Special Topics

Diagnostics

Residual Analysis

As with linear regression, an analysis of residuals is necessary to assess the model adequacy. The same techniques may be employed such as plotting residuals against predicted values and regressors as well as univariate summaries on residuals. These should be examined for any obvious patterns or trends which may suggest heterogeneity or lack of fit. PROC NLIN allows for the output of either standard (**R=** option) or studentized (**STUDENT=**option) residual values.

Nonlinearity

In nonlinear regression, the properties of the estimated solutions is not as clear cut as the linear case where it may be reasonable to assume that the parameter estimates follow a multivariate normal distribution. Nonlinear estimates may exhibit deviations (called nonlinearity in the solution) which results in parameter confidence intervals and regions which are distorted and unsymmetrical compared to their linear counterparts. Since NLIN reports the linear approximations for statistical inference (asymptotic confidence intervals) it is necessary to assess the degree of this distortion. Graphical aids available in assessing the nonlinearity of parameter estimates include profile t-plots and profile pair sketches. The necessary program codes to compute these measures are somewhat complex and thus, not provided here. Users interested in implementing such techniques are encouraged to do so with the help of a statistician.

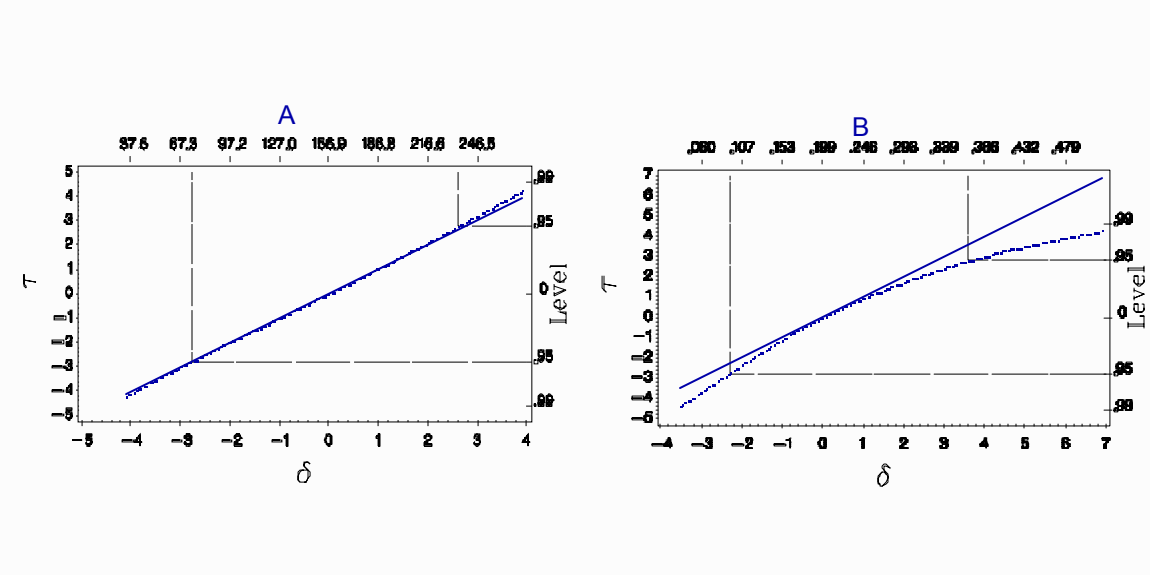
The following references may be useful:

Bates, D. M. and D. G. Watts. 1988. Nonlinear Regression Analysis and its Applications. John Wiley and Sons. New York.

