Pergamon

0097-8485(95)00007-0

Computers Chem. Vol. 19, No. 2, pp. 75–84, 1995

0097-8485/95 $9.50 + 0.00

EVALUATION OF MODEL DISCRIMINATION, PARAMETER ESTIMATION AND GOODNESS OF FIT IN NONLINEAR REGRESSION PROBLEMS BY TEST STATISTICS DISTRIBUTIONS

W. G. BARDSLEY,[1]* N. A. J. BUKHARI,[1] M. W. J. FERGUSON,[1] J. A. CACHAZA[2] and F. J. BURGUILLO[2]

[1]School of Biological Sciences, University of Manchester, Stopford Building, Oxford Road, Manchester M13 3PT, England (email: w.g.bardsley@man.ac.uk)
[2]Universidad de Salamanca, Departamento de Quimica Fisica, Apartado 449, Salamanca, Spain

(Accepted 23 January 1995)

Abstract—There are many programs for fitting nonlinear models to experimental data, and the use of this type of software is now widespread. After fitting a model or sequence of models, these programs usually calculate χ^2, run, sign, F and t statistics as an aid to model discrimination and parameter estimation. The distribution of such statistics from linear regression is well known, but these random variables do not have the stated named distribution after fitting nonlinear models.

First we describe a set of programs that can be used to study the distribution of these well known test statistics from nonlinear regression. Then we present the results from a study of two models that are frequently employed in the life-sciences, and summarize our results from more extensive simulations. Finally, we explain how these programs can be used to create the appropriate cumulative distribution functions, so that exact probability levels can be calculated, given the models of interest, the design points and error structure of a data set.

1. INTRODUCTION

Frequently experiments are performed where there is uncertainty about choosing the correct model from a sequence of possible models, and the F-test for excess variance is widely used to aid such model discrimination (Burguillo *et al.*, 1989). For example, it might be important to identify the correct number of exponentials generating a drug elimination profile (Bardsley *et al.*, 1986), or to pin down the statistically significant number of classes of receptors in a ligand binding experiment (Bardsley & McGinlay, 1987). Then, having selected a model, the χ^2, run and sign tests are often used to estimate goodness of fit, while the t-test is generally relied on to assess parameter redundancy. These statistical tests are of course exact when a linear model is appropriate and the errors are uncorrelated and normally distributed with zero means and known variances (Draper & Smith, 1981). The use of these methods in biochemistry without including correction terms was introduced by Petterson & Pettersson (1970) and Reich (1970), and a useful summary is to be found in the book edited by Endrenyi (1981). However, in the life sciences, the models fitted are seldom linear, the errors are not always uncorrelated or even normally distributed and the variances are never known exactly. Important attempts have been made to correct some of these test statistics for nonlinearity (Beale, 1960; Bates & Watts, 1980; Seber & Wild, 1989). However, in this paper we shall describe software that can be used to examine any given model and error structure in order to make accurate probability statements using the classical test statistics.

The weighting to be used in nonlinear regression is problematical and a number of procedures are routinely used:

- The assumption of constant variance. This assumes that the variances of the observed responses are effectively constant over the range of measurement and independent of the values of the observed responses. That is, weights equal to 1 are employed and the value of the objective function, i.e. the sum of squared residuals at the solution point, is used to form a variance estimate from the whole sample. This is the most widely used approach, but it is unlikely to be appropriate since, in most cases, the variance of measurements is an increasing function of the absolute value of the response.
- The assumption of constant relative error. This assumes that the variances of the observed responses are proportional to the square of the theoretical, i.e. error free responses. So,

* To whom all correspondence should be addressed.

weights derived from the best fit function are used, but this approach leads to serious bias if the wrong model is fitted. A variation is to use weights calculated from the measured response but this leads to uneven weighting with noisy data. In addition, both techniques tend to lead to unreasonably small variance estimates where the measured responses are small.

- Sample replicates are obtained at each design point and these are used to provide estimates of variances. Unfortunately, such estimates are likely to be very unreliable with small numbers of replicates, so sometimes a weighting function is fitted to smooth out the variance estimates and provide a more even weighting.

