International Journal of Statistical Sciences *ISSN 1683–5603*
Vol. 11(Special Issue), 2011, pp 189-206

Dose-Response Modeling For Continuous Responses: Alternative Variance Models

John F. Fox [1]
National Center for Environmental Assessment
US Environmental Protection Agency
Washington, D.C. 20460, USA

Evelyn Rochelle Frazier and Bimal Sinha
Department of Mathematics and Statistics
University of Maryland, Baltimore County
Baltimore, MD 21250, USA

[Received December 28, 2010; Revised April 16, 2011; Accepted May 25, 2011]

Abstract

In the context of modeling dose-response data for continuous outcomes, a common procedure is to summarize the data based on sample means and sample variances, and fit a suitable model for the means, $\mu(d)$, as a function of dose (d), assuming either a constant variance model or a power model for variance such as $\sigma^2 = \alpha[\mu(d)]^\rho$. However, in some cases the standard deviation is neither constant nor well modeled using a power function of the mean, and we find that practical inference for the benchmark dose can be based on a model with mean and standard deviation as distinct functions of dose. Herein, we explore the idea of fitting $\mu(d)$ and $\sigma(d)$ as separate functions of dose d, and use them to infer the benchmark dose (BMD) and its lower confidence limit (BMDL). We present several examples to demonstrate that this alternative approach has advantages over the existing practice in some particular cases.

Keywords and Phrases: benchmark dose (BMD), dose-response model, maximum likelihood estimate, profile likelihood.

AMS Classification: 62P10.

[1]The views expressed in this report are those of the authors and do not necessarily reflect the views or policies of the EPA. The U.S. Environmental Protection Agency (EPA) partially funded and collaborated in the research described here under Inter-Agency Agreement No. DW-89-92298301-0 with DOE, under which Dr. Bimal Sinha held an ORISE fellowship at EPA. The publisher acknowledges that this contribution is not subject to copyright protection in the United States, because it is a "United States Government Work"as described in the U.S. Copyright Act.

Introduction

This report considers problems that arise occasionally when fitting dose-response models to continuous response data and when it is assumed that variance is functionally related to the mean. While there are many reasonable choices for modeling the mean as a function of dose d, σ is often taken as either a constant or a suitable power of the mean rather than a direct function of dose (Filipsson et al., 2003; USEPA, 2006) in dose-response modeling. In what follows, we compare models for σ that are functions of dose with the customary model, $\sigma^2 = \alpha[\mu(d)]^\rho$, using data sets that were not well described by the customary model.

For modeling a continuous response Y recorded at a given dose d, it is generally assumed that $Y \sim N[\mu(d|\theta), \sigma^2(\mu|\psi)]$. The underlying likelihood based on the assumed normal data is used to compute the maximum likelihood estimates (MLEs) of the parameters θ and ψ, and subsequently to infer the benchmark dose (BMD, also denoted herein by d^*).

The likelihood function (apart from a constant, and here with σ a function of dose d) is given by

$$L(\theta, \psi|data) \sim [\prod_{i=0}^{k} \sigma_\psi(d_i)^{-n_i}] e^{-\frac{1}{2}\sum_{i=0}^{k}[n_i(\bar{y}_i - \mu_\theta(d_i))^2 + (n_i-1)s_i^2]/\sigma_\psi^2(d_i)}. \quad (1)$$

While for special cases (e.g., constant variance) the parameters can be obtained by least squares, in general for nonlinear models the maximum likelihood solution must be obtained iteratively.

The nature of data for a continuous dose-response model is as follows. At the dose level d_i, n_i observations are recorded: y_{ij}, $j = 1, \cdots, n_i$, $i = 1, \cdots, k$, along with a set of n_0 observations recorded at the control dose 0. Often instead of the raw data, summary statistics such as sample means $\bar{y}_0, \cdots, \bar{y}_k$ and sample standard deviations $s_0, \cdots, s_k$ are reported. As mentioned before, normality of the data observed at dose d_i with mean $\mu(d_i)$ and standard deviation $\sigma(d_i)$ is usually assumed. Herein, we use the conventional definition of sample standard deviation (sd) such that $(n-1)(sd)^2 \sim \sigma^2 \times \chi^2_{n-1}$.

While a rich variety of models (power, Hill, exponential, polynomial; Slob, 2002; Filipsson et al., 2003; USEPA, 2006) has been developed for modeling $\mu(d)$, the current practice is to model the standard deviation σ as either a constant or a power of the mean $\mu(d)$. This may not work properly in some dose-response scenarios, and hence a choice of the *direct* functional relationship between $\sigma(d)$ and d may be advantageous. Following common statistical practice, the sample means $[\bar{y}_0, \cdots, \bar{y}_k]$ and the sample standard deviations $[s_0, \cdots, s_k]$, observed at different dose levels, should be used to shed some light on the relation between μ and σ and their dependence on the dose rather than assuming *à priori* that σ is either a constant or a power of the mean $\mu(d)$.

Towards this end, a first step is to plot the sample means against dose and also the sample standard deviations against dose. For most reasonable sample sizes, the sample mean and the sample standard deviation being consistent estimates of the population

mean and population standard deviation, respectively, these graphs should suggest the nature of their dependence on the dose, albeit with sampling fluctuations. We have also plotted the logarithm of the sample standard deviation against the logarithm of the sample mean. The familiar power law: $\sigma^2 = \alpha[\mu(d)]^\rho$ can be justified only if this plot reveals a linear relationship at least approximately.

The BMD d^* may be defined[2] as the solution of the equation: $\mu(d^*) = \mu(0) \pm \sigma(0)$ with a positive (negative) slope of $\mu(d)$ at 0. Naturally, d^* is unknown, being a function of the parameters involved in the functions $\mu(d)$ and $\sigma(d)$. Here $d = 0$ corresponds to the *control* dose.

