

Practical Considerations in Applied Nonlinear Regression Modelling: Design, Estimation and Testing

Timothy E. O'Brien
Department of Mathematics and Statistics
Loyola University Chicago

NICASA-2007

Chicago, Illinois

22 March 2007

Talk Outline

- A. Nonlinear Models – Definitions and Illustrations
- B. Confidence Intervals and Curvature
- C. Correlated and Heteroskedastic Nonlinear Models
- D. Generalized Linear and Nonlinear Regression Models
- E. Design Considerations
- F. Comments, Recommendations and Conclusions

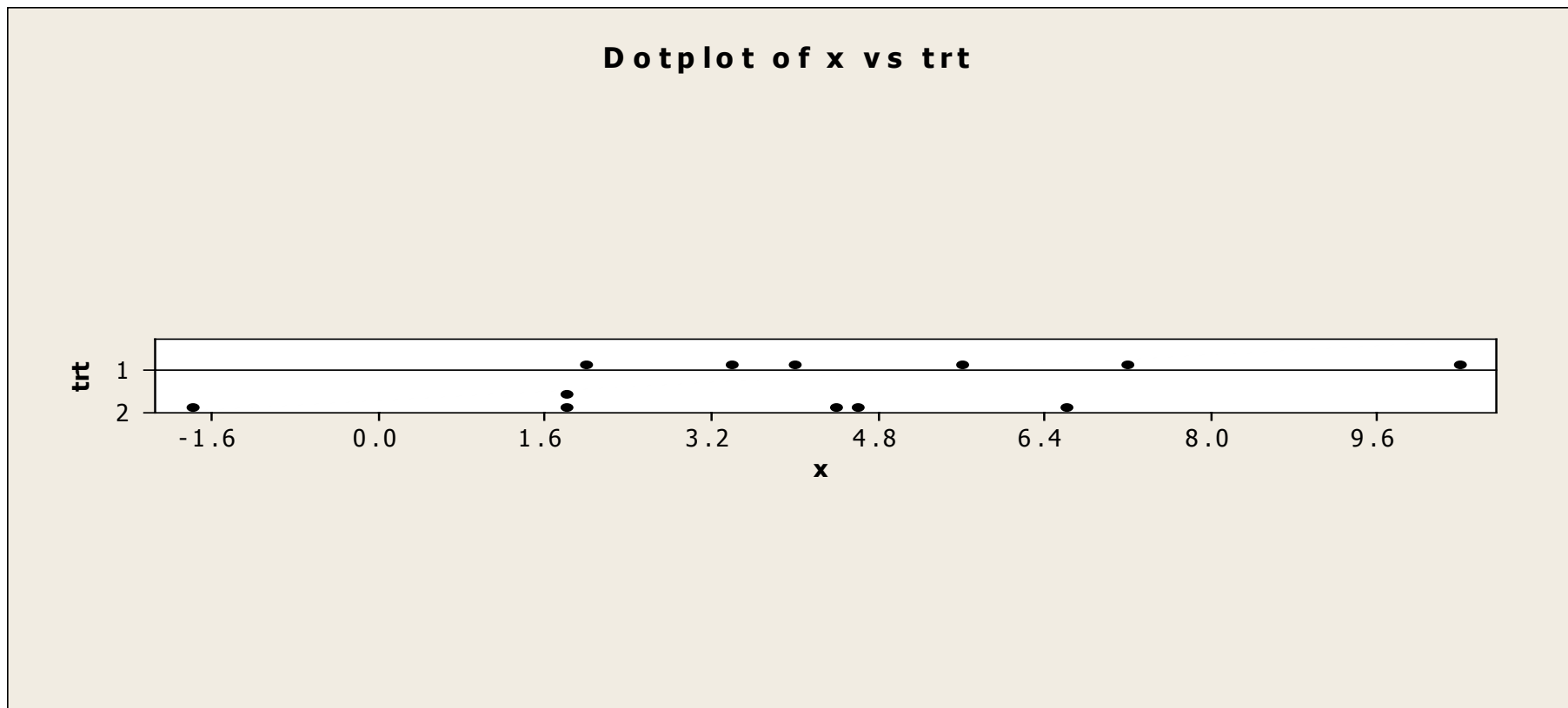
A. Nonlinear Models – Definitions and Illustrations

- *The model* –
 - assumed distribution
 - link function (e.g., identity, logit, log, etc.)
 - model function $\eta(\mathbf{x}, \theta)$ modelling the mean
 - variance function (perhaps depending on θ and/or additional parameters)
- *The goal* –
 - efficiently estimate the p parameters in θ (point or interval) – maybe only interested in a subset of θ
 - conduct hypothesis tests
 - make predictions
- *Choose an efficient design* (discussed later).

Examples:

1. Ratio of Two Normal Means (Fieller-Creasy Problem)

Variable	trt	N	Mean	StDev
x	1	6	5.40	3.07
	2	6	2.82	2.94



The NLIN Procedure					
Source	DF	Sum of Squares	Mean Square	F Value	Approx Pr > F
Model	1	20.0208	20.0208	2.22	0.1672
Error	10	90.2483	9.0248		
Corrected Total	11	110.3			

Parameter	Estimate	Std Error	Approximate 95% Confidence Limits
th1	5.4000	1.2264	2.6673 8.1327
th2	0.5216	0.2562	-0.0492 1.0924

SAS/IML	PLCI's: for th1	2.6673361	8.1326639
	for th2	0.0154953	1.3868296

Note that the “Approximate” Wald CI (WCI) for $\theta_1 = \mu_1$ is the same as the Profile Likelihood CI (PLCI), but that these intervals differ for $\theta_2 = \mu_2/\mu_1$. We discuss why and what to do later.

2. Binary Logistic Regression and Extensions

Assume $Y_x \sim \text{Binomial}(n_x, \pi_x)$ iid over x , and use the “LOGIT” link function, $g(\pi) = \log\left(\frac{\pi}{1 - \pi}\right)$ – which is the log-odds; then, set

$$\log\left(\frac{\pi_x}{1 - \pi_x}\right) = \beta_0 + \beta_1 x; \text{ i.e., } \pi = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

Collett (2003:6) reports data from a study in which sets of 40 mice were injected with an infecting dose of a culture of pneumococci, the bacteria associated with the occurrence of pneumonia. Each mouse was then injected with one of five doses of a given anti-pneumococcus serum so as to assess its efficacy.

Dose of serum in cc	0.28	0.56	1.12	2.25	4.50
Number of deaths out of 40	35	21	9	6	1

Here, we take $x = \log(\text{dose})$.

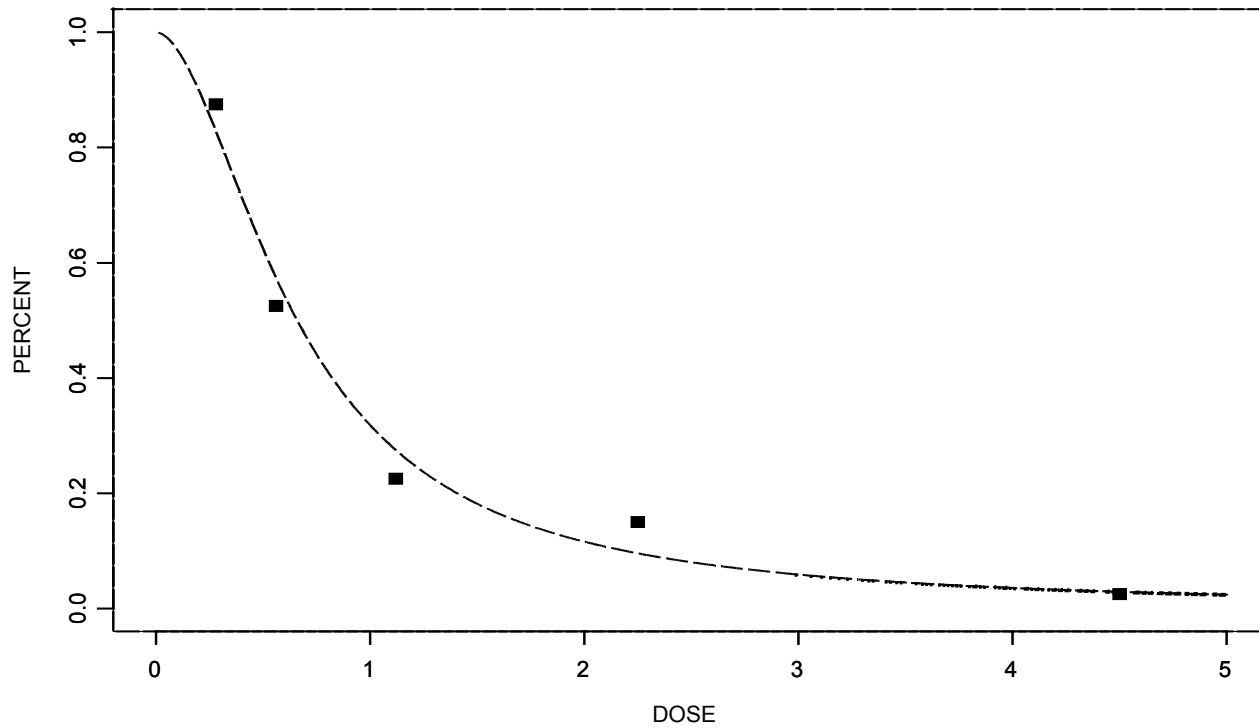
The GENMOD Procedure

GOF Criterion	DF	Value	Value/DF
Deviance	3	2.8089	0.9363

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-0.7637	0.2009	-1.1574	-0.3699	14.45	0.0001
ldose	1	-1.8296	0.2545	-2.3285	-1.3307	51.66	<.0001

Binary LOGLogistic Fit to Pneumonia data



What if our interest is in the $LD_{50} = \phi$? It turns out that $\phi = \exp\{-\beta_0 / \beta_1\}$, which we estimate as $\exp\{-0.7637/1.8296\} = 0.6588$, but what if we want an interval estimate?

