Journal of Data Science **7**(2009), 235-253

Encouraging Students to Think Critically: Regression Modelling and Goodness-of-Fit

Timothy E. O'Brien[1], Suree Chooprateep[2] and Gerald M. Funk[1]
[1]*Loyola University Chicago and* [2]*Chiang Mai University*

Abstract: This note underscores important considerations that should be taken into account when teaching students to check for inadequacies of a given linear, nonlinear or logistic regression models. Key illustrations are provided which underscore the shortcomings of currently used procedures. A brief overview of nonlinear regression models is given in order to lay the foundation for testing for lack of fit in nonlinear models. This paper also introduces a new 'scaled' binary logistic regression model to highlight potential problems with the usual logistic model, and implications for choosing a robust optimal experimental design are also underscored and discussed.

Key words: Lack of fit, logistic regression, nonlinear regression, optimal design, robustness, scaled logistic regression.

1. Introduction

Box (1979) reminds us that although no statistical model is ideal, some models are useful and beneficial for accurately representing diverse phenomena and mechanisms. Thus, in our statistics courses we underscore that scientific researchers often find that linear, generalized linear, nonlinear, and survival regression models are helpful for modelling various biological, chemical and medical processes. We also point out that once such a model is fit to a given set of data, it is incumbent upon the researcher or statistician to check the assumed model for inadequacies - that is, the so-called 'lack of fit' of the model. Indeed, standard regression textbooks such as Seber and Wild (1989); Lindsey (1997, 2001); Draper and Smith (1998); Krzanowski (1998); Rawlings et al (1998); Collett (2003a,b); and Seber and Lee (2003) stress the importance of checking for such model mis-specification. Additional regression and modelling references include Ratkowsky (1983, 1990); Bates and Watts (1988); Huet et al (1996); Harrell (2001); and Agresti (2002, 2007).

Through a series of key illustrations, this paper helps students see the potential inadequacies associated with the usual goodness of fit tests for popular regression models, and proposes important steps to take to guard against incorrect conclusions. These methods are given for simple linear regression models in Sections 2 and 3. Background and suggestions to detect lack of fit in nonlinear models are given in Sections 4 and 5. In Section 6, the assessment of model misspecification in binary logistic regression is discussed after we introduce a new methodology to help determine the correct scale to use for this model. Finally, the implications for the prudent choice of an efficient experimental design is given and discussed in Section 7.

2. Testing for Lack of Fit in Linear Models

We adopt here the usual notation used for linear regression models in introductory textbooks. Thus, the homoskedastic Gaussian linear model is written $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where the vector of errors are assumed to follow a $N_n(\mathbf{0}, \sigma^2\mathbf{I})$ distribution. In this setting, the estimator $\mathbf{b}$ that minimizes the error sums of squares function,

$$S(\boldsymbol{\beta}) = \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 = [\mathbf{y} - \mathbf{X}\boldsymbol{\beta}]^T[\mathbf{y} - \mathbf{X}\boldsymbol{\beta}]$$

is both the least-squares estimator and maximum likelihood estimator. The predicted response vector is then $\mathbf{y}^* = \mathbf{X}\mathbf{b} = \mathbf{P}_X\mathbf{y}$. In this expression, $\mathbf{P}_X$ is an idempotent matrix, and since $\mathbf{P}_X\mathbf{X} = \mathbf{X}$, this projection matrix projects onto $C(\mathbf{X})$, the column space of $\mathbf{X}$. Further, the residual vector is $\mathbf{e} = \mathbf{y} - \mathbf{y}^* = (\mathbf{I} - \mathbf{P}_X)\mathbf{y}$, and the residual sum of squares is denoted here $RSS = S(\mathbf{b}) = \mathbf{e}^T\mathbf{e} = ||\mathbf{y} - \mathbf{X}\mathbf{b}||^2 = \mathbf{y}^T(\mathbf{I} - \mathbf{P}_X)\mathbf{y}$.

In many instances, it is prudent to write the above model as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = [\mathbf{X}_1|\mathbf{X}_2]\boldsymbol{\beta} + \boldsymbol{\epsilon} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$$

Here, $\mathbf{X}$ is of dimension $n \times p$, $\mathbf{X}_1$ is $n \times (p - q)$, and $\mathbf{X}_2$ is $n \times q$. When our interest centers on testing the hypothesis $H : \boldsymbol{\beta}_2 = 0$ versus $A : \boldsymbol{\beta}_2 \neq 0$, the relevant test is the so-called Full-and-Reduced F test,

$$F^* = \frac{(RSS_H - RSS)/q}{RSS/(n - p)} \tag{2.1}$$

In this expression, RSS is the above residual sum of squares whereas RSS_H is the value of RSS under the constraint imposed under the null hypothesis H; when the H is true, F^* in equation (2.1) has a central F distribution with q and $(n - p)$ degrees of freedom. Following Seber and Lee (2003, p.100), we can write $RSS_H - RSS = \mathbf{y}^T\mathbf{P}_2\mathbf{y} = ||\mathbf{P}_2\mathbf{y}||^2$ and $RSS = \mathbf{y}^T(\mathbf{I} - \mathbf{P}_X)\mathbf{y} = ||(\mathbf{I} - \mathbf{P}_X)\mathbf{y}||^2$. Here $\mathbf{P}_2$ is the projection matrix associated with $\mathbf{R}_2 = (\mathbf{I} - \mathbf{P}_{X_1})\mathbf{X}_2$, the subspace of $C(\mathbf{X}_2)$ which is orthogonal to $C(\mathbf{X}_1)$.