Our procedure is to draw inference about the BMD as follows. Assume

$$\mu(d) = \mu(d|\theta), \;\; \sigma(d) = \sigma(d|\psi) \tag{2}$$

where θ and ψ are two distinct sets of unknown parameters. Therefore, the functions $\mu(d)$ and $\sigma(d)$ can be specified using a variety of available models. The normal likelihood, coupled with the parametric forms of $\mu(d|\theta)$ and $\sigma(d|\psi)$, is then used for estimation of θ and ψ. The ordinary least squares method can be used to obtain initial estimates of the model parameters. Having obtained MLEs for the model parameters, we next solve for the BMD, d^*, from its defining equation:

$$\mu(d^*) = \mu(0|\theta) \pm \sigma(0|\psi) \tag{3}$$

resulting in $d^*(\theta, \psi)$. In the above, $+$ or $-$ is used depending on whether $\mu(d)$ is increasing or decreasing in d at $d = 0$.[3] A point estimate of $d^*(\theta, \psi)$ is then readily obtained as

$$\widehat{[d^*(\theta, \psi)]} = d^*(\hat{\theta}, \hat{\psi}). \tag{4}$$

The familiar Wald method (Rao, 1973) could be used to derive symmetric confidence intervals for BMD. However, we will use the preferred method based on profile likelihood (Crump and Howe, 1985). The lower confidence limit of BMD is denoted BMDL. In what follows, we always use the 90% (2-sided) profile confidence interval, and treat the BMDL as having nominal 1-sided coverage 0.95 (Crump and Howe, 1985; USEPA, 2006).

[2]This is only one of several useful ways to define the BMD (USEPA, 2006). Here we specify the benchmark response as the control mean, $\mu(0)$, plus or minus the control standard deviation, $\sigma(0)$.

[3]There may be more than one solution when $\mu(d)$ is not monotone in d

Examples

In this section we present several data sets and explain our suggested procedure. For all of these data sets, both Bartlett's test and the BMDS likelihood ratio test "Test 2" indicated heterogeneous variances. We assumed normality; this assumption was not tested because only grouped data (i.e., dose-group sample means and standard deviations) were available. In all cases, parameters of $\mu(d)$ were unconstrained (in BMDS, it is customary to constrain means model parameters to be either non-negative or non-positive).

DATA SET 1

This data set came from George et al. (1986), who studied the effects of trichloroethylene (TCE) on reproduction and fertility in F344 rats. The rats were exposed to TCE in their feed for 18 weeks; dose has units mg/(kg-day). Table 1a gives summary statistics for litter sizes produced by the female rats.[4] The pattern of variation is complex. There was variation of litter size among litters within a female and among females; females differed in number of litters, number of litters declined with dose, and there is some indication that litter size varied with number of litters.

Table 1a. Mean litter size for female rats exposed to TCE in feed

dose (d)	0	72	186	389
sample size	39	20	20	20
sample mean	10.36	10.09	9.39	8.66
sample sd	2.248	1.52	1.565	2.862

Figure 1 shows plots of the sample means versus dose, sample standard deviations versus dose, and $\log(sd)$ versus $\log(\bar{y})$. It is clear from these figures that the standard deviation does not seem to be either a constant or to follow the routinely used power law. On the other hand, a quadratic model seems to fit σ quite well. Either a quadratic or linear model can describe μ well; based on parsimony, one would select a linear model. The quadratic model for the means is also shown to illustrate the sort of tradeoff (between fitting μ and σ) that can occur when σ is modeled as a function of μ and when the power law does not hold (compare upper and lower rows in Figure 1).

The quadratic model is

$$\begin{aligned} \mu(d) &= \theta_0 + \theta_1 d + \theta_2 d^2 \\ \sigma(d) &= \psi_0 + \psi_1 d + \psi_2 d^2. \end{aligned} \tag{5}$$

Using the normal likelihood given by (4), the MLEs of the parameters θ and ψ can

[4]The mean litter size was determined for each female rat. Females bore 1-5 litters during the study. The mean and standard deviation of these per-female means were reported by George et al. (1986) for each dose level.

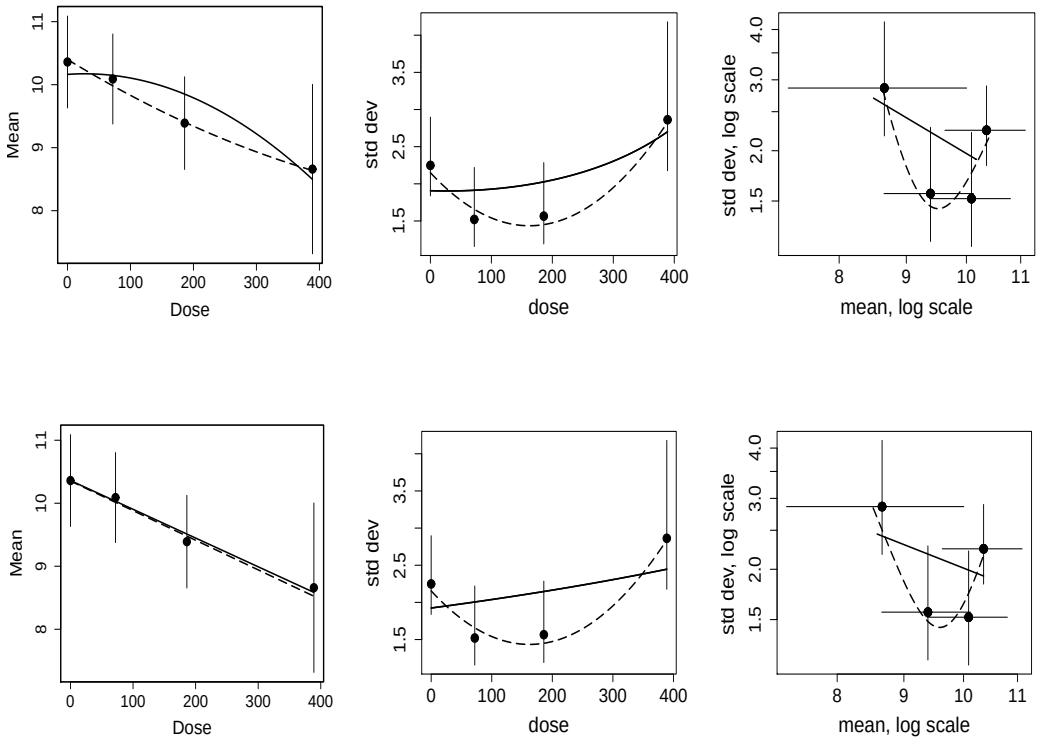

Figure 1: Quadratic (top row) and linear (bottom row) models for the means. Solid line: BMDS model, with power model for $\sigma(\mu)$. Dashed line: New model, with quadratic model for $\sigma(d)$. Error bars show 95% confidence intervals.

be readily obtained. The BMD d^* which is a solution of (3) is given by

$$d^* = \frac{-\theta_1 - [\theta_1^2 - 4\theta_2\psi_0]^{1/2}}{2\theta_2}. \tag{6}$$

The MLE of d^* is obtained by plugging in the MLEs of the relevant parameters. The BMD for the linear model is -ψ_0/θ_1.