→ We can use the NLMIXED procedure in SAS.

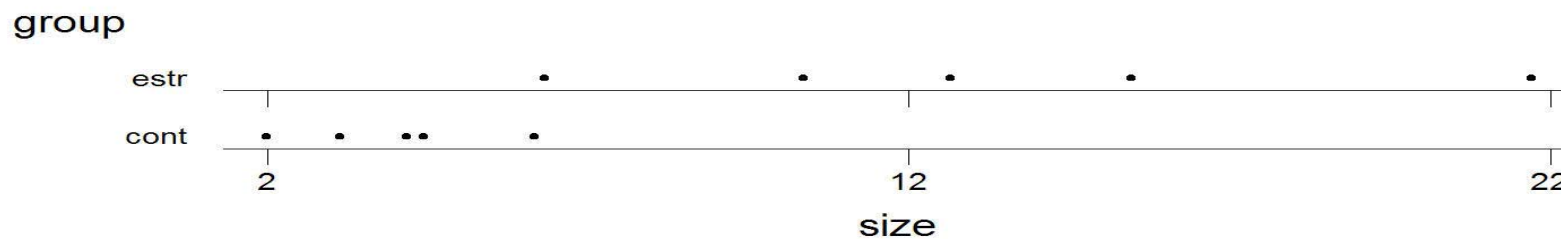
Note that this is now a *Generalized Nonlinear Model* since the RHS is now $\beta_1 \{\log(x) - \log(\phi)\}$ (nonlinear in the parms.), and the distribution is Binomial (Exponential Family but not Normal).

The NLMIXED Procedure									
Fit Statistics									
		-2 Log Likelihood							19.6
		AIC (smaller is better)							23.6
		AICC (smaller is better)							29.6
		BIC (smaller is better)							22.8
Parameter Estimates									
Parameter	Estimate	Standard Error	DF	t Value	Pr > t	Alpha	Lower	Upper	
phi	0.6588	0.06932	5	9.50	0.0002	0.05	0.4806	0.8370	
beta1	-1.8296	0.2545	5	-7.19	0.0008	0.05	-2.4840	-1.1753	

3. Relative Potency

- Direct Assay – Ratio of two means, medians, etc.
Pikounis (2001), prostate data for dogs: relative potency of estradiol to the control

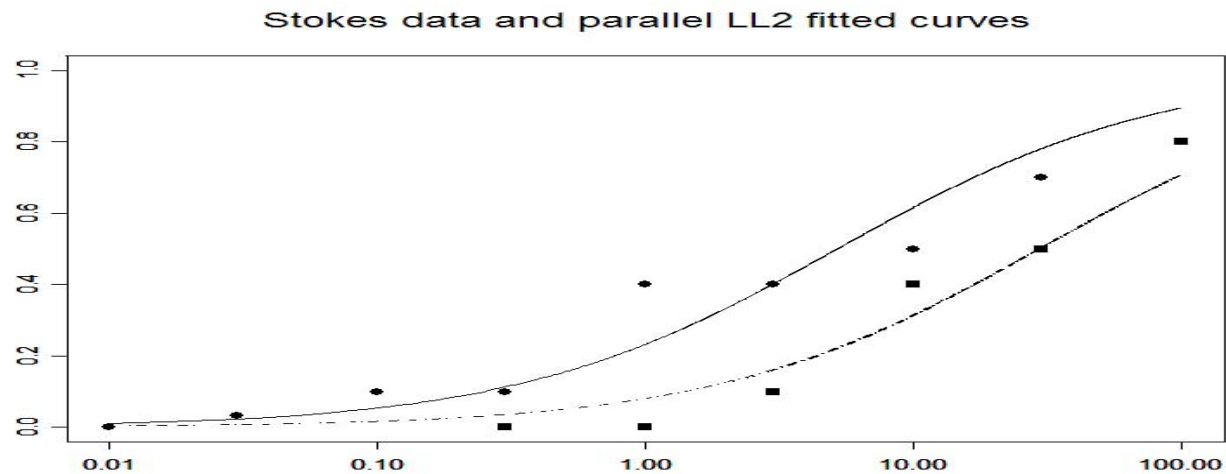
Prostate size versus tmt. group in dogs



Homoskedasticity assumption is unwarranted.

We assume normality, $\mu_{\text{ESTR}} = \rho\mu_{\text{CONT}}$, and that σ^2 is the variance for the Control group, whereas $\rho^2\sigma^2$ is the variance for the Estradiol group. Here ρ is estimated to be 3.42 and with 95% WCI (1.84,5.00).

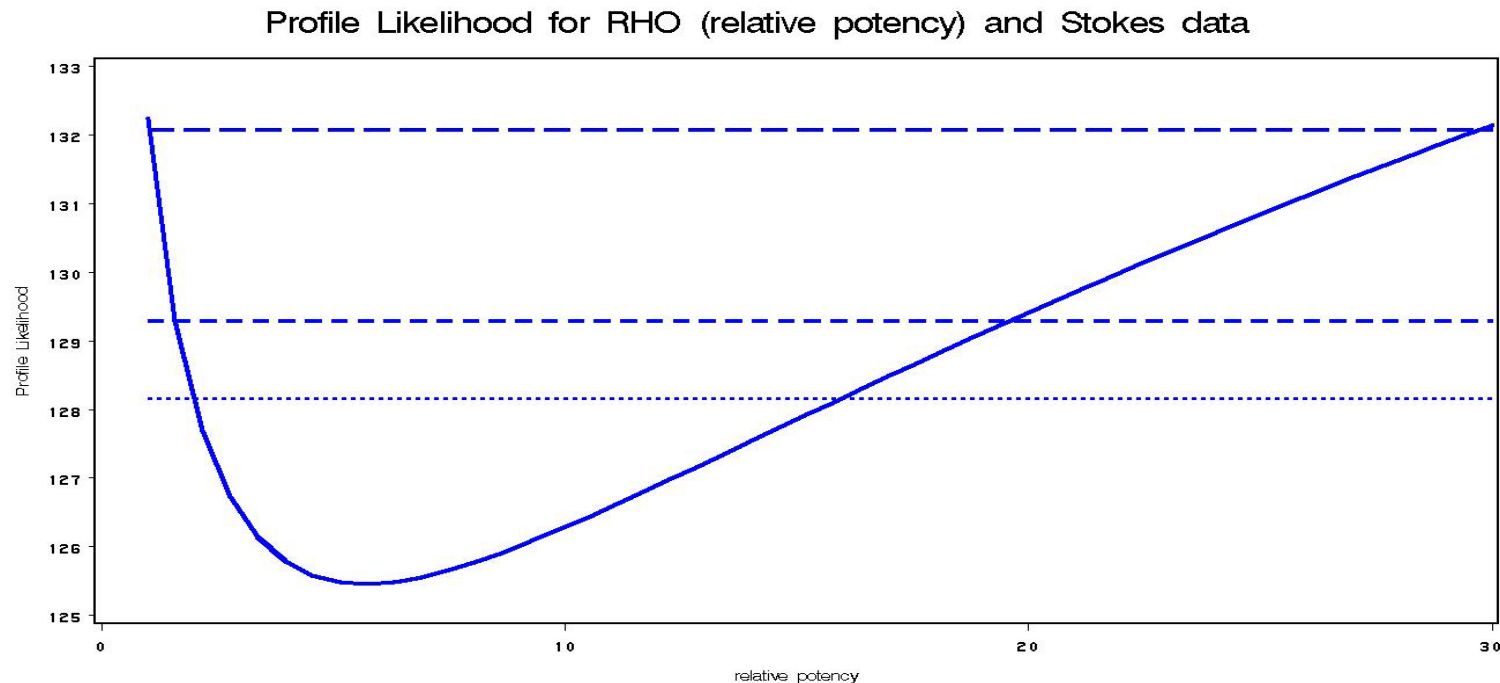
- Indirect Assay – Ratio of two LD_{50} 's ($\rho = \theta_{21}/\theta_{22}$)
Stokes *et al* (2000:331) – two peptides (neurotensin and somatostatin): groups of a fixed number of mice were exposed to one of several doses of either of these drugs and the number of deaths was noted.



The estimated relative potency is about 5.66, meaning that a dose of somatostatin must be approximately 5.66 times higher than that of neurotensin to have the same effect.

95% WCI for ρ is (-1.89,13.22) – inadequate

95% PLCI for ρ is (3.0,19.5): PL curve, graphed below, shows considerable curvature.