Interestingly, for datasets involving at least one repeat observation, students are often initially surprised to learn that the usual Lack of Fit (LOF) test uses the same F test statistic as the one given in equation (2.1); see for example in Draper and Smith (1998, p.47). But, in the case of the LOF test, there is an important distinction with the above idea of attempting to find a more parsimonious sub-model: for LOF, we are looking at how our assumed linear model fares *when compared with the highest order polynomial that can be fit to the data.* Thus, if our assumed model is $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}$, the LOF test assesses whether the additional terms, $\mathbf{X}_2\boldsymbol{\beta}_2$, are necessary so that really the model should be $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$ instead.

Example 1. Consider the data plotted in Figure 1, wherein the simple linear regression model is assumed to accurately describe these data.

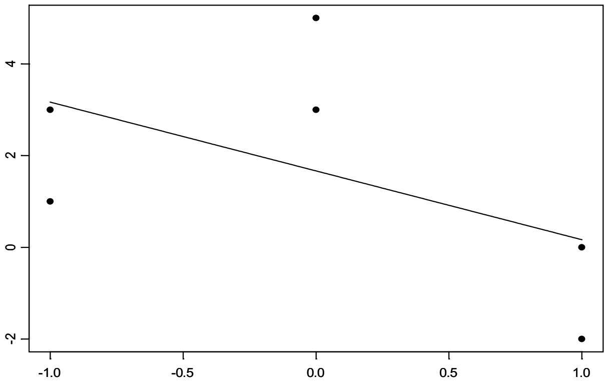

Figure 1: Plot of data and fitted line for Example 1

Since these data involve repeated measurements and since the design involves three support points, the highest order polynomial that can be fit to these data is a quadratic regression model, $y_k = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_k + \boldsymbol{\beta}_2 x_k^2 + \epsilon_k$. As our students learn, in general, the so-called "Full model" for a repeated measurement-design with S support points is an $(S-1)^{st}$-degree polynomial, which is equivalent to the one-way ANOVA model with S levels of the factor. Thus, for these data, the LOF test is equivalent to testing $H : \boldsymbol{\beta}_2 = 0$ in the larger (quadratic) model. Here, the LOF F statistic ($F_{1,3} = 8.17, p = 0.065$) is just the Full and Reduced F statistic comparing the assumed line with the highest order polynomial (i.e., quadratic here). Although the p-value of 6.5% (barely) exceeds the usual cut-of of 5%, one might be somewhat suspicious of the assumption of linearity for these data in light of the above graph.

It is also important to point out to students that the Lack of Fit test compares the assumed linear model with the highest order polynomial, and so *depends to a large degree upon the chosen design.* This point is discussed further in Section 7 and illustrated in the next section.

3. Some Cautions Related to Linear Models and Lack of Fit

It turns out that sometimes the usual Lack of Fit test — which compares the model function $\eta = \mathbf{X}_1\boldsymbol{\beta}_1$ with the highest-order polynomial model function $\eta = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2$ — may miss some important intermediate models, and thus may miss the inadequacy of the assumed model function $(\eta = \mathbf{X}_1\boldsymbol{\beta}_1)$. Here is an illustration.

Example 2. A line is fit to the data plotted in the following graph, involving a 4-point design with 2 replicates at each support point.

For these data, the assumed model (the line) looks inadequate, but the lack of fit test below indicates otherwise $(F_{2,4} = 3.73, p = 0.122)$. It turns out that what is masked here is that the quadratic effect is significant, and this is not detected in the above Lack of Fit (LOF) test *since the LOF test lacks power to detect intermediate departures from the assumed line.* We see this by using orthogonal polynomials, which remove the inherent confounding so that the linear, quadratic and cubic effects are each estimated in an orthogonal or "independent" manner. When we do this for these data, the quadratic effect appears to be marginally non-significant since $p = 0.053 > 0.05$. However, this underscores our second warning: that all available degrees of freedom (and corresponding sums of squares) should be used in the estimation of σ^2 in small studies such as this one. Since it is clear that the cubic effect is not significant here $(p = 0.767)$, its sums of squares and single degree of freedom should be absorbed into the estimation of σ^2. When this is done, our estimate of σ^2 changes from 0.495 on 4 degrees of freedom (dfs) to 0.5075 on 5 dfs, and this increased power gives a p-value for the quadratic term of $p = 0.030$.

In sum, for these data the LOF test indicates that the line is adequate even though the quadratic model fits the data much better. This situation arises since the LOF test here compares the line with the cubic model, and thus misses the important intermediate (quadratic) model.

The above example points out another important consideration, demonstrated using the following ANOVA table associated with this data.

Source	DF	SS	MS	F	p
Regression	1	16.200	16.200	68.51	0.000
Residual Error	6	1.419	0.236		
Lack of Fit	2	0.924	0.462	3.73	0.122
Pure Error	4	0.495	0.124		
Total	7	17.619			

The worse-case scenario would be one where all of the “0.924” LOF sum of squares (SS) is associated with only one degree of freedom (df) and zero SS with the other df. Then, the *worse case LOF test statistic* would be $F_{1,5} = (0.92375/1)/(0.4950/5) = 9.33$, and the corresponding p-value would be $p = 0.0283$. That is, we need to be mindful of the matrix $\mathbf{R}_2 = (\mathbf{I} - \mathbf{P}_{X_1})\mathbf{X}_2$ and of this worse-case setting, which provides a *lower bound for the LOF p-value.* If this lower-bound p-value is high (e.g., over 0.05), then the model is probably adequate; if it is not, then further analysis is needed wherein the some intermediate models should be examined and compared with the assumed model.

4. Overview of Nonlinear Modelling

Before discussing the assessment of model mis-specification for nonlinear models, we first provide students with some helpful background in nonlinear regression models; additional results are given in Ratkowsky (1983, 1990); Bates and Watts (1988); Seber and Wild (1989); and Huet *et al.* (1996).