Here are the maximum likelihood estimates and related quantities under the quadratic means model:

$$\begin{aligned} \hat{\theta}_0 = 10.400,\ \ \hat{\theta}_1 = -0.00610,\ \ \hat{\theta}_2 = 4.0402e-06, \\ \hat{\psi}_0 = 2.1500,\ \ \hat{\psi}_1 = -0.008800,\ \ \hat{\psi}_2 = 2.7106e-05, \\ BMD = 561,\ \ BMDL = 270,\ \ AIC = 248. \end{aligned} \tag{7}$$

The maximum likelihood estimates and related quantities under the linear means model are:

$$\begin{aligned} \hat{\theta}_0 = 10.3500,\ \ \hat{\theta}_1 = -0.0044606, \\ \hat{\psi}_0 = 2.1500,\ \ \hat{\psi}_1 = -0.008800,\ \ \hat{\psi}_2 = 2.7106e-05, \\ BMD = 459,\ \ BMDL = 327,\ \ AIC = 246. \end{aligned} \tag{8}$$

To infer a lower confidence bound for the BMD d^* under the profile likelihood method, upon substitution of θ_1 by $-\frac{\psi_0}{d^*}-\theta_2 d^*$, thus reducing the effective number of parameters from six to five, we rewrite the likelihood function (4) and maximize it with respect to the remaining parameters, obtaining the log-likelihood value, LL_{d^*}. The constrained maximization was carried out for various values of d^*, to find a close approximation to the root[5] of $LL_{d^*} - LL_{mle} - 0.5\chi^2_{1,0.90}$.

Table 1b. Comparison of models

Models for μ and σ	AIC	BMD	BMDL	P(V)	P(M)
BMDS: quadratic(d), power(μ)	253	414	282	0.017	0.546
BMDS: quadratic(d), constant	254	514	233	0.010	0.809
NEW: quadratic(d), quadratic(d)	248	561	270	0.418	0.749
BMDS: linear(d), power(μ)	252	422	262	0.017	0.530
BMDS: linear(d), constant	252	470	305	0.010	0.949
NEW: linear(d), quadratic(d)	246	459	327	0.418	0.922

Observation 1.1. First, we compare our results with those obtained using the power model for σ from the BMDS software (Table 1b). The P-values for the new models are based upon the likelihood ratio tests (LRTs) used in BMDS, with Test 3 and Test 4 (corresponding to P(V) and P(M)) being tests of fit for the variance and means models, respectively. The observed standard deviations are not well described by either model provided in BMDS (Fig.1, Table 1b) and require a non-monotonic model (a quadratic function of either dose or μ would serve in this case).

Observation 1.2. This data set illustrates how some data sets (having a non-monotone relation between sample mean and sd) may force an undesirable compromise

[5]We used the univariate root-solver, *uniroot*, in R

between fitting μ and σ when σ is modeled as a monotone function of μ. Referring to Figure 1 (top row), there was a pronounced distortion in the fit of the quadratic model for the means when σ was modeled as a function of μ; a slightly better fit for σ was obtained at the cost of a worse fit for μ. A similar effect for μ has been seen with a few other data sets.

Observation 1.3. When σ is modeled as a function of μ, the BMDS models might be rejected because the variance is not adequately fitted (based upon the LRTs, Table 1b), yet a linear dose-response relation fits the means well (Fig. 1). This problem is illustrated below for two more data sets. This points to the need for a more flexible approach to modeling the variance, such as that illustrated above.

DATA SET 2

This data set (Schlosser et al., 2003, Table II) represents a measure of cell proliferation in the nasal tissues of rats exposed to formaldehyde by inhalation. The "dose" in Table 2a is average concentration in the air inhaled, in ppm. The index of cell proliferation is termed ULLI (for unit length labeling index).

The design used 6 rats at each combination of exposure concentration and exposure duration. The eight exposure durations ranged from one day to 78 weeks. For each rat, the index was measured at 6 nasal sites and averaged. Indices at the four earliest exposure durations were adjusted by the authors to account for a difference of measurement method. A time-weighted average of the per-rat values was constructed by the authors using exposure durations, producing the means and standard deviations in Table 2a. The labeling index (at each nasal site within each rat) consisted of a count of number of radio-labeled cells divided by length of basal membrane examined microscopically (the count data are not reported). Thus, there is reason to expect the variance to increase with the mean, but the sources of variation are evidently complex.

Table 2a. Cell proliferation index 'ULLI' in nasal tissues of rats

dose (d)	0	0.7	2	6	10	15
n	48	46	47	48	48	47
mean	10.9	8.2	7.7	15.0	43.8	70.7
sd	3.2	2.3	2.7	15.6	17.6	19.4

Figure 2 shows the plots of the sample means versus dose, sample standard deviations versus dose, and $\log(sd)$ versus $\log(\bar{y})$. For the means, a cubic model fit substantially better than quadratic or linear models (based upon AIC). For the standard deviations, it seems that the power law between $\mu(d)$ and $\sigma(d)$ may hold (see Figure 2, right panel), but with large deviations owing to the nearly constant and low sd for the first three doses, and the high sd for the highest three doses.