4. Synergy Models – two drugs or compounds (A and B) in amounts x_1 and x_2 , then the ‘*effective dose*’ is

$$z = x_1 + \theta_4 x_2 + \theta_5 (\theta_4 x_1 x_2)^{1/2}$$

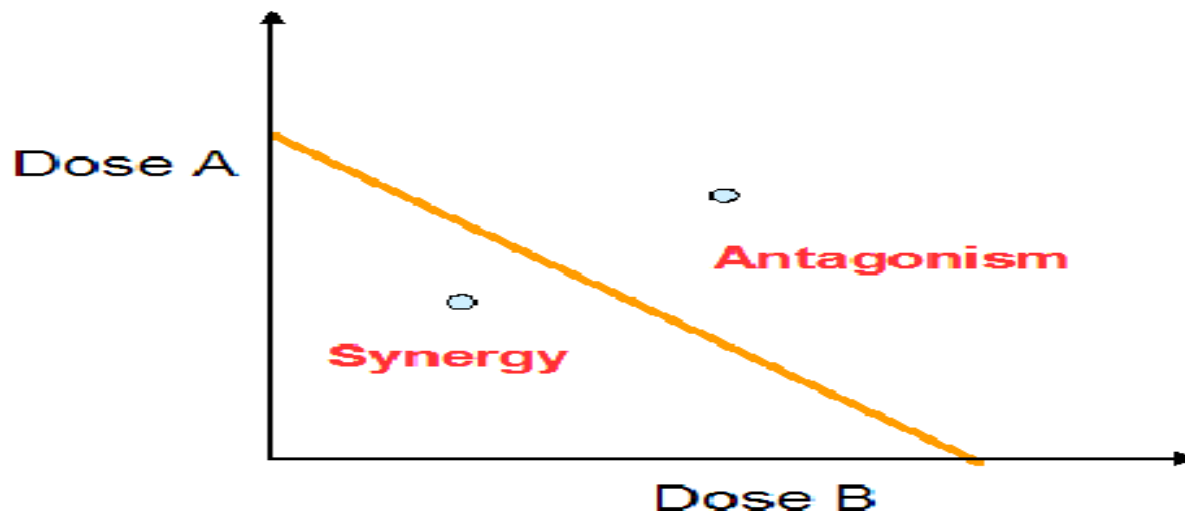
θ_4 is the relative potency parameter;

θ_5 is the *coefficient of synergy* ($\theta_5 < 0$: antagonism,

$\theta_5 \approx 0$: independent action, $\theta_5 > 0$: synergy).

Then, in Gaussian situations, we relate Y to z using a parametric model such as the LL3 model function,

$$E(Y) = \eta = \theta_1 / (1+t), \quad t = (z/\theta_2)^{\theta_3}$$



B. Confidence Intervals and Curvature

- Wald CR's use a Linear Approximation:

(1- α)*100% WCR

$$\{ \theta \in \Theta: (\theta - \theta_*)^T \mathbf{V}_*^T \mathbf{V}_* (\theta - \theta_*) \leq ps^2 F_\alpha \}$$

(1- α)*100% LBCR

$$\{ \theta \in \Theta: S(\theta) - S(\theta_*) \leq ps^2 F_\alpha \}$$

where $S(\theta) = \varepsilon^T \varepsilon = \|y - \eta(\theta)\|^2 = [y - \eta(\theta)]^T [y - \eta(\theta)]$

Linearization

So, if we write $\eta(\theta) \approx \eta(\theta_*) + \mathbf{V}_*(\theta - \theta_*)$, then

$\varepsilon = \varepsilon_* - \mathbf{V}_*(\theta - \theta_*)$, and $S(\theta) - S(\theta_*) \approx (\theta - \theta_*)^T \mathbf{V}_*^T \mathbf{V}_* (\theta - \theta_*)$.

- The extent to which $\eta(\theta) \approx \eta(\theta_*) + \mathbf{V}_*(\theta - \theta_*)$ is captured in *curvature/nonlinearity* – and is discussed below (after an example).

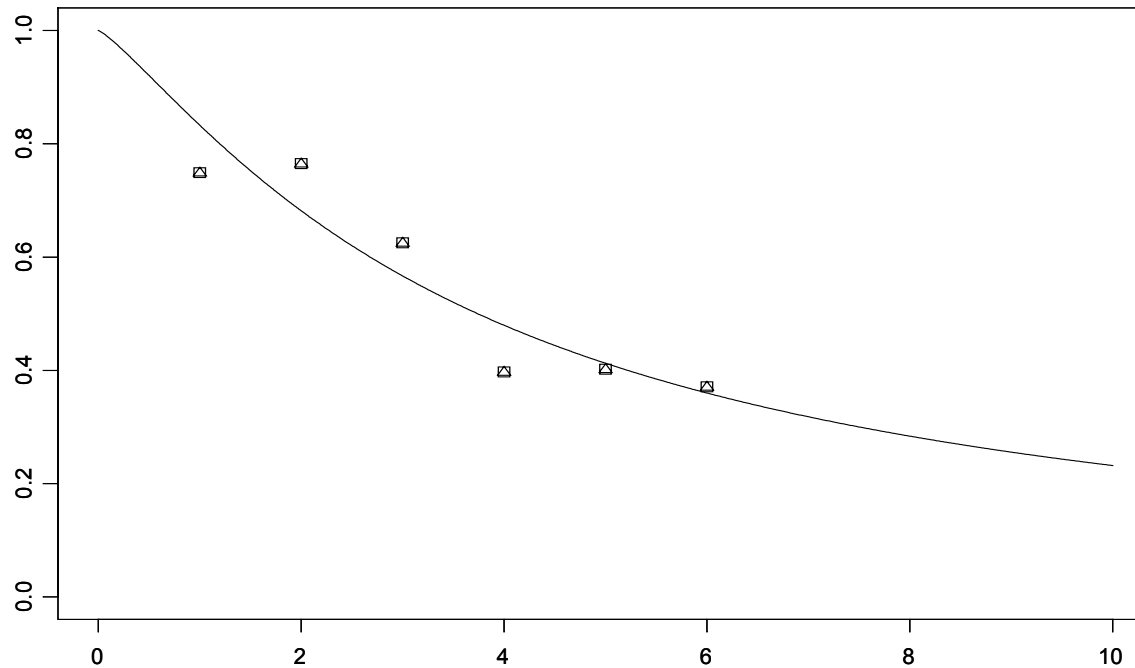
- Sigmoidal Curve (LL2) Example -

Model Function:

$$\eta(x, \theta) = \frac{1}{1 + (x/\theta_2)^{\theta_3}}$$

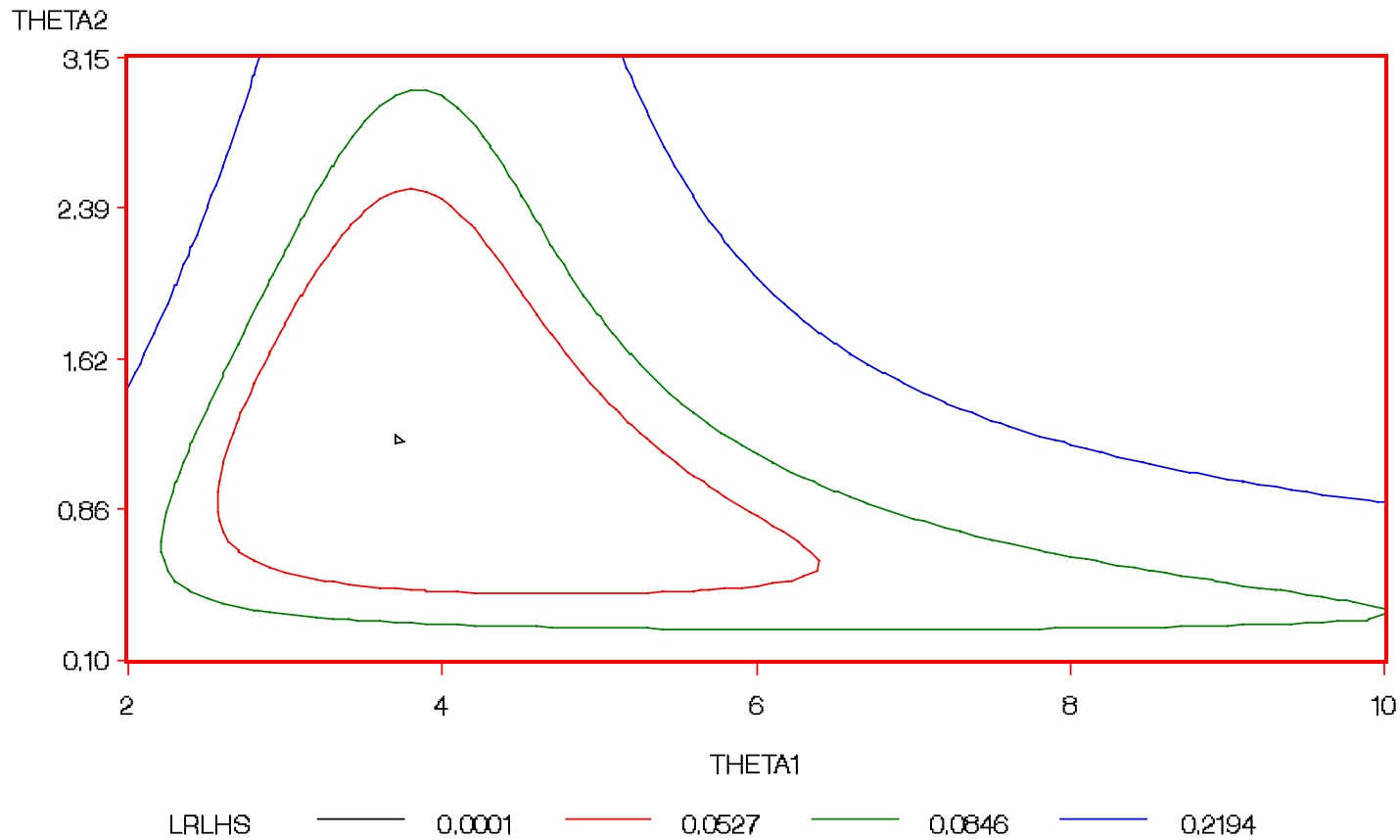
2 Model parameters: θ_2 : LD₅₀ and θ_3 : Slope