Indeed, whereas linear models may be used at the preliminary stages of studying a given process, nonlinear models are often more appropriate and useful as one’s knowledge and understanding of the mechanism deepens. Interestingly, nonlinear models also arise in situations involving linear models but where interest centers on a nonlinear function of random variables; examples include Pearson’s correlation coefficient in simple linear regression and relative potency involving the ratio of two Gaussian sample means. Thus, in contrast with linear models, model parameters in nonlinear models are often paramount and have important practical interpretations. Since the class of nonlinear regression models is quite large, we focus in this Section only on some important two-parameter nonlinear models.

Two popular examples of concave asymptotic growth models are the SE2 and MM2 model functions given in the following table. The latter model has widespread application in modelling enzyme kinetics. In both models, θ_1 is the upper asymptote for each of these model functions. In the MM2 model, θ_2 is the so-called LD_{50} parameter, so that θ_2 is the value of the explanatory variable whereby the expected response is one-half the value of the upper asymptote. In both models, the expected response is assumed to be zero at $x = 0$; when this is unreasonable, a third parameter can be added to each of these models to shift the curve up or down.

In contrast, the IP2 model function with $\theta_1 > \theta_2$ is useful for modelling the movement of drugs through the body or chemicals through 'compartments'; see for example, Box and Lucas (1959) and Atkinson and Donev (1992, p.193ff). The corresponding curve starts at the origin indicating the belief that the expected concentration in the target organ is zero when the tablet is ingested. Key functions of the model parameters are the value of time (x) where the maximum concentration is reached (t_{MAX}), the expected maximal concentration at this time point (C_{MAX}), and the area under the IP2 curve ($AUC = 1/\theta_2$).

Commonly-Used 2-parameter Nonlinear Models

Model Function Name Model	Function Equation
Simple Exponential (SE2)	$\eta_A = \theta_1\,(1 - \exp(-\theta_2 x))$
Michaelis-Menten (MM2)	$\eta_B = \theta_1 x/(\theta_2 + x)$
Intermediate product (IP2)	$\eta_C = \frac{\theta_1}{\theta_1 - \theta_2}\,(\exp(-\theta_2 x) - \exp(-\theta_1 x))$
Log-Logistic (LL2)	$\eta_D = \frac{(x/\theta_2)^{\theta_4}}{1+(x/\theta_2)^{\theta_3}}$

Another key two-parameter growth model is the LL2 model function, wherein θ_2 is the LD_{50} and $\theta_3 > 0$ is the slope. This expression can be rewritten as $\eta = t/(1+t)$ for $t = (x/\theta_2)^{\theta_3}$, with the corresponding decay function is thus given by $\eta = 1/(1+t)$. The LL2 model function is equivalent to the MM2 model function when θ_3 in the LL2 model function and θ_1 in the MM2 model function are equal to unity. This is an important generalization of the MM2 model since sometimes growth takes place first at an increasing rate and then at a decreasing rate. Of course, if the upper asymptote is not known to be equal to one, then an upper asymptote (θ_1) can be added to the LL2 model, resulting in the LL3 model given below in Equation (5.4).

Technical details for preferring the F-statistic approach in Equation (2.1) to other methods for nonlinear models are given in Seber and Wild (1989, chap. 5). Briefly, since intrinsic curvature is usually negligible (Bates and Watts, 1988, chapter 7), these likelihood-based methods are considered nearly exact in terms of coverage of nominal confidence intervals. Further, Gallant (1987, chap. 1) argues for their superiority based on statistical power.

5. Testing for Lack of Fit in Nonlinear Models

As might be expected, testing for lack of fit (LOF) is somewhat more involved in nonlinear models than in linear models since the corresponding intermediate models (between the assumed and the saturated models) are less obvious. The model assumptions here are similar to those for linear models — that the response variables (denoted Y) are uncorrelated, from a Gaussian distribution and with common variance — but with the exception that the mean function here is a specified nonlinear function (in the parameters). Of course, the global LOF test — wherein the assumed model is compared with the highest order polynomial (i.e., the one-way ANOVA model) — is the same as for linear models.

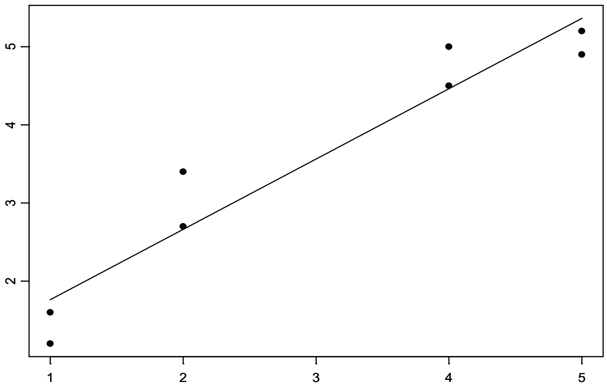

Figure 2: Plot of data and fitted line for Example 2

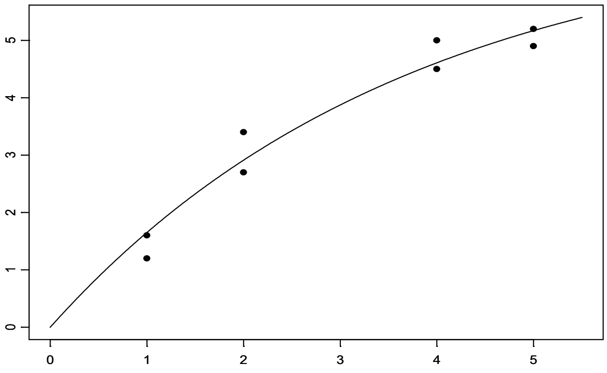

Figure 3: Plot of data and fitted SE2 model function for Example 2 (continuation)

Example 2 (continuation). To give a practical context to these data examined above, if the model here relates the expected yield (i.e., $E(Y)$) to the amount of fertilizer applied (x), then we might then posit the homoskedastic Gaussian SE2 model function. As pointed out above, this is a concave growth curve, and use of this model is predicated upon the belief that crop yield increases at a decreasing rate with the additional application of more and more fertilizer. Thus, nonlinear modelling in practice is partly based on choosing a model that fits the data and partly on (scientific or otherwise) '*common sense*'. The data and fitted curve are graphed above.