We tried two additional models for σ, as follows. Our first alternative was to fit a two-component (piecewise-constant) variance model for $\sigma(d)$ with a cubic model for

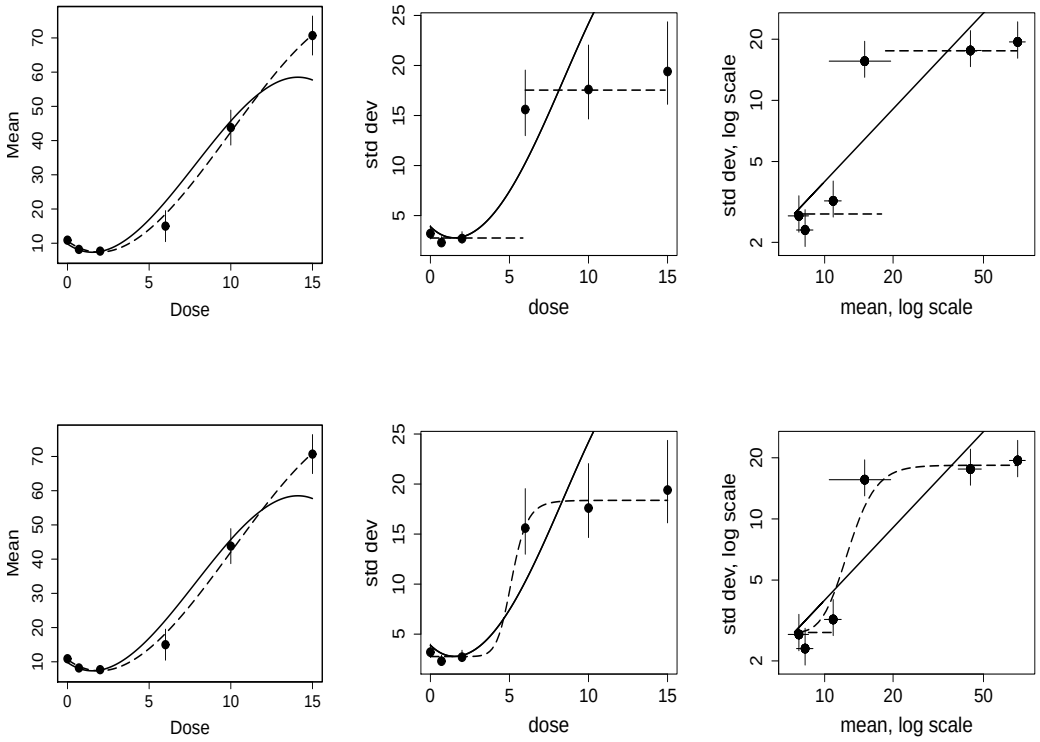

Figure 2: Two alternative models for $\sigma(d)$: a two-component model (top row) and Hill model (bottom row). A cubic model for μ is used in both cases. Solid line: BMDS model (μ cubic, $\sigma(\mu) = \alpha\mu^{\rho}$). Dashed line: new model. Error bars show 95% confidence intervals.

$\mu(d)$ as described below and shown in Figure 2 (top row):

$$\begin{aligned} \mu(d) &= \theta_0 + \theta_1 d + \theta_2 d^2 + \theta_3 d^3 \\ \sigma(d) &= \sigma_1 \ \ for \ \ doses \leq 2, \quad \sigma_2 \ \ for \ \ doses > 2. \end{aligned} \tag{9}$$

The MLEs of the above six parameters, with related quantities, are

$$\begin{aligned} \hat{\theta}_0 = 10.671,\ \ \hat{\theta}_1 = -3.5171,\ \ \hat{\theta}_2 = 1.00615,\ \ \hat{\theta}_3 = -0.033603 \\ \hat{\sigma}_1 = 2.7582,\ \ \hat{\sigma}_2 = 17.539, \\ BMD = 1.145,\ \ BMDL = 0.819,\ \ AIC = 1401. \end{aligned} \tag{10}$$

The BMD d^* is a root of $0 = \theta_1 d^* + \theta_2 d^{*2} + \theta_3 d^{*3} - \sigma(0)$. There are two solutions, 1.145 and 2.753, and we take the smaller as the BMD.

To infer a lower confidence bound for the BMD d^* under the profile likelihood method, we substitute θ_1 by $\frac{\sigma(0)-\theta_2 d^{*2}-\theta_3 d^{*3}}{d^*}$, thus reducing the effective number of parameters from six to five. We next rewrite the likelihood function (4) and maximize it with respect to the remaining parameters. Carrying this out for various values of d^*, using a univariate root-solver, results in the BMDL = 0.819.

As a second approach to modeling the unusual relation of σ to μ, we applied the Hill model (USEPA, 2006; Filipsson et al., 2003),

$$\sigma(d) = g + \frac{v * d^m}{k^m + d^m}, \tag{11}$$

in which parameters g, v, m and k are all positive.

The MLEs of the eight parameters and related quantities are

$$\begin{aligned} \hat{\theta}_0 = 10.678,\ \ \hat{\theta}_1 = -3.5020,\ \ \hat{\theta}_2 = 0.99020,\ \ \hat{\theta}_3 = -0.032587, \\ \hat{g} = 2.7604,\ \ \hat{v} = 15.612,\ \ \hat{m} = 10.618,\ \ \hat{k} = 5.1589, \\ BMD = 1.145,\ \ BMDL = 0.821,\ \ AIC = 1404. \end{aligned} \tag{12}$$

Observation 2.1. None of the dose-response models we evaluated fit especially well according to the likelihood ratio tests we used (Table 2b), although one fits 'adequately' by the criterion $P \geq 0.10$. Data Set 2 exhibits an abrupt increase in variance that could well be related to the (unknown) details of measurement and design, requiring either a piecewise-constant or an abruptly sigmoidal model.

Observation 2.2. Data Set 2 presents an interesting exception to taking the sign of the dose-response curve slope at $d = 0$ to find the BMD. In the case of the quadratic model for μ, no real solution exists when the BMR is taken to be one standard deviation *below* the control response. In the case of the cubic model for μ, there are two solutions for BMD. For a non-monotonic dose-response function, two solutions, $\mu(0) \pm \sigma(0)$, must be evaluated, and the smallest real solution should be accepted.