LL2 fit to 6-point uniform design data

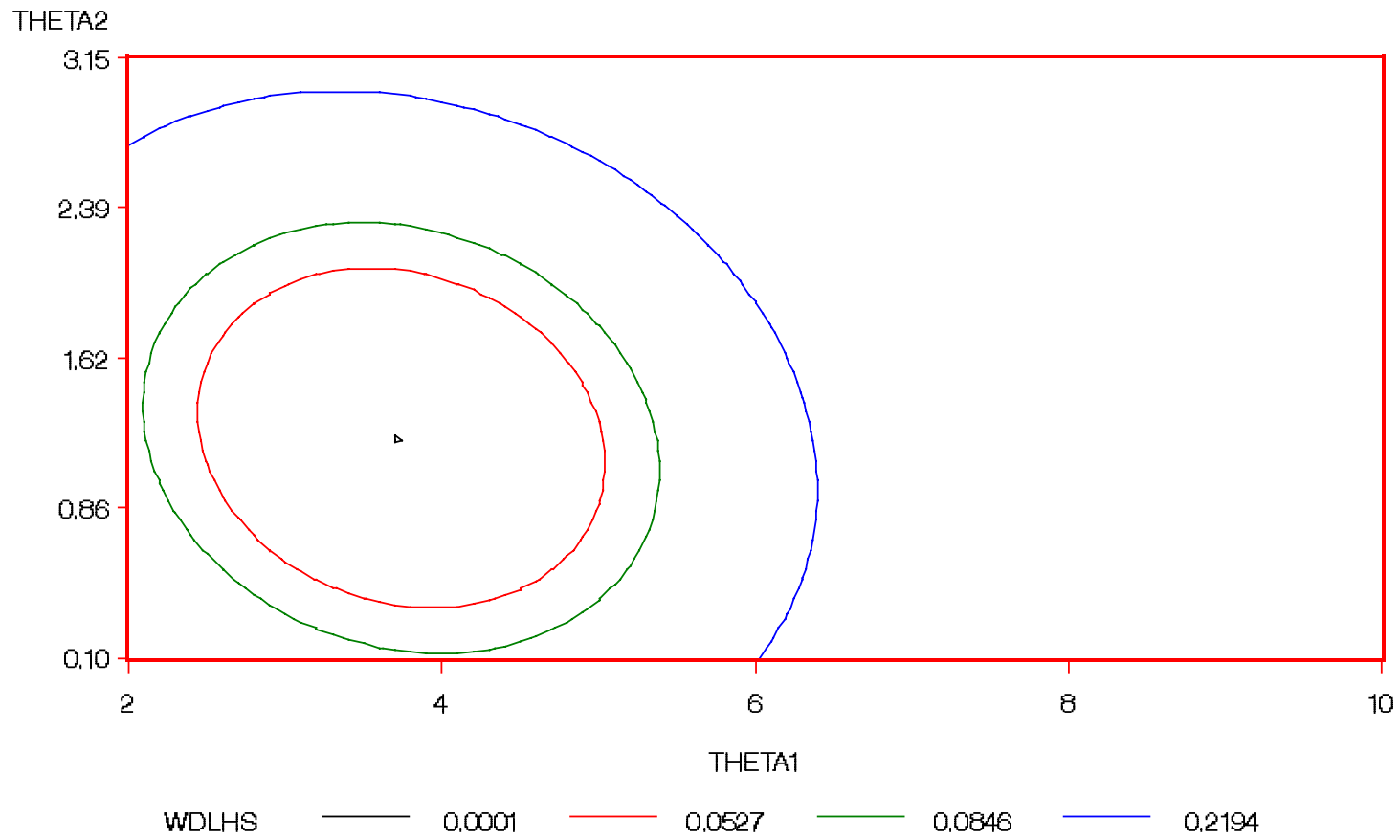


Confidence Regions – 99%(outer), 95%, 90%(inner)

Likelihood Based (LB) CR's

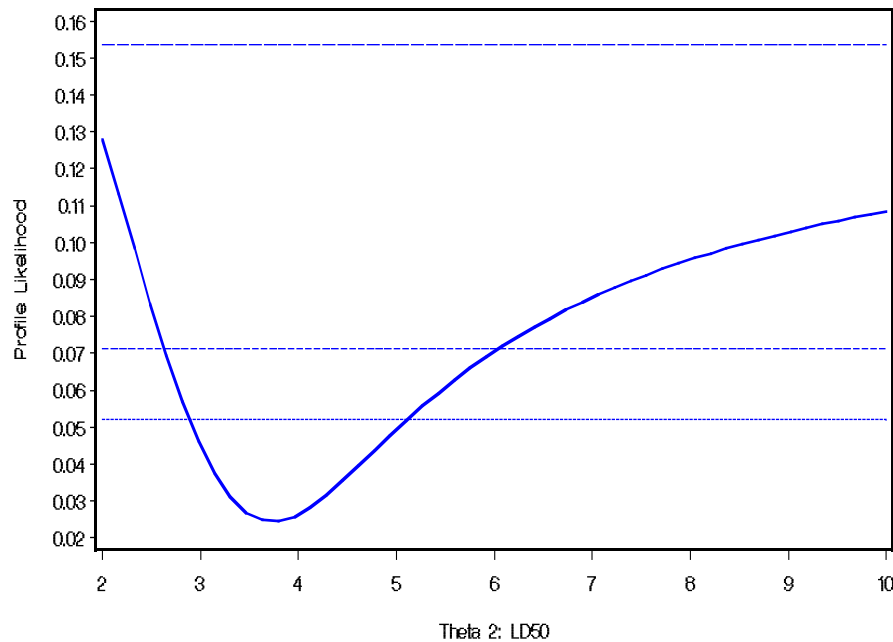


Wald (approximate) CR's

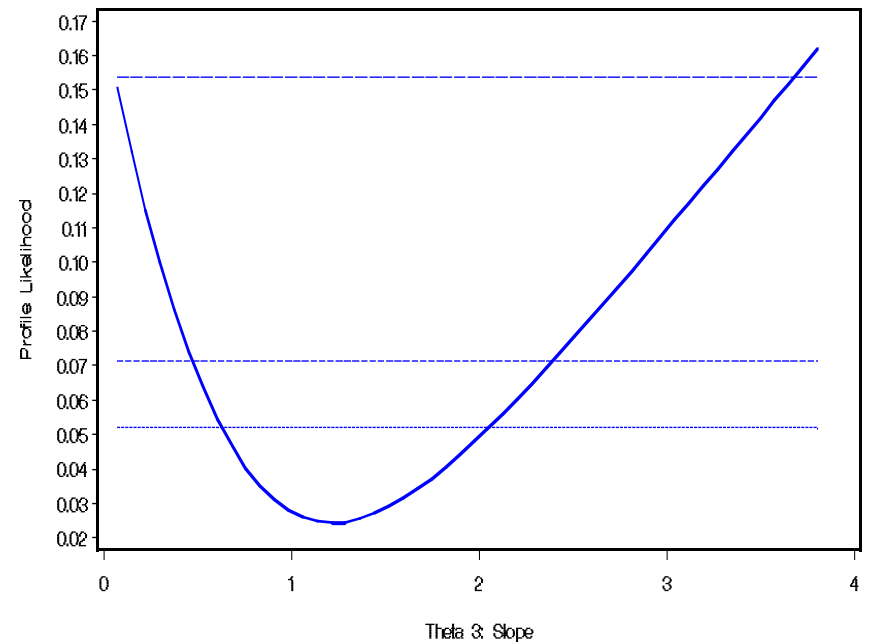


Confidence Intervals – 99%(top line), 95%, 90%(bottom line)

For θ_2



For θ_3



	WCI ($\alpha = 5\%$)		PLCI ($\alpha = 5\%$)	
$\theta_2 - LD_{50}$	2.512	4.968	2.631	6.046
$\theta_3 - Slope$	0.401	2.033	0.471	2.390

- Curvature and Nonlinearity notation:

Θ = the p-dimensional parameter space

E = expectation surface, p-dimensional in n-dim'l. SS

IN (intrinsic) curvature = degree of flatness of E

PE (parameter effects) curvature = degree to which

SPEL in Θ are mapped onto *SPEL* on E

SPEL = straight, parallel, equi-spaced lines (a rectangular grid)

Many of the conventional curvature measures are spurious (indicate a problem when there is none and v.v.): see Cook & Witmer. Clarke (1987): these [IN, PE] “measures suffer from the practical defect, however, of attempting to measure a multidimensional phenomenon by a single quantity.”