For these data, the estimates of θ_1 and θ_2 are 6.98 and 0.27 respectively and the residual sum of squares (RSS) is 0.7279. In this case, the format of the LOF test is as for linear models: here we calculate $F_{2,4} = [(0.7279 - 0.4950)/2]/[0.4950/4] = 0.9410(p = 0.4625)$. More importantly, the worse case LOF test statistic, $F_{1,5} = [(0.7279 - 0.4950)/1]/[0.4950/5] = 2.35(p = 0.1857)$, indicates the adequacy of this assumed SE2 model function.

Of course, the above F test is only appropriate for *nested models*. To illustrate, the linear model ($\eta = \beta_0 + \beta_1 x_k$) is nested in the quadratic model ($\eta = \beta_0 + \beta_1 x_k + \beta_2 x_k^2$) since the latter model reduces to the former when $\beta_2 = 0$. We also say that this quadratic model generalizes the linear one. Three- and four-parameter generalizations of the SE2 model include the WEIB3 and WEIB4 (three- and four-parameter Weibull) model functions:

$$\eta_E = \theta_1 \left(1 - \exp\{-(\theta_2 x)^{\theta_3}\}\right) \tag{5.1}$$

and

$$\eta_F = \theta_4 + (\theta_1 - \theta_4)\left(1 - \exp\{-(\theta_2 x)^{\theta_3}\}\right) \tag{5.2}$$

These model functions have an inflection point whenever $\theta_3 > 1$. In the WEIB4 model, θ_4 is the lower asymptote, which is taken to equal zero in the WEIB3 model. Also, the WEIB4 model reduces to the SE2 model when $\theta_3 = 1$ and $\theta_4 = 0$. Fitting the WEIB4 model is equivalent to the Full (cubic) model in Example 2 since there are four support points.

On the other hand, there may be other — non-nested — models which perform better than the SE2 model function in cases such as Example 2. In the case of non-nested models, regression textbooks point out that one can chose the preferred model on the basis of the so-called Akaike's Information Criterion, $AIC = 2f(\theta^*) + 2p$, where f is the negative of the marginal log-likelihood function and p is the number of model function parameters; discussion of the AIC can be found in Harrell (2001) wherein it is recommended that a model be chosen which produces the lowest value of AIC. For example, using the AIC measure for the above illustration, the SE2 model function is preferred to the MM2 and LL3 model functions, but the quadratic curve is preferred to the SE2.

Issues of model selection are sometimes challenging for nonlinear models — especially when contrasted with linear models — since one is never completely assured that the 'best' model is even one under consideration. Our approach has always been to be very well versed in models in use, with any shortcomings associated with these models, and with connections between these various models; this latter point is illustrated in Section 7 below. Additional comments regarding model selection are given in standard texts such as Lindsey (2001, chap. 2).

For the above example, even though the AIC is lower for the fitted quadratic model than for the SE2 model for these data, the estimated quadratic model predicts negative yield for $x < 0.27$ and for $x > 9.73$ — which is not feasible; the model also proposes a decline in yield for $x > 5$, which may also be unreasonable. That is, for nonlinear models, the assessment of '*lack of fit*' must be performed using the calculated measures in conjunction with some degree of *common sense* or *expert knowledge* related to the process or phenomenon under study. Thus, for these data, we would probably want to choose the SE2 model function as the most reasonable function and deem it the '*best fit*' for these data.

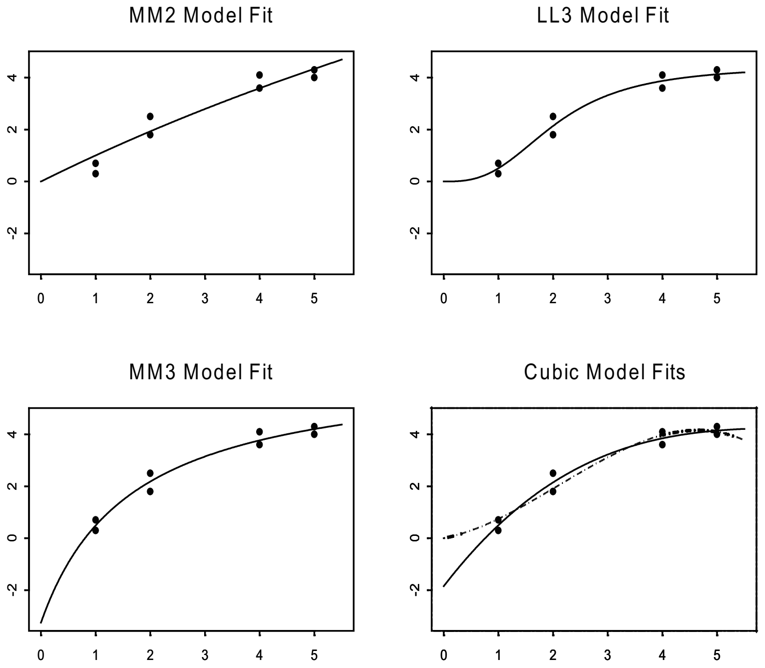

Figure 4: Plots of fitted MM2, LL3, MM3 models and two Cubic fits and the modified Example 2 data

To go one step further along this line of reasoning and to deepen our understanding of lack of fit for nonlinear models, we now change of our original data in Example 2 in the following manner: the new yield values (Y's) are obtained by subtracting 0.90 from each of the original Y's; the fertilizer amounts (X's) remain the same.

