.E. Simmons. 2004. Detecting interaction(s) and assessing the impact of component subsets in a chemical mixture using fixed-ratio mixture ray designs. *Journal of Agricultural, Biological, and Environmental Statistics* 9 (3): 339–361.

Christensen, E.R., and N. Nyholm. 1984. Ecotoxicological assays with algae: Weibull dose-response curves. *Environmental Science & Technology* 18 (9): 713–718.

Cook, R.D., and C.L. Tsai. 1985. Residuals in nonlinear regression. *Biometrika* 72 (1): 23–29.

Crofton, K.M., E.S. Craft, J.M. Hedge, C. Gennings, J.E. Simmons, R.A. Carchman, W.H. Carter, and M.J. DeVito. 2005. Thyroid-hormone-disrupting chemicals: Evidence for dose-dependent additivity or synergism. *Environmental Health Perspectives* 113 (11): 1549–1554.

Dette, H., N. Neumeyer, and K.F. Pilz. 2005. A note on nonparametric estimation of the effective dose in quantal bioassay. *Journal of the American Statistical Association* 100 (470): 503–510.
Dinse, G.E., and D.M. Umbach. 2011. Characterizing non-constant relative potency. *Regulatory Toxicology and Pharmacology* 60: 342–353.
Efron, B., and R. Tibshirani. 1993. *An introduction to the bootstrap*. Boca Raton: CRC Press.
EPA (Environmental Protection Agency). 2016. The ToxCast analysis pipeline: An R package for processing and modeling chemical screening data. Date of access: 12 December 2017. https://www.epa.gov/sites/production/files/2015-08/documents/pipeline_overview.pdf.
Fang, Q., W.W. Piegorsch, S.J. Simmons, X. Li, C. Chen, and Y. Wang. 2015. Bayesian model-averaged benchmark dose analysis via reparameterized quantal-response models. *Biometrics* 71 (4): 1168–1175.
Finney, D.J. 1971. *Probit analysis*. Cambridge: Cambridge University Press.
Freedman, D.A. 2006. On the so-called "Huber sandwich estimator" and "robust standard errors". *The American Statistician* 60 (4): 299–302.
Guardabasso, V., D. Rodbard, and P.J. Munson. 1987. A model-free approach to estimation of relative potency in dose-response curve analysis. *The American Journal of Physiology* 252 (3): E357–E364.
Guardabasso, V., P.J. Munson, and D. Rodbard. 1988. A versatile method for simultaneous analysis of families of curves. *The FASEB Journal* 2 (3): 209–215.
Harrell, F.E. 2001. *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*. New York: Springer.
Haseman, J.K., and M.D. Hogan. 1975. Selection of the experimental unit in teratology studies. *Teratology* 12 (2): 165–171.
Hertzberg, R.C., Y. Pan, R. Li, L.T. Haber, R.H. Lyles, D.W. Herr, V.C. Moser, and J.E. Simmons. 2013. A four-step approach to evaluate mixtures for consistency with dose addition. *Toxicology* 313: 134–144.
Hill, A.V. 1910. The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. *The Journal of Physiology* 40 (Suppl): iv–vii.
Hunt, D.L., and D. Bowman. 2004. A parametric model for detecting hormetic effects in developmental toxicity studies. *Risk Analysis* 24 (1): 65–72.
Kauermann, G., and R.J. Carroll. 2001. A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association* 96 (456): 1387–1396.
Kelly, C., and J. Rice. 1990. Monotone smoothing with application to dose-response curves and assessment of synergism. *Biometrics* 46 (4): 1071–1085.
Kim, S.B., S.M. Bartell, and D.L. Gillen. 2016. Inference for the existence of hormetic dose-response relationships in toxicology studies. *Biostatistics* 17 (3): 523–536.
Margolin, B.H., N. Kaplan, and E. Zeiger. 1981. Statistical analysis of the Ames Salmonella/microsome test. *Proceedings of the National Academy of Sciences of the United States of America* 78 (6): 3779–3783.
Montgomery, D.C., E.A. Peck, and G.G. Vining. 2012. *Introduction to linear regression analysis*. 5th ed. Hoboken: Wiley.
Mood, A.M., and F.A. Graybill. 1963. *Introduction to the theory of statistics*. 2nd ed. New York: McGraw Hill.
Nottingham, Q.J., and J.B. Birch. 2000. A semiparametric approach to analyzing dose-response data. *Statistics in Medicine* 19: 389–404.
Piegorsch, W.W., and J.K. Haseman. 1991. Statistical methods for analyzing developmental toxicity data. *Teratogenesis, Carcinogenesis, and Mutagenesis* 11: 115–133.
Ramsay, J.O. 1988. Monotone regression splines in action. *Statistical Science* 3 (4): 425–461.
Reeve, R., and J.R. Turner. 2013. Pharmacodynamic models: Parameterizing the Hill equation, Michaelis-Menten, the logistic curve, and relationships among these models. *Journal of Biopharmaceutical Statistics* 23: 648–661.
Schwarz, G.E. 1978. Estimating the dimension of a model. *Annals of Statistics* 6 (2): 461–464.
Searle, S.R., G. Casella, and C.E. McCulloch. 2006. *Variance components*. New York: Wiley.

Seber, G.A.F. 1977. *Linear regression analysis*. New York: Wiley.
Seber, G., and C. Wild. 1989. *Nonlinear regression*. New York: Wiley.
Shockley, K.R. 2016. Estimating potency in high-throughput screening experiments by maximizing the rate of change in weighted Shannon entropy. *Scientific Reports* 6: 27897. https://doi.org/10.1038/srep27897.
Simmons, S.J., C. Chen, X. Li, Y. Wang, W.W. Piegorsch, Q. Fang, B. Hu, and G.E. Dunn. 2015. Bayesian model averaging for benchmark dose estimation. *Environmental and Ecological Statistics* 22: 5–16.
St. Laurent, R.T., and R.D. Cook. 1992. Leverage and superleverage in nonlinear regression. *Journal of the American Statistical Association* 87 (420): 985–990.
St. Laurent, R.T., and R.D. Cook. 1993. Leverage, local influence and curvature in nonlinear regression. *Biometrika* 80 (1): 99–106.
Stephens, M.A. 1974. EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association* 69 (347): 730–737.
Wheeler, M.W., and A.J. Bailer. 2007. Properties of model-averaged BMDLs: A study of model averaging in dichotomous response risk estimation. *Risk Analysis* 27 (3): 659–670.
———. 2008. Model averaging software for dichotomous dose response risk estimation. *Journal of Statistical Software* 26 (5): 1–15. https://doi.org/10.18637/jss.v026.i05.
———. 2009. Comparing model averaging with other model selection strategies for benchmark dose estimation. *Environmental and Ecological Statistics* 16: 37–51.
Wilk, M.B., and R. Gnanadesikan. 1968. Probability plotting methods for the analysis of data. *Biometrika* 55 (1): 1–17.
Williams, D.A. 1982. Extra-binomial variation in logistic linear models. *Applied Statistics* 31 (2): 144–148.
Zorrilla, E.P. 1997. Multiparous species present problems (and possibilities) to developmentalists. *Developmental Psychobiology* 30 (2): 141–150.