Table 2b. Comparison of Models

Models for μ and σ	AIC	BMD	BMDL	P(V)	P(M)
BMDS: quadratic(d), constant	1748	6.443	5.528	< 0.001	< 0.001
BMDS: quadratic(d), power(μ)	1539	3.998	3.321	< 0.001	< 0.001
NEW: quadratic(d), Hill(d)	1424	5.76	5.13	0.063	< 0.001
NEW: quadratic(d), σ_1 & σ_2	1424	5.723	1.502	0.119	< 0.001
BMDS: cubic(d), constant	1731	7.260	6.644	< 0.001	0.310
BMDS: cubic(d), power(μ)	1484	4.294	3.926	< 0.001	0.006
NEW: cubic(d), Hill(d)	1404	1.145	0.821	0.063	0.080
NEW: cubic(d), σ_1 & σ_2	1401	1.145	0.819	0.119	0.099

DATA SET 3

This data set comes from a study of reproductive toxicity of butyl benzyl phthalate (BBP) in rats (Tyl et al., 2004). BBP was administered in the feed at 0, 750, 3750, and 11,250 ppm to 30 rats per sex per dose group; dose in Table 3a has units mg/(kg-day). Rats were exposed to BBP for 10 weeks (pre-breeding period), through a 2-week breeding period, then through gestation and lactation. After weaning, the females were necropsied and organ weights were recorded.

Table 3a. Ovary weight as a percentage of final body weight in rats

dose (d)	0	50	250	750
n	30	30	30	30
mean	0.0495	0.0480	0.0460	0.0400
sd	0.001095	0.007120	0.004930	0.03834

Figure 3 shows the plots of the sample means versus dose, sample standard deviations versus dose, and $\log(sd)$ versus $\log(\bar{y})$. It seems unclear whether sigma is better represented by a linear or quadratic function of dose, so we used a linear model for $\mu(d)$ and compared quadratic and linear models for $\sigma(d)$:

$$\begin{aligned} \mu(d) = \theta_0 + \theta_1 d, \quad and \\ \sigma(d) = \psi_0 + \psi_1 d \quad versus \quad \sigma(d) = \psi_0 + \psi_1 d + \psi_2 d^2. \end{aligned} \tag{13}$$

The MLEs and related quantities for the quadratic $\sigma(d)$ model are

$$\begin{aligned} \hat{\theta}_0 = 0.049, \quad \hat{\theta}_1 = -0.0000118, \\ \hat{\psi}_0 = 0.0103, \quad \hat{\psi}_1 = -0.0000524, \quad \hat{\psi}_2 = 0.00000012, \\ BMD = 397.8, \quad BMDL = 18.27, \quad AIC = -1012. \end{aligned} \tag{14}$$

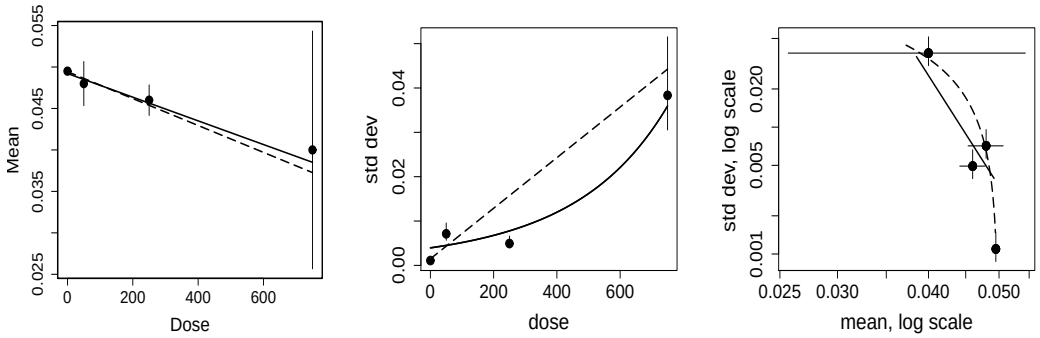

Figure 3: Solid line: BMDS model. Dashed line: new model (linear models for both μ and σ). Error bars show 95% confidence intervals.

The expression for the BMD d^* from (2) is given by

$$d^* = -\frac{\psi_0}{\theta_1} \tag{15}$$

and its MLE is $\hat{d}^* = 397.8$. To infer the BMD d^* under the profile likelihood method, we substitute $-\frac{\psi_0}{d^*}$ for θ_1, reducing the effective number of parameters from five to four. We next rewrite the normal likelihood function (4) and maximize it with respect to the remaining parameters. Carrying out the constrained maximization for various values of d^* yields the solution $BMDL = 18.27$.

Similar results after fitting linear models for both $\mu(d)$ and $\sigma(d)$ are

$$\begin{aligned} \hat{\theta}_0 = 0.04944, \quad \hat{\theta}_1 = -0.00001622, \\ \hat{\psi}_0 = 0.001475, \quad \hat{\psi}_1 = -0.00005703, \\ BMD = 90.90, \quad BMDL = 53.1, \quad AIC = -1026. \end{aligned} \tag{16}$$

Observation 3.1. Our results are compared with those obtained from BMDS software in Table 3b. The model with constant variance is obviously inappropriate and was rejected by the likelihood ratio tests provided by BMDS. Once again, we see that modeling σ as a power of μ leads to acceptance of the variance model and rejection of the means model by the LRTs. Modeling σ as a polynomial of dose leads to the reverse: rejection of the variance model and acceptance of the means model. The quadratic model for $\sigma(d)$ over-estimates $\sigma(0)$ (0.0051 vs. observed sd = 0.0011) while the linear model estimate (0.0015) for $\sigma(0)$ is closer to the observed sd; perhaps this, and the quadratic function shape, account for the low BMDL of 18 for this model. The model with lowest AIC, linear in dose for both μ and σ, provided the best fit. Despite

the LRTs, it appears reasonable to infer BMD and BMDL from the model linear for μ and σ, because these models agree well with the observations near the BMD and at the control dose. As shown later, the 2^{nd} BMDS model in Table 3b also fits the means adequately.