Example 2 (modification). For the new data, the LOF test for the line is as for the original data since the data have just been shifted downward; however, the predicted line (and quadratic curve) is negative over a larger range, and hence would be unreasonable for the practical reasons given above.

Using the AIC measure for these modified data, the MM2 model function is preferred to the SE2 model, although it too is inadequate as shown in the top left of the following graph.

In the above graph, we plot the MM2, the LL3 (with the inflection point), the MM3, and two cubic model fits (the dashed one forced through the origin) along with the modified data. The MM3 expression is given by the equation,

$$\eta_G = \theta_4 + \frac{(\theta_1 - \theta_4)x}{\theta_2 + x} \tag{5.3}$$

We also easily see now how important is our '*expert knowledge*' about the behavior of $E(Y)$ — expected yield here — when $x = 0$ since the estimate of θ_4 in this MM3 model is negative. In the present context, even though the MM3 model fits the data well, it would thus be rejected as being unreasonable since negative yield is impossible. The same logic leads us to reject the cubic models since the unrestricted cubic model predicts negative yields whereas the restricted cubic model shows a declining curve for $x > 5$.

As indicated above, the LL3 model function is given by the equation

$$\eta_H = \frac{\theta_1 (x/\theta_2)^{\theta_3}}{1 + (x/\theta_2)^{\theta_3}} \tag{5.4}$$

This model function generalizes the MM2 model function by adding a slope parameter (denoted θ_3 here); thus, when $\theta_3 = 1$, the two model functions coincide, demonstrating that the MM2 model is nested in the MM3 model. Using the corresponding SAS computer output given in the Appendix and noting that the value of one is in the "Approximate 95% Confidence Interval" for θ_3, we are lead to believe that we could accept the MM2 model. This confidence interval is the so-called 'Wald Interval', and is based on a linear approximation of the model function. Equivalently, the corresponding Wald test statistic here is $t_5^* = (2.8325 - 1)/0.8028 = 2.283$, yielding the p-value $p = 0.0713$.

On the other hand, the Full-and-Reduced F test — that is, the likelihood-based test statistic in Equation (2.1) - yields

$$F = (1.3014 - 0.4974)/0.0995 = 8.0804,$$

and the corresponding p-value of $p = 0.0361$. This discrepancy underscores the fact that the Wald linear approximation can be and often is inappropriate, and emphasizes the superiority (increased power) of the Likelihood Full-and-Reduced F test. Our conclusion here is that these data are probably best modeled using the LL3 model; this conclusion is based both on the LOF and the Full-and-Reduced F tests, and on scientific common sense.

The above example underscores the intricate nature of testing for LOF in nonlinear models, and how this test must be used in conjunction with an understanding of the specific subject matter under investigation. Caution also needs to be exercised with binary logistic regression modeling, considered next.

6. Testing for Lack of Fit in Binary Logistic Regression

As for Gaussian nonlinear models, testing for model adequacy in binary logistic regression can also be challenging and more involved than for linear models. As pointed out in introductory regression texts, binary logistic regression typically involves taking measurements at a series of x values denoted here by $x_1, \ldots, x_w$. Then, for a given value $x = x_k$, an independent set of n_k binary experiments are performed, and the binary logistic regression model posits that the response random variable has a binomial distribution with parameters n_k and π_k and with

$$\pi_k = \frac{\exp(\alpha + \beta x_k)}{1 + \exp(\alpha + \boldsymbol{\beta} x_k)} = \frac{\exp(\beta(x_k - \gamma))}{1 + \exp(\beta(x_k - \gamma))} \tag{6.1}$$

Equivalently, this model is usually expressed as

$$\log\left(\frac{\pi_k}{1 - \pi_k}\right) = \alpha + \beta x_k = \beta(x_k - \gamma) \tag{6.2}$$

Thus, whereas students recognize that the middle term in this expression indicates that this parameterization of the model is a generalized linear model, they also observe that the parameterization on the right in Equation (6.2) corresponds to a generalized nonlinear model. More importantly, since γ is the corresponding LD_{50}, its interpretation is often paramount, and so the latter parameterization is often preferred.

The usual assessment of lack of fit for the binary logistic regression model, discussed for example in Agresti (2002, 2007) and Collett (2003a), involves examining the corresponding standardized residuals for outliers or general lack of fit, and checking for under- or over-dispersion which can arise with the binomial assumption since with fixed sample sizes (the n_k), this Binomial distribution has just one parameter (π_k) for modelling both the mean and variance. Of course, one is also interested in assessing whether the linearity assumption in Equation

(6.2) is suitable; for example, Hastie and Tibshirani (1990, p.282) and Agresti (2007, p.124) give examples in which quadratic models fit their data better than do linear ones.

We underscore here another important manner in which the binary logistic regression model can be mis-specified and demonstrate 'lack of fit'. This arises in situations where the wrong scale is chosen for the independent variable — that is, where for example dose or concentration values may be used but the model may fit better using log-doses, or vice-versa. Instead of fitting the logistic model with several scale choices for the independent variable, we propose here instead that a so-called Box-Cox approach be used — with one very important distinction. Whereas the transformation given in Box and Cox (1964) was applied to the dependent variable, here it is applied to the x-variable in order to choose the appropriate scale.