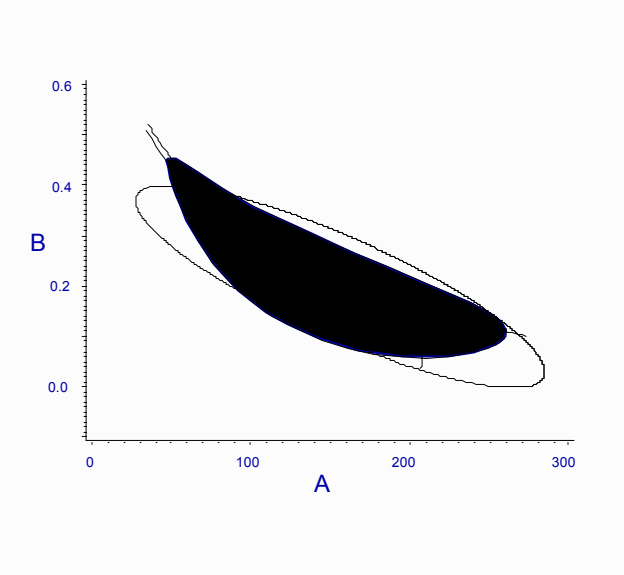
Shafii, B., W. J. Price, J. B. Swensen, and G. A. Murray. 1991. *Nonlinear Estimation of Growth Curve Models for Germination Data Analysis*. *In* Applied Statistics in Agriculture, G. A. Milliken and J. R. Schwenke eds., pp 19-42. Kansas State University, Manhattan, KS.

Price, W. J. and B. Shafii. 1992. *Graphical Aids fir Summarizing Inferential Results of Nonlinear Regression*. Proceedings of the Seventeenth Annual SAS User's Group International Conference. pp. 1329-1334. SAS Institute Inc, Cary NC.

An example for the use of t-plots and pair sketches can be made with the exponential data. Each parameter produces a t-plot in which the degree of nonlinearity present is proportional to the deviation of the profile function (dashed line) from the solid diagonal reference line which represents the linear case. If the nonlinearity is small the two lines will coincide, as is the case for parameter **a**. True or nominal confidence intervals can also be read from the t-plot by selecting the confidence level on the right side and tracing it through the profile function to the top axis. Thus, a t-plot such as the one shown for parameter **b** will give a more unsymmetrical confidence interval.



The profile pair sketch gives an idea as to the distortion in the joint distribution of the parameters. In this example, the 95% nominal region is bent and distorted compared to the overlaid elliptical normal approximation, suggesting that joint inferences may not be accurate at the 5% significance level.



Linearization

Before modern computing ability was common place, it was standard practice to transform nonlinear models to linear forms when possible, since linear regression is computationaly simpler. For example, the exponential model given earlier can easily be linearized by taking the logarithm of each side:

Original Model: $\mathbf{y = ae^{-bx}}$

Linearized Model: $\mathbf{Log(y) = log(a) - bx}$

or: $\mathbf{y^* = a^* - bx}$

Transformations like this will usually give estimates close to those of nonlinear regression, but a trade off is made on inferences. Any statistical inference must be made on the log scaled estimates. In addition, linearized models may result in estimates that are not biologically meaningful. Although this technique is widely practiced in some disciplines, nonlinear regression along with greater computing capabilities provide users with a better solution (literally!).

Parameterization

Nonlinear estimation may be sensitive to the parameterization of the expectation function (how the model is expressed). In certain forms, the parameter correlations may be significantly high or the nonlinearity of the solution may be unworkable. In such cases the model should be rewritten (reparameterized). The logistic model used earlier provides a good example. The original model was given as:

$$y = M/(1 + \exp(-B(time - L)))$$

which could be reparameterized by multiplying the **B** parameter in to the **(time - L)** section yielding:

$$y = M/(1 + \exp(L^{*} - Btime)) \quad .$$

Intuitively, this may seem simpler, however, for estimation (and interpretation) purposes, this model is not good. The parameter estimates and correlations for both model forms are given below:

Original Model: Biotype R

Parameter	Estimate	Asymptotic Std. Error	Asymptotic 95 % Confidence Interval Lower	Asymptotic 95 % Confidence Interval Upper
M	99.27603715	0.37137504235	98.544410201	100.00766410
L	11.36767872	0.04082659073	11.287248347	11.44810910
B	1.38998060	0.06903338146	1.253981468	1.52597973