Table 3b. Comparison of Models

Models for μ and σ	AIC	BMD	BMDL	P(V)	P(M)
BMDS: linear(d), constant	-821	1594	879	< 0.001	0.984
BMDS: linear(d), power(μ)	-1019	277	172	0.227	< 0.001
NEW: linear(d), quadratic(d)	-1012	398	18.3	< 0.001	0.787
NEW: linear(d), linear(d)	-1026	90.9	53.1	< 0.001	0.655

DATA SET 4

This data set also comes the study of reproductive toxicity of butyl benzyl phthalate (BBP) in rats by Tyl et al. (2004). These data represent testis weight in grams for male offspring of the adult rats exposed to BBP. Dosage is expressed as mg/(kg-day).

Table 4a. Testis weight in F1 offspring of rats exposed to BBP

dose (d)	0	50	250	750
n	30	29	28	28
mean	3.598	3.649	3.623	2.858
sd	0.2739	0.2531	0.6032	0.9472

Figure 4 shows the plots of the sample means versus dose, sample standard deviations versus dose, and $\log(sd)$ versus $log(\bar{y})$. Based upon comparisons of AIC and graphics, a quadratic model for $\mu(d)$ and a linear model for $\sigma(d)$ were selected:

$$\mu(d) = \theta_0 + \theta_1 d + \theta_2 d^2, \qquad \sigma(d) = \psi_0 + \psi_1 d. \tag{17}$$

The MLEs and related quantities for this model are

$$\begin{aligned} \hat{\theta}_0 = 3.6055, \quad \hat{\theta}_1 = 6.7988e-04, \quad \hat{\theta}_2 = -2.2439e-06, \\ \hat{\psi}_0 = 0.25289, \quad \hat{\psi}_1 = 0.001020119, \\ BMD = 520, \quad BMDL = 362, \quad AIC = -63.5. \end{aligned} \tag{18}$$

Observation 4.1. Our results are compared with those from BMDS software in Table 4b. Modeling σ as a power of μ leads to rejection of the means model by the LRTs, while in this case the power model for variance is also rejected. In our alternative approach, modeling σ as a linear function of dose leads to a marginally acceptable ($P = 0.123$) variance model and acceptance of the means model. The

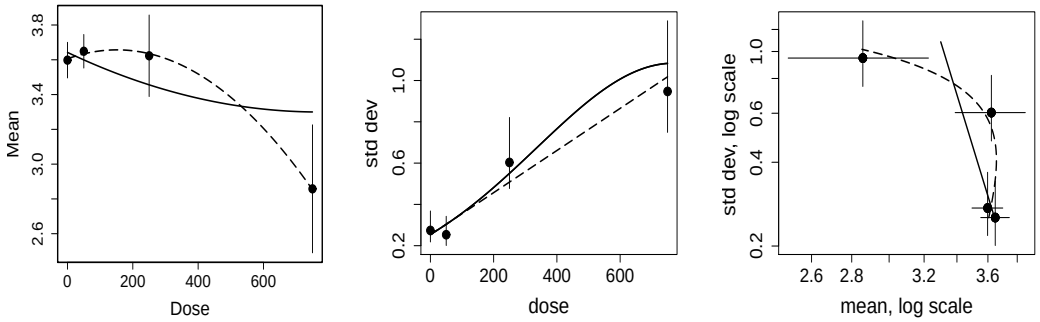

Figure 4: Solid line: BMDS model. Dashed line: new model (linear models for σ). Quadratic models are used for the means. Error bars show 95% confidence intervals.

BMDS quadratic means model is accurate when variance is constant, but then the estimated SD is 0.575, much larger than that observed for the control and first dose (0.274, 0.253), resulting in a overly high BMD and BMDL (678 and 542). When variance is modeled as a power of μ, the BMDS quadratic means model is wide of the mark, but this is corrected when σ as modeled as a linear function of dose (Fig. 4).

Table 4b. Comparison of models

Models for μ and σ	AIC	BMD	BMDL	P(V)	P(M)
BMDS: quadratic(d), constant	-4.3	678	542	< 0.0001	0.858
BMDS: quadratic(d), power(μ)	-55.2	380	170	0.013	0.017
NEW: quadratic(d), linear(d)	-63.5	520	362	0.123	0.735

Discussion

In the BMDS software (USEPA, 2006) and more generally (Filipsson et al., 2003), dose-response models for continuous responses employ one of two models for the variance: (1) constant variance common to all groups, or (2) variance as a power of the predicted group mean. In our experience, these models provide an adequate fit in the great majority of cases. Occasionally, however, one encounters data sets for which the variance is not constant (as determined by a LRT) and is also not well described by a power of the mean, sometimes not even as a monotonic function of dose. We show a few such data sets and explore alternative models for σ as a function of the dose rather than μ, uncoupling $\hat{\mu}$ from $\hat{\sigma}$.

These examples demonstrate the utility of examining a log-log plot of the sample

mean vs. standard deviation to evaluate the suitability of the power model $\sigma^2 = \alpha\mu^\rho$, and, when appropriate, modeling $\mu(d)$ and $\sigma(d)$ separately as a function of the dose d rather than tying σ to μ. When it is necessary to use such data for dose-response modeling, these alternative models can provide acceptable and workable descriptions of the data.

In some cases, the likelihood ratio tests (LRT) for goodness of fit employed by BMDS indicate a conflict or trade-off: the model for means fits adequately when variance is constant, but a constant variance model is ill-fitting (and rejected by the LRT); however when variance is modeled (adequately, according to the LRT) as a power of the mean, the model for the means is rejected by a LRT. In these cases, graphical examination may show that there is indeed a trade-off between fitting models for mean vs. variance. There are other cases, without an apparent trade-off, in which the mean can be fitted adequately by some model, but the variance is not modeled adequately by the two 'standard' choices. Practitioners then might be lead to reject the data as not amenable to modeling (using the available choices provided in BMDS). However, the LRT "Test 4" can be misleading in this case and the model for the means may be adequate, as explained below.