- Marginal Curvatures - Clarke (1987) expands θ_K in powers of σ^2 so as to adjust the endpoints of the WCI to bring them more in line with the PLCI.

$$\text{Wald: } [W_L^K, W_U^K],$$

$$W_L^K = \theta_K - t * SE_K$$

$$W_U^K = \theta_K + t * SE_K$$

$$\text{MC: } [M_L^K, M_U^K],$$

$$M_L^K = \theta_K - t * (1 - \Gamma_a t + \beta_a t^2) * SE_K$$

$$M_U^K = \theta_K + t * (1 + \Gamma_a t + \beta_a t^2) * SE_K$$

Γ_a and β_a are functions of the second and third derivatives of η with respect to θ ; note: first derivative (Jacobian matrix) is $n \times p$, second derivative is an $n \times p \times p$ array, etc.

	Wald	MC (Clarke)
Length	$2*t*SE_K$	$2*t*SE_K*(1 + \beta_a t^2)$
Skewness	0	$2*\Gamma_a*t^2*SE_K$

So, like PLCI's, MCCI's can be skewed (determined by Γ_a) and/or widened or narrowed (determined by β_a).

Note: the Hougaard skewness measure given in SAS/NLIN is directly related to Γ_a but ignores β_a ; this is not always a good idea: see Haines *et al* 2004.

LL2 Example continued. Here for θ_2 , $\Gamma_a = 0.0903$, $\beta_a = 0.0281$. For $\alpha = 5\%$, we obtain for θ_2 :

Type	Confidence interval	Overlap to PLCI
Wald	(2.512 , 4.968)	66.2%
MC	(2.554 , 5.542)	83.4%
PL	(2.631 , 6.046)	----

Empirical and theoretical evidence shows that MCCI's perform better in approximating PLCI's than do WCI's. Note that since calculating PLCI's is cumbersome, after finding a WCI and a MCCI, and distinguishing one of four cases (MWWM, MWMW, WMWM, WMMW), we use a function (denoted **f₁**) *which assesses the overlap of the WCI to the MCCI* as an indicator of when a problem exists. Practitioner is provided an indication: (1) use WCI from package (e.g., SAS/NLIN), (2) use MCCI found in SAS/IML, or (3) go through the hassle of finding PLCI; details in Haines *et al* (2004).

Uses LL4 model function,

$$\eta = \theta_4 + \frac{\theta_1 - \theta_4}{1 + (x / \theta_2)^{\theta_3}}$$

Model Variances as well, $\text{Var}(Y) = \sigma^2 \eta^\rho(x)$,

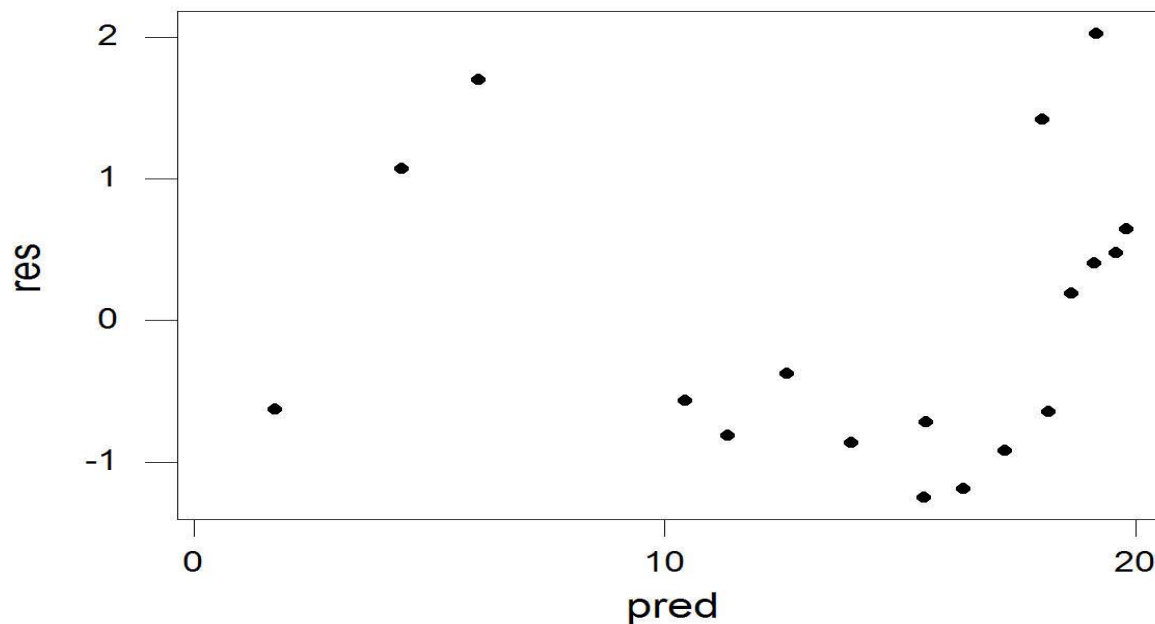
where ρ is an additional model parameter to be estimated; model parameters are thus σ^2 and ρ , and those in θ .

2. Atkinson *et al* (1993) fit the IP3 model

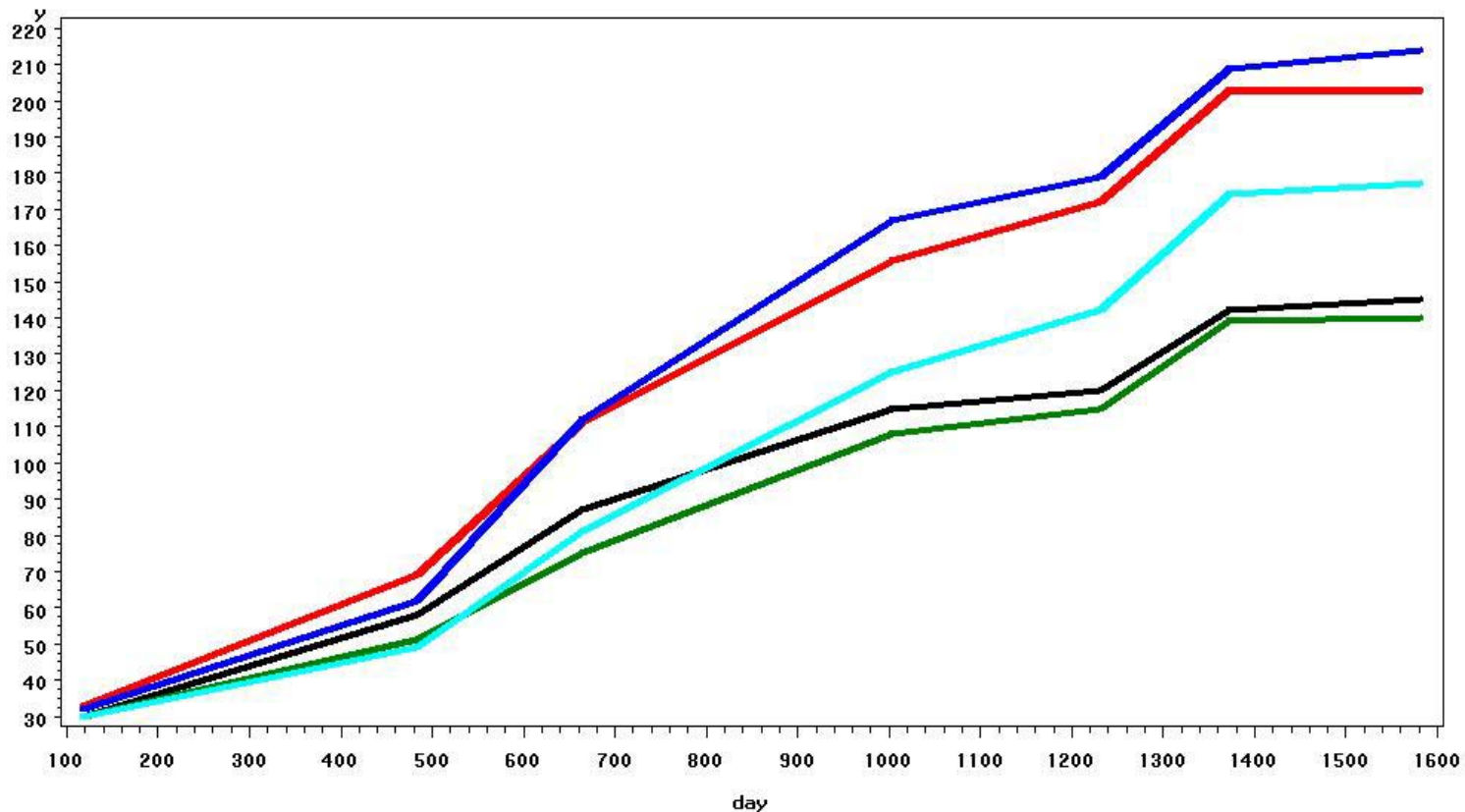
$$\begin{aligned} \eta(t, \theta) &= \theta_3 \left\{ e^{-\theta_2 t} - e^{-\theta_1 t} \right\} \\ &= \frac{\phi \theta_1 \theta_2}{\theta_1 - \theta_2} \left\{ e^{-\theta_2 t} - e^{-\theta_1 t} \right\} \end{aligned}$$

The latter parameterization so that ϕ represents the *area under the curve (AUC)*; Y = drug concentration, t = time. The residual pattern also shows a possible first-order autoregressive (AR_1) structure.