Asymptotic Correlation Matrix

Corr	M	L	B
M	1	0.198073651	-0.175030454
L	0.198073651	1	-0.035966444
B	-0.175030454	-0.035966444	1

Reparameterized Model: Biotype R

Parameter	Estimate	Asymptotic Std. Error	Asymptotic 95 % Confidence Interval Lower	Asymptotic 95 % Confidence Interval Upper
M	99.27603713	0.37137504211	98.544410186	100.00766408
L	15.80085307	0.78476012414	14.254837349	17.34686879
B	1.38998062	0.06903338285	1.253981484	1.52597975

Asymptotic Correlation Matrix

Corr	M	L	B
M	1	-0.160704791	-0.175030453
L	-0.160704791	1	0.9973853966
B	-0.175030453	0.9973853966	1

The reparameterized model gives reasonable parameter estimates, but the asymptotic parameter correlations between parameters **B** and **L** has increased severely. This indicates that the model is not expressed adequately. Furthermore, the interpretation of the parameter estimates of this model is not as direct. For instance, parameter **L** now estimates the product of the rate parameter **B** and the time to 50% germination (T_{50}). Thus, in order to estimate the T_{50} value, the estimate of **L** must be divided by the estimate for **B**. That is:

$$T_{50} = L/B = 15.8/1.39 = 11.37 \ .$$

Although this value is similar to the original model estimate, inferences such as confidence intervals on such a ratio would be difficult with standard statistical techniques. Overall, the original parameterization was better and provided more easily interpreted estimates.

Model Comparison

One of the most powerful uses of nonlinear regression is the ability to examine structural differences between data classes, such as treatments, environments, genetics, etc. The process involves fitting a model to each class, carefully diagnosing

each model, and finally comparing the models either by parameters or overall. This requires that the models be of the same form, e.g. all logistic models. Similar to dummy variable regression illustrated in PROC REG (SAS workshop #4), the technique is to build a full model containing all treatments, and then conduct contrasts among the parameters. As with diagnostics, the programming required is lengthy and case specific, so it is not given here. An illustration can be done, however, using the prickly lettuce data.

Example 3

One question that may be asked about the resistant biotype seedlings relates to their fitness relative to the susceptible biotype. Thus, a comparison of the respective germination curves may provide some answers. The two models are:

Resistant Model

$$\mathbf{germ = M_R/(1 + \exp(-B_R(time-L_R)))}$$

Susceptible Model

$$\mathbf{germ = M_S/(1 + \exp(-B_S(time-L_S)))}$$

Some appropriate hypotheses would be:

1) Are the final levels of germination the same?

H_0: $\mathbf{M_R = M_S}$

2) Are the rates of germination the same?

H_0: $\mathbf{B_R = B_S}$

3) Is there any delay or lag in the germination?

H_0: $\mathbf{L_R = L_S}$

4) Is the overall germination process different?

H_0: $\mathbf{M_R = M_S}$,

$\mathbf{B_R = B_S}$,

$\mathbf{L_R = L_S}$

The results of the hypothesis testing are given below. This type of analysis can be more informative than say, an analysis of variance, because details of differences among classes can be highlighted (i.e. germination levels, rates, and lags). ANOVA can only ascertain differences in overall means with no specifics on where the differences may originate.

Note that the following are based on asymptotic F tests which rely on the linear approximations mentioned earlier. Therefore, the prior use of nonlinear diagnostics is necessary to prevent erroneous conclusions.

Estimates

M_r	99.276037
L_r	11.367679
B_r	1.3899806
M_s	98.325116
L_s	14.756919
B_s	0.9584168

Contrasts:

Mr vs Ms

F1	P1
2.2278141	0.1362108

Br vs Bs

F2	P2
2051.5915	0

Lr vs Ls

F3	P3
21.843078	3.8617E-6

Curves

F4	P4
743.31004	0

From this, we conclude that the final levels of germination are equivalent, but that the rate of germination and the lag time to 50% germination are significantly different. Biotype R germinates sooner and at a faster rate than biotype S. Overall, the two biotypes possess different germination processes.