Alternative models for variance

Carroll and Ruppert (1988) provide a much more comprehensive discussion of approaches to variance modeling that might be applied in the context of dose-response models. We used simple functions of dose and chose quadratic and linear models for σ as a matter of convenience. When there are more than four dose groups, it may be useful to consider 4-parameter models for σ such as the 'exponential' models (USEPA, 2006) of Slob (2002), the Hill model (illustrated above using Data Set 2), or a modification of the "multistage" model having an asymptote term, i.e., $a*$[1 - exp($-q_0$ $-q_1 * d$ $-q_2 * d^2$)]. It is also possible to model σ as a power of dose rather than μ, using a third parameter, e.g., $\sigma(d) = \rho[d + \gamma]^\eta$. If desired, all of these models can be made monotone with suitable restrictions on parameters. Non-monotonic variance models may be reasonable if variance heterogeneity is confirmed, the model fits well, and one does not extrapolate beyond the observed doses. One would prefer to have a clear explanation of the pattern of variation in terms of sources of variation and how these relate to means and doses, but that is seldom possible.

Generally for dose-response modeling of continuous endpoints, variance should be bounded away from zero (in particular, at the control, unless the control response is known to be identically zero). If a polynomial model fits well, it will necessarily predict positive variances over the region of interest; however, the profile likelihood search would need to incorporate lower bounds to the parameter space or a nonlinear constraint to maintain $\sigma(0) > 0$. The other models just mentioned require 'background' parameters that insure $\sigma(0) > 0$, but other parameters can vary freely in a profile likelihood search.

Non-proportionality between variance and mean

Practical experience with hundreds of data sets and various lines of reasoning suggest that variance should be proportional to the mean for many types of biological response variables (endpoints) that are subject to dose-response modeling. Thus, when variance is not well described as a power of the mean, it seems plausible that the exceptions are likely related to issues of design, measurement error, or how the endpoint is quantified. Data sets 1 and 2 have endpoints that were constructed as averages of random variates that have moderately complex hierarchical structure. Also, individuals differ in tolerance to dosage, so the proportion of individuals responding adversely may increase with dosage, causing increased variability at intermediate or higher doses (depending on the range of doses used). In data set 3, the endpoint, organ weight as a fraction of body weight, may not change proportionally to the mean because both organ and body weight are changing with dosage. In data set 4, the abrupt increase in variance of organ weight could perhaps be explained if individuals responded divergently at the two higher doses. In other cases, measurement error might obscure the inherent biological relation between variance and mean, resulting in a non-monotonic relation or a monotone relation not well described by a power of the mean. In the event of non-proportionality, it may be necessary to consider models for the variance other than the customary ones, including non-monotonic models.

Without detailed knowledge about the processes producing the observed variances, choice of a smooth model, either monotonic or non-monotonic, is merely an expedient, albeit often a necessary one. A similar expedient is exercised in fitting models to the mean responses: usually, a variety of models are fitted and a selection is made based upon various goodness of fit statistics and other criteria (Filipsson et al., 2003; USEPA, 2006). [6]

Goodness of Fit to the Means

Lastly, we address the problem with the likelihood ratio test for goodness of fit of the means model that was illustrated above for Data Set 4. "Test 4" reported in BMDS consists of the LRT

$$-2[LL(A3) - LL(fit)] > \chi^2(1-\alpha, (df_{A3} - df_{fit})), \tag{19}$$

where LL(fit) is the log-likelihood for the user-selected, or fitted model, and LL(A3) is the log-likelihood for Model A3, in which means are unique to groups and variance is a power of those means. When the variances are indeed a power of the means,

[6] Model shapes for the mean response are generally constrained by a toxicological presumption that the response curve is monotonic, and for some responses, that biological principles require an upper bound or asymptote (Filipsson et al., 2003; USEPA, 2006). This may also apply to variances when the design and sources of variation are simple and measurement error is relatively small. However, monotonicity of variances in relation to doses or means may not always hold, as we have found.

Model A3 should provide a reasonable reference model for goodness of fit to means of the user-selected model. Using sample statistics for the means ($\bar{y}_i$) and standard deviations (s_i), indexed by dose groups (i = 0, 1, ..., k), these log-likelihoods are both described by

$$2LL = -\sum_{i=0}^{k} n_i \log(\widehat{\sigma}_i^2) - \sum_{i=0}^{k} \frac{(n_i - 1)s_i^2}{\widehat{\sigma}_i^2} - \sum_{i=0}^{k} \frac{n_i(\bar{y}_i - \widehat{\mu}_i)^2}{\widehat{\sigma}_i^2} \tag{20}$$

with $\widehat{\sigma}_i^2 = \alpha(\widehat{\mu}_i)^\rho$ for both models. In Model A3, the means are estimated uniquely for each dose group by $(\hat{\mu}_0, ..., \hat{\mu}_i, ..., \hat{\mu}_k)$. In the user-selected or 'fitted' model, the means are modeled by some function $\mu(d) = f(d, \theta)$. Estimation is by maximum likelihood and in general there is no explicit closed-form solution.

LRT Test 4 can be partitioned into contributions from lack of fit to variance and mean:

$$-2[LL(A3) - LL(fit)] =$$

$$+\sum_{i=0}^{k} n_i \left[\log(\widehat{\sigma}_{i,fit}^2) - \log(\widehat{\sigma}_{i,A3}^2)\right] \quad + \sum_{i=0}^{k}(n_i - 1)s_i^2 \left[\frac{1}{\widehat{\sigma}_{i,fit}^2} - \frac{1}{\widehat{\sigma}_{i,A3}^2}\right] \tag{21}$$

$$+\sum_{i=0}^{k} \frac{1}{\widehat{\sigma}_{i,fit}^2}(\bar{y}_i - \hat{\mu}_{i,fit})^2 - \sum_{i=0}^{k} \frac{1}{\widehat{\sigma}_{i,A3}^2}(\bar{y}_i - \hat{\mu}_{i,A3})^2$$

Evidently, Test 4 is not a test of goodness of fit for the means alone, unless the variance estimates under the fitted model and Model A3 are very nearly the same; it can be inflated by lack of fit to the variances.

When variance is modeled as a constant, the variance-related terms on the first line of Eq. 21 vanish; Model A3 is replaced by Model A2, in which the means and variances are estimated uniquely for each dose group by their MLEs (i.e., $\hat{\mu}_i = \bar{y}_i$, and $\hat{\sigma}_i^2 = ((n_i - 1)/n_i)s^2$). When variance is modeled as a function of dose, the MLEs for variance and mean are not functionally tied together, making it easier to achieve a good fit using an appropriate model. In contrast, when variance is modeled as a power of the mean but that model does not fit well, Test 4 can be misleading.