Specifically, we define the scale transformation function (for $\lambda \neq 0$)

$$x = z(\text{dose}) = \frac{\text{dose}^{\lambda} - 1}{\lambda} \tag{6.3}$$

Since the limit of this expression is the natural log dose, log(dose), when $\lambda \to 0$, we also define $z(\text{dose}) = \log(\text{dose})$ for $\lambda = 0$. In similar manner, we let $\gamma = z(\theta_2)$ in equation (6.2) using this same scale transformation function so that θ_2 is therefore the LD_{50}. This follows since then the right-hand side of Equation (6.2),

$$\beta\{\text{dose}^{\lambda} - \theta_2^{\lambda}\}/\lambda \tag{6.4}$$

equals zero if and only if dose $= \theta_2$, whence $\pi = 1/2$ for this value.

Thus, whereas Box and Cox transformed the dependent variable so as to achieve approximate normality, here we transform the independent variable (dose) so as to permit the logistic curve to better fit the data. Our goal is therefore to use the data to estimate the scale parameter in addition to the original two model parameters ($\boldsymbol{\beta}$ and θ_2). If a set of data indicates that $\lambda = 0$, then the log-dose scale is indicated; the original dose scale should be used for $\lambda = 1$. The estimation method for all three parameters employed here is that of maximum likelihood with an underlying Binomial distribution.

Note that when $\lambda = 0$ (so that the log-dose is indicated), the right-hand side in Equation (6.4) is $\boldsymbol{\beta}\{\log(\text{dose}) - \log(\theta_2)\} = \boldsymbol{\beta}\log(\text{dose}/\theta_2)$, so π_k in equation (6.1) becomes $t_k/(1 + t_k)$ for $t_k = (x_k/\theta_2)^{\theta_3}$ as in the LL2 equation in Section 4 (using θ_3 in place of $\boldsymbol{\beta}$); hence, the name "log-logistic" really applies to the logistic model but using the log-scale for the independent variable. On the other hand, when $\lambda = 1$, then the right hand sides of Equations (6.2) and (6.4) simply become $\theta_3(\text{dose} - \theta_2)$, and the usual logistic model is indicated; this later model function is denoted LOG2 here.

We call this new three-parameter model — with parameters θ_2, θ_3, and λ — the '*scaled binary logistic regression model*', and illustrate its use next.

Example 3. The following data are reported in Collett (2003a, p.6) related to groups of 40 mice infected with a bacterium and injected with one of five doses of a given anti-pneumococcus serum.

Dose	0.0028	0.0056	0.0112	0.0225	0.0450
Number of mice surviving out of 40	5	19	31	34	39
Percent surviving	12.5%	47.5%	77.5%	85.0%	97.5%

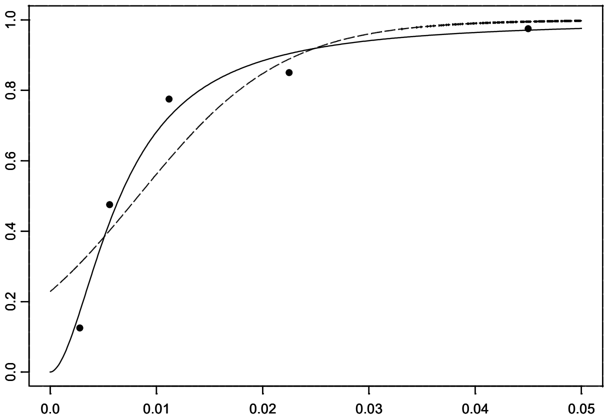

Figure 5: Plot of data (percent survival versus dose), fitted Binary LL2 (solid curve) and Binary LOG2 (dashed curve) model functions for Example 3

Although when the binary logistic regression model in equation (6.2) is fit using the LOGISTIC procedure in SAS no indication of model inadequacy is reported, the Minitab procedure does indicate significant model lack of fit (using each of the Pearson, Deviance and Hosmer-Lemeshow methods). Thus, the practitioner may wonder whether a quadratic term should be added to the model. Furthermore, when these data are examined using the GENMOD procedure in SAS, the calculated deviance statistic is more than five times the corresponding degrees of freedom, leading many students and statistical modelers to suspect over-dispersion. But, it turns out that neither of these diagnoses is appropriate here - rather, these data are best modeled using the log-dose scale instead of the dose scale.

To see this, note that when our three-parameter scaled binary logistic regression model is fit to these data, the scale parameter is estimated to be -0.37, with a confidence interval that includes zero but excludes one. Also, when this model is fit with $\lambda = 0$ — that is, the binary LL2 model fit — the corresponding AIC value is lower than it is for the three-parameter model, thereby reflecting the parsimony and preference of the LL2 model for these data.

For purposes of comparison, these data are plotted in Figure 5 along with the fitted LL2 model function (solid curve) and the fitted LOG2 model function (dashed curve). The superiority of the LL2 model fit for these data is then readily apparent.

As this illustration points out, we advocate that before any other lack of fit diagnostic tests is performed for the binary logistic model, the proper scale be determined for a given dataset using the three-parameter scaled logistic model. This example clearly points out inadequacies associated with logistic regression lack-of-fit tests and diagnostics, and highlights the importance of using our scaled logistic model to first determine the correct scale.

7. Implications for Efficient Experimental Design

As noted in Atkinson and Donev (1992) and O'Brien (1996), so-called optimal designs for models with p model parameters often have only p support points, and are therefore of little usefulness to test for model mis-specification. Thus, whereas these designs may provide maximal information to estimate the model parameters, they provide low or no information for detecting lack of fit. As pointed out above at the end of Section 2, choice of the experimental design is thus paramount when, after the model parameters have been estimated, one wishes to test for goodness-of-fit. This follows since the choice of the design dictates the nature of the departure of the assumed model that can be detected. Practitioners, therefore, are typically interested in choosing near-optimal (so-called robust optimal) designs useful for both estimation and testing for model adequacy.