Residual Plot for ACHJ data with heterosk. IP3 fit



3. Littell *et al* (1996), Draper & Smith (1998:559) orange tree growth data, Y = trunk circumference over time for 5 trees.



Model function (Log3):

$$\eta = \frac{\theta_1}{1 + e^{-(t - \theta_2) / \theta_3}}$$

Under three settings:

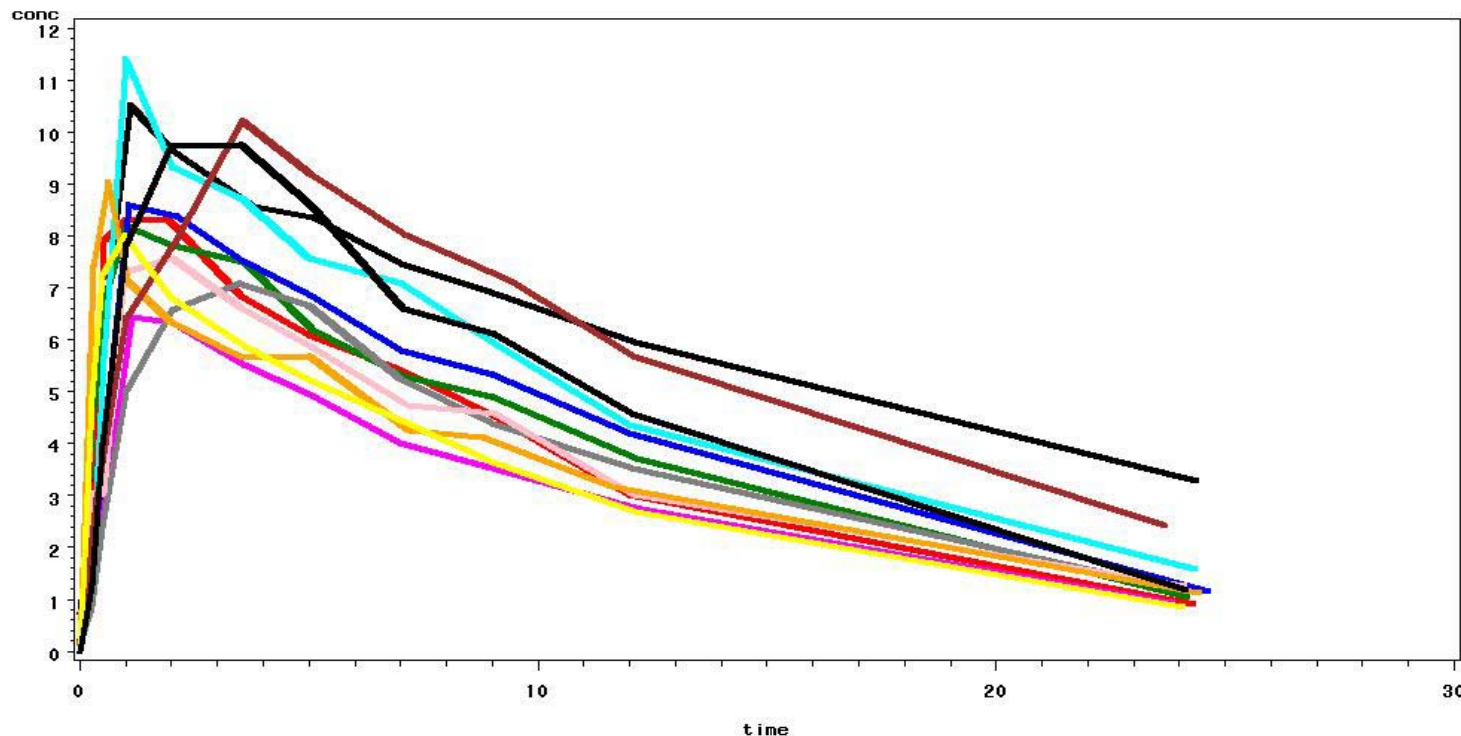
- homoskedasticity assumption
- modelling variances as in Example 1
- hierarchical (HNLM) structure: $\theta_{1k} \sim \mathbf{N}(\theta_{1*}, \sigma_u^2)$ (k^{th} tree)

Setting	-2LL	AIC
Homoskedastic	316.8	324.8
$\text{Var}(Y) = \sigma^2 \eta^p(x)$	291.1	301.1
HNLM	263.1	273.1

Another Example: Pharmacokinetics of theophylline ($Y =$ concentration over time); from Pinheiro & Bates (2000:352). Model function (IP3) for k^{th} (of 12) individual is

$$\eta = \frac{D\theta_{1k}\theta_{2k}}{Cl_k(\theta_{2k} - \theta_{1k})} \left\{ e^{-\theta_{1k}t} - e^{-\theta_{2k}t} \right\}$$

$D = \text{dose given}, \beta_3 = \log(\theta_{1k}), b_{1k} = \log(Cl_k) \sim N(\beta_1, \sigma_1^2), b_{2k} = \log(\theta_{2k}) \sim N(\beta_2, \sigma_2^2)$



- $\log(Cl_k)$ and $\log(\theta_{2k})$ are modeled as bivariate Normal (Cl_k and θ_{2k} are modeled as bivariate Log-Normal) due to skewness
- test of zero covar. between these variance components is retained
- estimated clearance rate (Cl) is $\exp\{-3.2268\} = 0.0397$
- estimated absorption rate (θ_2) is $\exp\{0.4806\} = 1.6170$
- estimated elimination rate (θ_1) is $\exp\{-2.4592\} = 0.0855$

D. Generalized Linear and Nonlinear Regression Models

We consider here the 2-parameter Weibull distribution for Y , which has Survival Function

$$S(y) = e^{-(y/\alpha)^\beta} = 0.5(y/\phi)^\beta$$

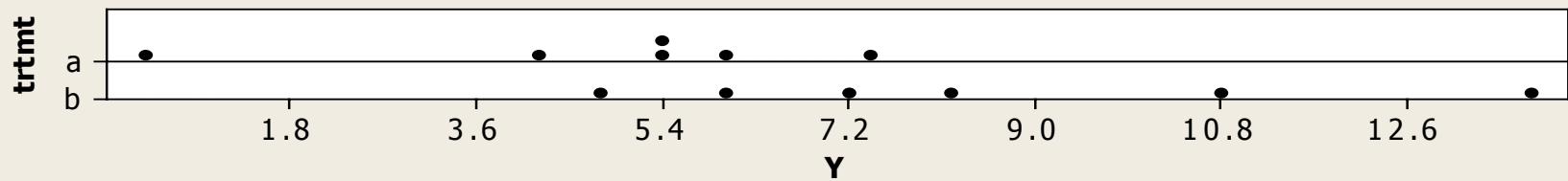
where ϕ is the median (i.e., LD_{50}). Since the median is so much easier to work with for the Weibull distribution than is the mean, we extend the two independent-sample t-test of means and the simple linear and nonlinear regression models to allow for the Weibull distribution (based on medians).

Example - We randomize 6 cold sufferers to receive cold treatment A and we randomize 6 cold sufferers to receive cold treatment B, and we measure the number of days that their colds lasted (Y). We obtain the following data (no censored measurements).

Descriptive Statistics: Y

Variable	trtmt	N	Mean	StDev	Median
Y	a	6	4.798	2.376	5.385
	b	6	8.470	3.370	7.690

Dotplot of Y vs trtmt



Our goal here is to perform the two-tailed test

$$H_0: \phi_A = \phi_B$$

We assume about β_A and β_B (akin to assuming that $\sigma_A^2 = \sigma_B^2$). Since no *pivot* exists for this test, we LL's (log-likelihood's) associated with the *Full* (no restrictions on ϕ_A and ϕ_B) and *Reduced* (imposing H_0) models, and form the *test statistic*, $-2\Delta LL$, which for large samples has a χ^2 distribution. Again, we can use NLMIXED in SAS to do this.