The lack of fit of the variance model $\sigma^2 = \alpha\mu^\rho$ in our examples illustrates how $\widehat{\sigma}_{i,fit}^2$ may differ from s_i^2, inflating the LR for Test 4. In Table 5, details are shown for Data Set 3: log-likelihoods are shown for model A3 and the fitted model, and the variance and mean components are shown for the Test 4 statistic. The first component, related to variance estimates, dominates when σ^2 is modeled as a power of μ for these data, resulting in $P(M) < 0.001$.

Table 5. Components of LR Test 4 for Models of Data Set 3

μ	σ	$2LL_{A3}$	$2LL_{fit}$	LRT4	P(M)
linear(d)	constant	413.528	413.512	$0.0322 = 0^{+} + 0.0322$	0.984
linear(d)	power(μ)	550.4	513.6	$73.7 = 76.3 - 2.6$	< 0.001
linear(d)	linear(d)	517.3841	516.9625	$0.843 = 0^{+} + 0.843$	0.656

An approximate F-test for goodness of fit of the means model is discussed by Seber and Wild (1989, particularly p. 82, Section 2.8.6). Applied to the three models in Table 5, it results in p-values of 0.984, 0.735, and 0.664, respectively. The second P-value contradicts the result in the middle row of Table 5 for LRT4, which was dominated by a contribution from lack of fit to the variances. The F-test consists of

$$\frac{(Q_H - Q)/(k' - p)}{Q/(N - k')} \sim F_{k'-p,N-k'} \tag{22}$$

with $k' = k + 1$,

$$Q_H = \sum_{i=0}^{k} \sum_{j=0}^{n_i} \hat{w}_i (y_{ij} - \hat{\mu}_i)^2,\ Q = \sum_{i=0}^{k} \sum_{j=0}^{n_i} \hat{w}_i (y_{ij} - \bar{y}_i)^2,\ \textit{and}\ N = (\sum_{i=0}^{k} n_i). \tag{23}$$

The estimated weights are $\hat{w}_i = 1/\hat{\sigma}_i^2$. The equation for Q_H expands into

$$Q_H = \sum_{i=0}^{k} \sum_{j=0}^{n_i} \hat{w}_i (y_{ij} - \bar{y}_i)^2 \quad + \sum_{i=0}^{k} \hat{w}_i n_i (\bar{y}_i - \hat{\mu}_i)^2. \tag{24}$$

so that the term $(Q_H - Q)$ in the F-ratio becomes $\sum_{i=0}^{k} \hat{w}_i n_i (\bar{y}_i - \hat{\mu}_i)^2$, while the denominator is obviously $Q = \sum_{i=0}^{k} \hat{w}_i (n_i - 1) s_i^2$.

These quantities can be calculated using the estimated standard deviations ($\hat{\sigma}_i^2$) and scaled residuals (r_i) reported by BMDS. $(Q_H - Q)$ is given by $\sum r_i^2$, because $r_i = (\bar{y}_i - \hat{\mu}_i)/(\hat{\sigma}_i/\sqrt{n_i})$. Q is obtained by $\sum w_i (n_i - 1) s_i^2 = \sum (1/\hat{\sigma}_i^2)(n_i - 1) s_i^2$. From the nominal d.f. for Q, N-k', we would subtract one to account for the power parameter used in the model for variances.

Having determined that the means model fits adequately, while the variance is not well described by the BMDS models, a user might go on to apply an alternative model for variances. For the data sets we have presented, this F-test, applied to models fitted in BMDS (using the power law for variances), yielded $P > 0.10$ for Data Sets 1 and 3 but not Date Sets 2 and 4.

Acknowledgments

Our sincere thanks are due to Paul White for his encouragement and suggestions. We thank Dihua Xu, a graduate student at UMBC, for programming the OLS and MLE solutions and graphics during the initial stages of this project.

References Cited

1. Carroll, R. J., and D. Ruppert. (1988). ***Transformation and Weighting in Regression***. Chapman and Hall

2. Crump, K. S.; R. Howe. (1985). *A review of methods for calculating statistical confidence limits in low dose extrapolation*, Chapter 9 in ***Toxicological Risk Assessment***. D. B. Clayson, D. Krewski, I. Munro, eds. Boca Raton: CRC Press, Inc.

3. Filipsson, A. F., S. Sand, J. Nilsson and K. Victorin (2003). *The benchmark dose method - review of available models, and recommendations for application in health risk assessment.* ***Critical Reviews in Toxicology*** **33**: 505-542.

4. George, J. D., J. R. Reel, C. B. Myers, A. D. Lawson and J. C. Lamb IV (1986). *Trichloroethylene: reproduction and fertility assessment in F344 rats when administered in the feed*. National Toxicology Program, National Institute of Environmental Health Sciences. Report NTP-86-085 (January 13, 1986).

5. Rao, C. R. (1973). ***Linear Statistical Inference and Its Applications***. New York: John Wiley.

6. Slob, W. (2002). *Dose-Response Modeling of Continuous Endpoints.* ***Toxicological Sciences*** **66**: 298-312.

7. Schlosser, P. M., P. D. Lilly, R. B. Conolly, D. B. Janszen, and J. S. Kimbell (2003). *Benchmark dose risk assessment for formaldehyde using airflow modeling and a single-compartment, DNA-protein cross-link dosimetry model to estimate human equivalent doses.* ***Risk Analysis*** **23**: 473-487.

8. Seber, G. A. F., and C. J. Wild (1989). ***Nonlinear Regression***. (reprinted 2003) John Wiley.

9. Tyl, R. W., C. B. Myers, M. C. Marr, P. A. Fail, J. C. Seely, D. R. Brine, R. A. Barter, and J. H. Butala (2004). *Reproductive toxicity evaluation of dietary butyl benzyl phthalate (BBP) in rats.* ***Reproductive Toxicology*** **18**: 241-264.

10. USEPA (U. S. Environmental Protection Agency) (2006). *Help Manual for Benchmark Dose Software - Version 1.4*. Office of Research and Development. EPA 600/R-00/014F, September, 2006.