Example 4. To illustrate, consider the situation in which we believe that the usual simple linear regression model ($y = \beta_0 + \beta_1 x_1 + \epsilon$) is appropriate, and where we wish to choose a design both to estimate the model parameters efficiently and to test for lack of fit of this assumed line; a related illustration is given in Draper and Smith (1998, p.88). Suppose further that due to practical constraints we can only choose as many as $n = 12$ observations, and that the independent variable must lie in the interval between 1 and 12. Arguably the most popular criterion for choosing an optimal design in regression settings is the D-optimality criterion (Silvey, 1980); for this model, the D-optimal design places six replicates at $x = 1$ and six replicates at $x = 12$. This design is denoted Design A in the following table. Although this design gives the maximal information in terms of estimating the linear models parameters (and hence has 100% D-efficiency), it provides absolutely no information for detecting lack of fit, and so is of only limited usefulness in practical settings.

Design	Design support or x-points (number of replicates)	D-Efficiency
A	1 (6); 12 (6)	100.0%
B	1 (4); 6.5 (4); 12 (4)	81.6%
C	1 (3); 4.67 (3); 8.33 (3); 12 (3)	74.5%
D	1 (2); 3.2 (2); 5.4 (2); 7.6 (2); 9.8 (2); 12 (2)	68.3%
E	1; 2; 3; 4; 5; 6; 7; 8; 9; 10; 11; 12	62.8%

On the other end of the spectrum lies Design E, which takes one observation at each of the integers between 1 and 12 inclusive. Since this design has no replicates, it provides no estimate of "pure error." So although it is helpful to check for linearity of the responses, it provides no ability to test for lack of fit in the usual sense (nor can it be used to check the constant variance assumption). Design B has four replicates and three equi-spaced support-points. This design results in an information loss of 18.4% (i.e., subtracting the D-efficiency from its maximal value of 100%), and is useful only to check for quadratic departures of the assumed linear model. By analogy, Design C — which results in only a marginally higher information loss - can be used to test for quadratic and cubic departures.

Hence no one sweeping rule-of-thumb can be provided to help practitioners choose the 'best' design: this choice depends on the degree of faith that the researcher has in the assumed model function and how much information (in the data) the researcher wishes to sacrifice in order to test for lack of fit. It also depends on the nature of the departures from the assumed (linear) model. That said, in many situations, one might be wise to choose either Design B or C in the above situation. Of course, a hybrid or compromise design such as the following is also possible: four replicates at $x = 1$ and $x = 12$ and two replicates at $x = 4.67$ and $x = 8.33$.

Not unexpectedly, the situation is more complicated for nonlinear models and generalized linear models such as the binary logistic model; so-called optimal designs are still of limited usefulness for these models since they usually have only as many support points as model parameters. As a result of this shortcoming, we have proposed two methods to find so-called robust optimal designs — that is, designs that are near optimal but which have extra support points to test for lack of fit. We have found it very beneficial to have interested students learn about these techniques as a part of class projects.

Our first robust design strategy, given in O'Brien *et al.* (2009), provides either uniform or geometric designs. Geometric designs are of the form with support points $x_1 = a, x_2 = a^*b, x_3 = a^*b^2, \ldots, x_{K+1} = a^*b^K$, and are chosen so that the loss in information is not too great — for example, less than 10%. Use of these types of designs is widespread in practical settings. For example, the design used above in Example 4 is a geometric design with $a = 0.0028125, b = 2$, and $K = 4$. Further details on the optimal choice of these parameters for the general setting are given in O'Brien *et al.* (2009); generally, computer algorithms provide optimal choices for 'a' and 'b' here.

The second design procedure that we have proposed is useful for specific departures from the assumed model function; see O'Brien (1996). Here, the assumed model function is embedded or nested into a larger model, called the 'super-model', so that one obtains the assumed model function from the super-model for specific parameter choices. Further, the super-model is chosen so that it has other meaningful sub-models (beyond the assumed model function) as additional special cases. For example, both the three-parameter Weibull (WEIB3) in Equation (5.1) and the three-parameter log-logistic (LL3) given in Equation (5.4) are rival sigmoidal models which asymptote to the line $y = \theta_1$ (as x gets large). Thus a larger four-parameter sigmoidal model is given in O'Brien (1996) that generalizes both of these models functions (has both of these models as special cases) and is used to provide robust designs.

To illustrate with another example, suppose that we feel that the homoskedastic Gaussian two-parameter log-logistic (LL2) model correctly describes a given process, and that we desire a robust optimal design. As noted in Section 4, this model function can be written $\eta = t/(1+t)$ for $t = (x/\theta_2)^{\theta_3}$. This model is equivalent to using the scaled logistic model from the previous section with $\lambda = 0$, but with the important distinction here that the response variable is believed to follow the Gaussian distribution (with constant variance) instead of the Binomial distribution used in the previous section. With the choice $\lambda = 0$ (and the LL2 model function), this means that we feel the model fits using the log-dose scale, but now suppose that we have some doubt about this choice of scale, and that we desire a design useful to confirm this choice in addition to efficiently estimating the model parameters.

In O'Brien *et al.* (2009), it is shown that the D-optimal design for this LL2 model function has two support points and these points are such that $\eta_1 = 0.26$ and $\eta_2 = 0.74$. Indeed, as noted above, this design is therefore problematic since it has only two support points and cannot be used to check for model adequacy. On the other hand, the D-optimal design for the scaled logistic model with faith that the LL2 model is correct (i.e., $\lambda = 0$) has three support points such that $\eta_1 = 0.12, \eta_2 = 1/2$ and $\eta_3 = 0.88$. This latter design is therefore suggested to test for lack of fit, and, more specifically, to test for departures from the assumed LL2 model in the direction of a Gaussian logistic fit using a scale other than the assumed log-dose one.