Full Model

The NLMIXED Procedure				<u>-2 Log Likelihood = 59.1</u>			
Parameter Estimates							
Parameter	Estimate	SError	DF	t Value	Pr > t	Lower	Upper
phi1	4.6551	0.8209	12	5.67	0.0001	2.8666	6.4436
phi2	7.9920	1.4308	12	5.59	0.0001	4.8745	11.1095
beta	2.4800	0.6088	12	4.07	0.0015	1.1535	3.8066

Reduced Model

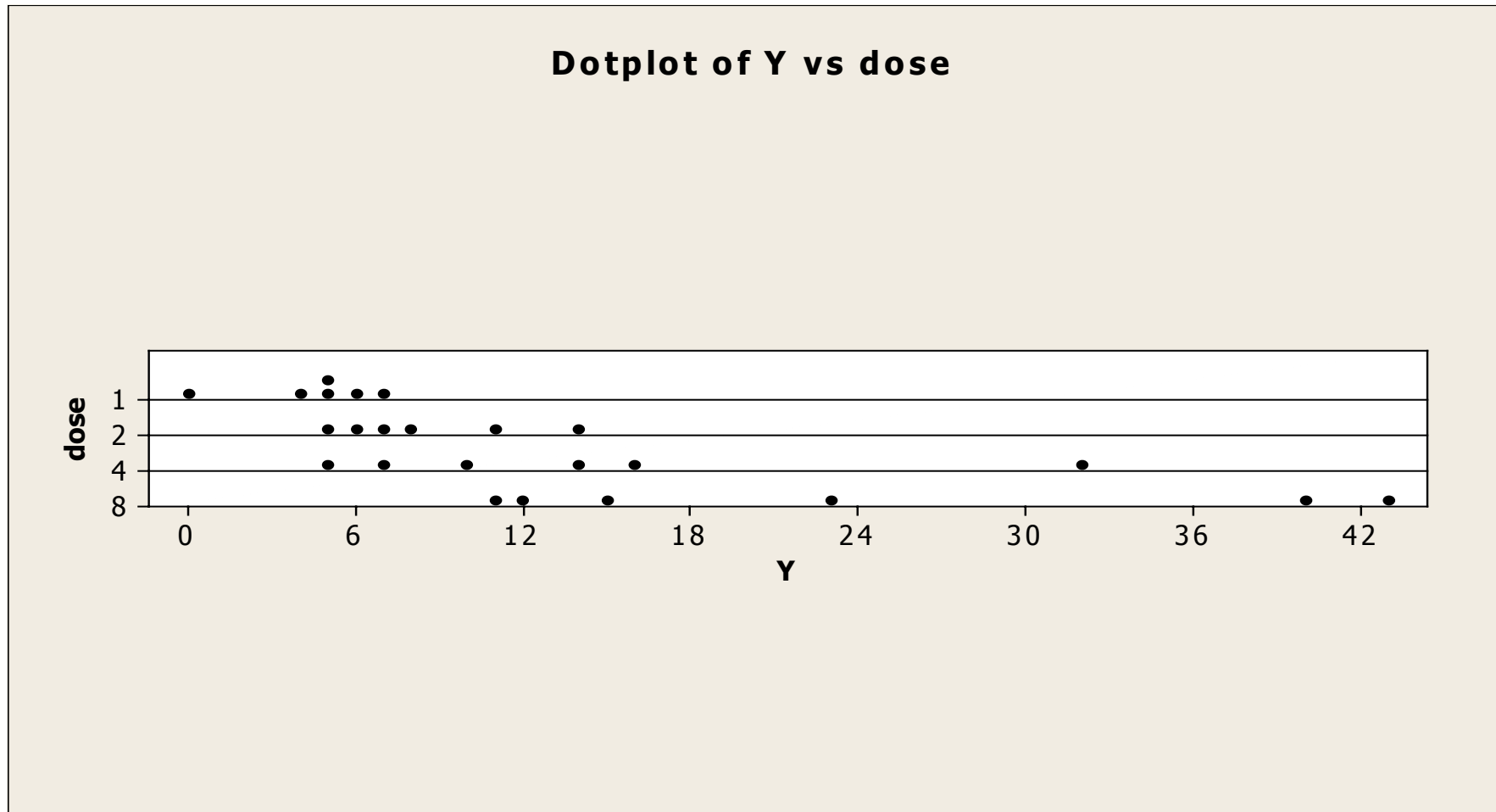
The NL MIXED Procedure				<u>-2 Log Likelihood = 63.1</u>			
Parameter Estimates							
Parameter	Estimate	SEerror	DF	t Value	Pr > t	Lower	Upper
phi	6.1355	1.0281	12	5.97	<.0001	3.8954	8.3757
beta	1.9919	0.4591	12	4.34	0.0010	0.9917	2.9921

Performing the test, we obtain the test statistic $\chi_1^2 = 63.1 - 59.1 = 4.0$, which has an approximate p-value of 0.0455, indicating a possible difference between the medians for the two treatments.

Another Example (Linear Regression)

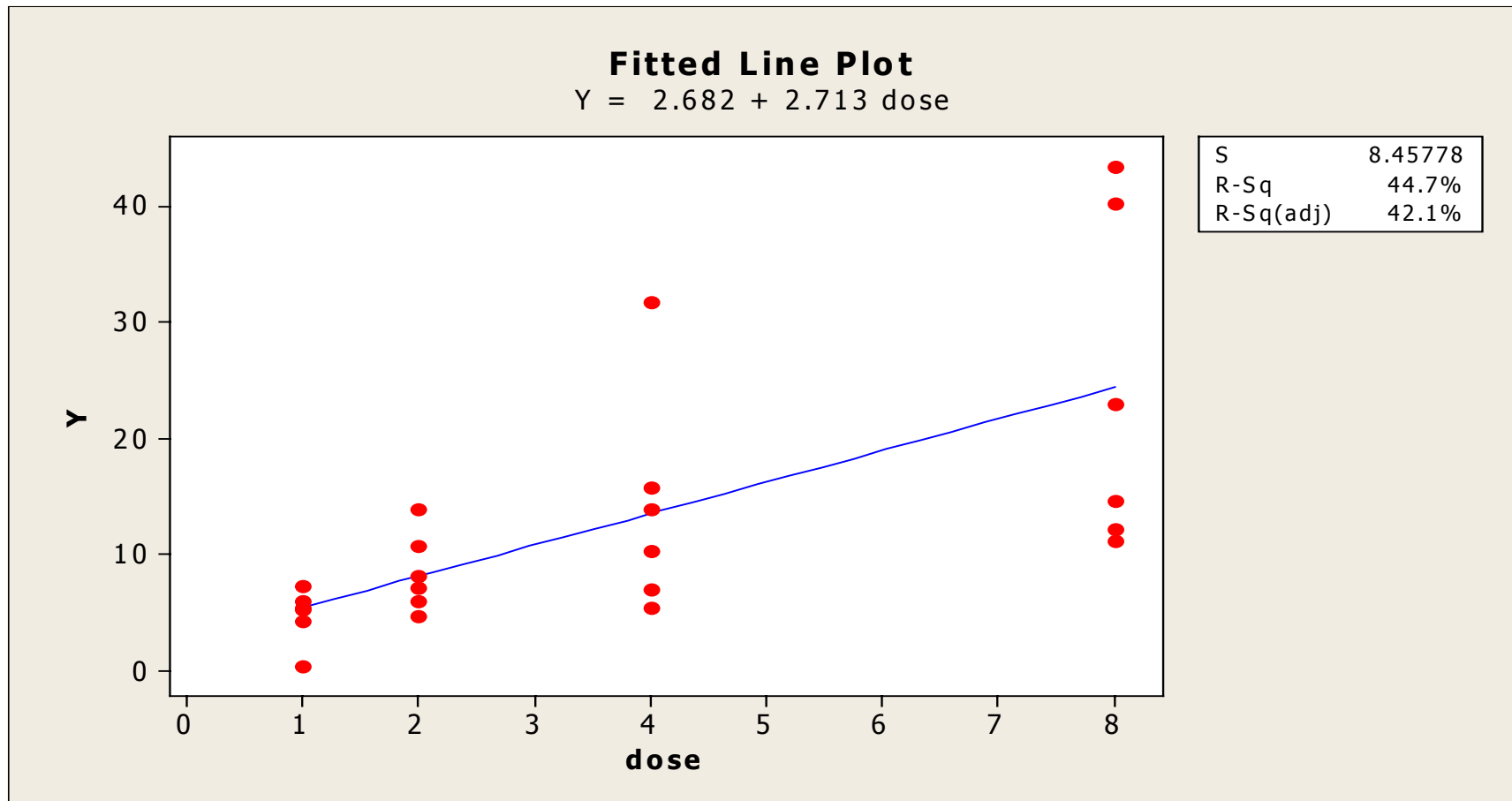
Suppose that a new drug, Zimodium, shows promise in prolonging life of cancer victims after diagnosis of aggressive type K cancer of the larynx. We randomize six larynx cancer patients to receive a dose = 1mg of Zimodium per day, six patients to receive 2mg per day, six patients to receive 4mg per

day, and six patients to receive 8mg per day, and we measure the time until the patient's cancer spreads to his or her throat (in days). Here are the data (no censored measurements).



Descriptive Statistics: Y

Variable	dose	Mean	Median
Y	1	4.798	5.385
	2	8.47	7.69
	4	14.03	12.16
	8	24.12	18.82



The above SLR is inappropriate here since it assumes Normality and constant variances. Instead, SLR holds that $\mu = \beta_0 + \beta_1 \cdot \text{dose}$ (where μ is the *mean* of the Normal distribution), we now extend this relation by modelling the Weibull *medians* and the relation, $\phi = \beta_0 + \beta_1 \cdot \text{dose}$. We also assume the same Weibull shape parameter β for each of the dose levels. SAS/NLMIXED gives:

The NLMIXED Procedure				<u>-2 Log Likelihood = 148.0</u>			
				Parameter Estimates			
Parameter	Estimate	SEError	DF	t Value	Pr > t	Lower	Upper
b0	1.6571	1.2298	24	1.35	0.1904	-0.8812	4.1953
b1	2.8767	0.6304	24	4.56	0.0001	1.5755	4.1778
beta	2.0631	0.3335	24	6.19	<.0001	1.3749	2.7514

The model fit with zero slope yields -2 Log Likelihood = 168.3, and the test statistic $\chi_1^2 = -2\Delta LL = 168.3 - 148.0 = 20.3$, p-value = 0.0000066; we thus reject the claim that the slope is zero.

➔ Extending Nonlinear models is also very straightforward.

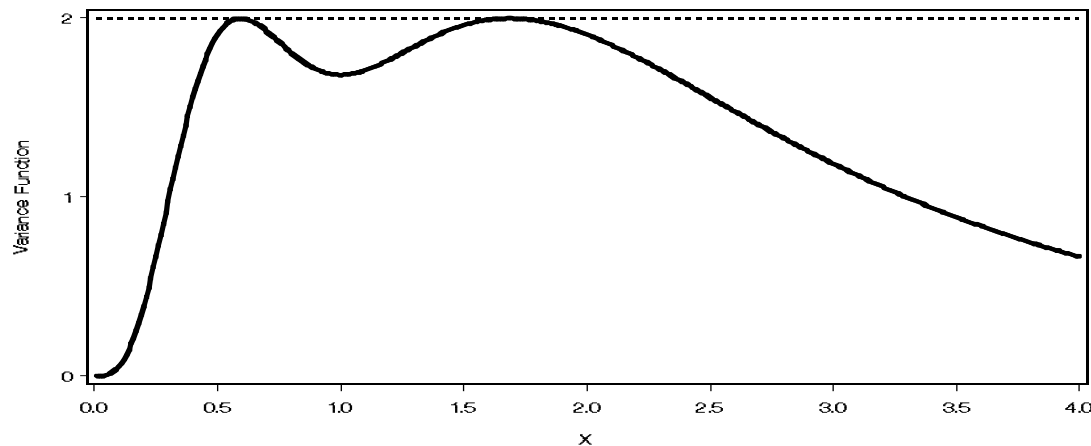
E. Design Considerations

The LL2 model function is written

$$\eta = \frac{1}{1+t} = \frac{1}{1+(x/\theta_2)^{\theta_3}}$$

The D-optimal design places the weight of $\omega = 1/2$ on only two support points regardless of the LD_{50} (θ_2) and slope (θ_3).

D-optimality of this design is confirmed by noting that the graph of the corresponding variance function does not exceed the line $y = 2$.

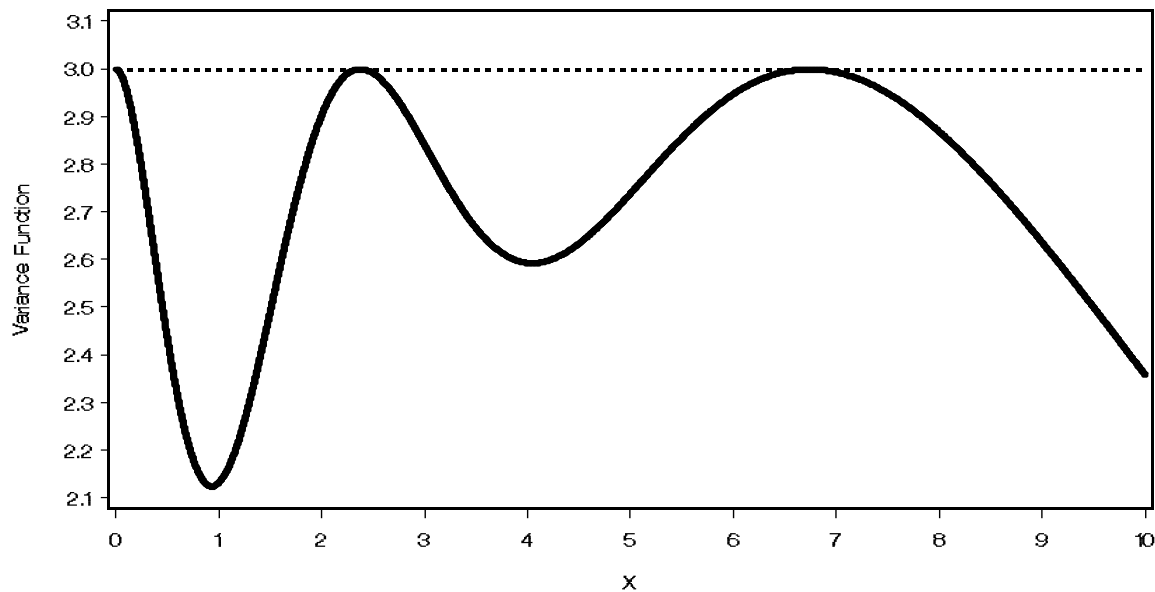


Note that optimal designs typically only have p ($= 2$ here) support points, and have thus been rightly criticized as providing no ability to test for lack of fit of the assumed model function (see below).

Further, the LL3 model function is written

$$\eta = \frac{\theta_1}{1+t} = \frac{\theta_1}{1 + (x/\theta_2)^{\theta_3}}$$

where the additional parameter here (θ_1) corresponds to the upper asymptote. Not surprisingly, the D-optimal design includes three points – the two indicated above – in addition to $x = t = 0$. The graph of the variance function confirms D-/G-optimality.

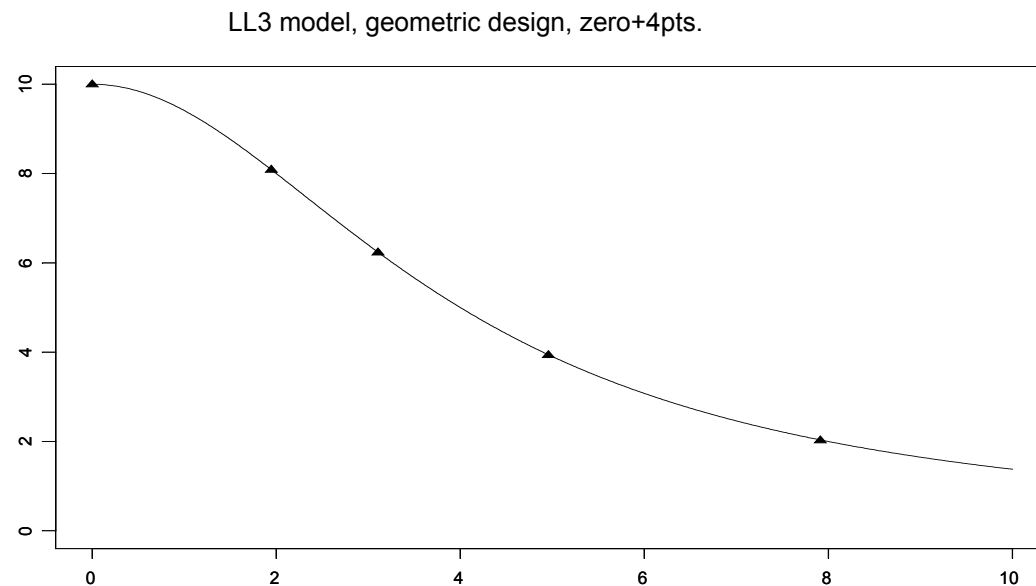


A rival strategy – provided the information loss is not too great – is to seek designs with a *geometric pattern*, as was used in the logistic example, and where the chosen design was $x = 0.28 \cdot (2)^k$, for $k = 0, 1, \dots, 4$. Here, for the LL3 model function, we choose

$$x_1 = 0, x_2 = a, x_3 = a \cdot b, x_4 = a \cdot b^2, x_5 = a \cdot b^3,$$

with the weight ω_1 placed at $x_1 = 0$ and the remaining $1 - \omega_1$ divided evenly across x_2 through x_5 . Choosing the D-optimality criterion, we thus seek designs to maximize $\det(\mathbf{M})$ over (a, b, ω_1) .

Regardless of the values of θ_2 and θ_3 , this approach produces designs for which the Y values are approximately 80%, 60%, 40% and 20% of the maximum value of Y, and are thus easily obtained provided one can draw a reasonable sketch of the anticipated model fit. This situation is graphed below for $\theta_2 = 4$ and $\theta_3 = 2$, for which $a = 1.945$, $b = 1.597$ and $\omega_1 = 0.326$; the x-values are then approximately 0, 2, 3, 5 and 8.



Note that the weight at $x = 0$ is approximately $1/3$, meaning that a final sample size of the form $N = 6n$ is indicated.

We can assess the loss of information of using the design ξ relative to the three-point D-optimal design (ξ_D) by using the D-efficiency,

$$D_{\text{EFF}} = [\det(\xi)/\det(\xi_D)]^{1/p}$$

In this case, it turns out that $D_{\text{EFF}} = 94.8\%$, meaning that the use of this 5-point geometric design only results in about a 5% information loss, yet provides us with a means to test for lack-of-fit and may be more practical. Incidentally, a rival geometric design – which takes measurements at $x = 0, 1, 2, 4, 8$ – results in an information loss around 12.5%.

F. Comments, Recommendations, and Conclusions

1. Nonlinear models are often more useful and appropriate than are linear ones;
2. Easy-to-use software packages and procedures now permit us to fit even Exponential Family nonlinear models with and without complicated error structures – this represents a big improvement over old methods such as transforming variables and wishful thinking;
3. We can thus more directly model our data and answer the relevant research questions;
4. Robust design strategies and “rules of thumb” need to be provided to practitioners to optimize information.

Thank You!