8. Conclusion

Students and applied statisticians observe that regression modeling represents a somewhat simplified distillation of reality. Nonetheless, it can be a very effective and meaningful tool when one chooses a good model and when good techniques are used to check, validate and/or modify this assumed model. Throughout this paper, we have highlighted important inadequacies with the commonly used checks for model mis-specification, and underscored the need to test for departures from the assumed model function in the direction of meaningful "intermediate" models. In the process, we have challenged students to think critically about statistical modeling. Whereas this is often a relatively easy task for linear models, it involves important subtleties and keen understanding of the nature of various nonlinear and generalized linear equations for more complex models. Also, since choice of the experimental design is paramount in testing for goodness-of-fit, we strongly suggest the use of robust experimental design procedures such as the geometric or model-nesting procedures discussed in Section 7.

Acknowledgements

T. O'Brien gratefully acknowledges the kind hospitality and support of the Department of Statistics, Chiang Mai University, which hosted him during his Fulbright Scholar visit in 2006 -7, as well as the generous financial support the J. William Fulbright Foreign Scholarship Program, Loyola University Chicago, the Thailand-United States Educational Foundation, and the United States Department of State.

Appendix. SAS PROC NLIN output for MM2 and LL3 model fits for modified Example 3 data

```
MM2 Model                    Sum of        Mean               Approx
Source               DF     Squares      Square    F Value    Pr > F
Model                 2     73.0286     36.5143     168.35    <.0001
Error                 6      1.3014      0.2169
Uncorrected Total     8     74.3300

Parameter   Estimate  Approx Std Error  Approx 95% Confidence Limits
```

```
th1          25.5456      35.4462          -61.1882       112.3
th2          24.4245      39.6693          -72.6428       121.5

LL3 Model                  Sum of        Mean               Approx
Source             DF     Squares      Square    F Value    Pr > F
Model               3     73.8326     24.6109     247.40    <.0001
Error               5      0.4974      0.0995
Uncorrected Total   8     74.3300

Parameter  Estimate  Approx Std Error  Approx 95% Confidence Limits
th1        4.4661         0.4633            3.2752      5.6570
th2        2.0600         0.2555            1.4033      2.7167
th3        2.8325         0.8028            0.7688      4.8962
```

References

Agresti, A. (2002). *Categorical Data Analysis*, 2nd edition. Wiley.

Agresti, A. (2007). *An Introduction to Categorical Data Analysis*, 2nd edition. Wiley.

Atkinson, A. C. and Donev, A. N. (1992). *Optimum Experimental Designs.* Clarendon Press.

Bates, D. M. and Watts, D. G. (1988). *Nonlinear Regression Analysis and Its Applications.* Wiley.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In *Robustness in Statistics* (Edited by R. L. Launer and G. N. Wilkinson), 201-236. Academic Press.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society B* **26**, 211-243.

Box, G. E. P. and Lucas, H. L. (1959). Design of experiments in non-linear situations. *Biometrika* **46**, 77-90.

Collett, D. (2003a). *Modelling Binary Data*, 2nd edition. Chapman and Hall.

Collett, D. (2003b). *Modelling Survival Data in Medical Research*, 2nd edition. Chapman and Hall.

Draper, N. R. and Smith, H. (1998). *Applied Regression Analysis*, 3rd edition. Wiley.

Gallant, A. R. (1987). *Nonlinear Statistical Models.* Wiley.

Harrell, F. E. (2001). *Regression Modeling Strategies.* Springer-Verlag.

Hastie, T and Tibshirani, R. (1990). *Generalized Additive Models.* Chapman and Hall.

Huet, S., Bouvier, A., Gruet, M-A. and Jolivet, E. (1996). *Statistical Tools for Nonlinear Regression: A Practical Guide with S-Plus Examples.* Springer.

Krzanowski, W. J. (1998). *An Introduction to Statistical Modelling.* Oxford University Press.

Lindsey, J. K. (1997). *Applying Generalized Linear Models.* Springer-Verlag.

Lindsey, J. K. (2001). *Nonlinear Models in Medical Statistics.* Oxford University Press.

O'Brien, T. E. (1996). Robust design strategies for nonlinear regression models. In *Versuchsplanung in der Industrie* (Edited by H. Toutenberg and R. Gössl). Prentice Hall, 41-52.

O'Brien, T. E., Chooprateep, S. and Homkham, N. (2009). Efficient geometric and uniform design strategies for sigmoidal models. *South African Statist. J.* **43**, to appear.

Ratkowsky, D. A. (1983). *Nonlinear Regression Modeling: A Unified Practical Approach.* Marcel Dekker.

Ratkowsky, D. A. (1990). *Handbook of Nonlinear Regression Models.* Marcel Dekker.

Rawlings, J. O., Pantula, S. G. and Dickey, D. A. (1998). *Applied Regression Analysis: A Research Tool.* Springer-Verlag.

Seber, G. A. F. and Lee, A. J. (2003). *Linear Regression Analysis*, 2nd edition. Wiley.

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

Silvey, S. D. (1980). *Optimal Design.* Chapman and Hall.

Received June 15, 2007; accepted October 9, 2007.

Timothy E. O'Brien
Department of Mathematics and Statistics
Loyola University Chicago
6525 N. Sheridan Road
Chicago, Illinois 60202, USA
tobrie1@luc.edu

Suree Chooprateep
Department of Statistics
Chiang Mai University
Chiang Mai, Thailand
aslsurch@chiangmai.ac.th

Gerald M. Funk
Department of Mathematics and Statistics
Loyola University Chicago
6525 N. Sheridan Road
Chicago, Illinois 60202, USA
gfunk@luc.edu