Clearly a determined investigator should undertake a thorough study of the error structure of the data in order to choose the best weighting scheme. However, in the laboratory, there are often limitations on the number of design points and replicates that can be employed. Also, poor reproducibility as a function of time is often sufficiently severe as to preclude the sort of retrospective repeat measurements recommended from considerations of optimal design, unless the design explicitly allows for such time dependence.

In this paper we shall describe software that can be used to study the relationship between the behaviour of test statistics predicted by the linear model and the same statistics resulting from nonlinear models with known variances, or variances estimated from replicates. This software is illustrated by simulations of some models, numbers of experimental points, number of replicates and types of error chosen to cover typical laboratory experiments.

The statistics predicted by the linear model are a sort of ideal, those generated by correct weights are the closest that could be achieved by nonlinear models, while the difference between known and estimated variances explores how many replicates are required to achieve statistics close to those for known variances.

In order to describe our programs to study the distribution of these widely used test statistics we shall use the following nomenclature:

n = number of distinct design points.
m = number of replicates at each design point,
$N = mn$, the total number of observations,
x_i = design points, $i = 1, 2, \ldots, n$,
ϵ_{ij} = experimental errors,
σ_i^2 = correct variances assuming constant relative error,
s_i^2 = variances estimated from sets of replicates.
$w_i(\sigma) = 1/\sigma_i^2$, the weighting factors using exact variances,
$w_i(s) = 1/s_i^2$, the weighting factors using sample variances,
v = number of parameters to be estimated,
$\Theta = (\theta_1, \theta_2, \ldots, \theta_v)$, the true parameter vector,
$\hat{\Theta}(\sigma)$ = parameters estimated using $w_i(\sigma)$,
$\hat{\Theta}(s)$ = parameters estimated using $w_i(s)$,
$f_k(x, \Theta)$ = an arbitrary model number k
$y_{ij} = f_k(x_i, \Theta) + \epsilon_{ij}$, the measured responses with additive random error,
$r_{ijk}(\sigma) = y_{ij} - f_k[x_i, \hat{\Theta}(\sigma)]$, residuals from fitting model k with weights $w_i(\sigma)$,
$r_{ijk}(s) = y_{ij} - f_k[x_i, \hat{\Theta}(s)]$, residuals from fitting model k with weights $w_i(s)$,
$WSSQ_k(\sigma) = \Sigma_{i=1}^{N} w_i r_{ijk}^2(\sigma)$, the minimized objective function with weights $w_i(\sigma)$,
$WSSQ_k(s) = \Sigma_{i=1}^{N} w_i r_{ijk}^2(s)$, the minimized objective function with weights $w_i(s)$.

This nomenclature will also be extended in obvious ways to include such derived statistics as $F(s)$ and $F(\sigma)$ for F statistics from fitting hierarchies of models with either estimated or known variances, and $t(s)$ and $t(\sigma)$ for the t-statistics used for estimating parameter redundancy.

2. DATA SIMULATION AND FITTING

SIMFIT is a set of computer programs written by one of us (WGB) for the purpose of data simulation, nonlinear regression, statistical analysis and graph plotting in the life sciences. The original version was a main frame suite relying on the NAG library for numerical analysis, but PC versions are now available and can be obtained from the author on request.

First of all we shall describe two programs that can be used to simulate data, namely MAKDAT and ADDERR. Program MAKDAT allows the user to choose a model from a library and then generate data either between fixed end points of the independent variable, or over a range of independent variables determined by chosen values for the dependent variable. Exact data values can then be generated with the independent variable in either an arithmetic or geometric progression, or a special spacing for graph plotting in transformed axes can be selected. Program ADDERR takes in exact data generated by program MAKDAT, then adds errors according to several possible error structures; including generating replicates and adding outliers. In this paper we shall only discuss results obtained with data points forming a geometric progression. This is the optimal design for model discrimination with these types of models (Bardsley *et al.*, 1989), in the sense that the probability of rejecting a deficient model is higher with data points forming a geometric progression than it is when the design points are in an arithmetic progression. Also, we shall only discuss errors

generated according to a normal distribution and constant relative error, with no outliers and with weights calculated either from the sample replicates generated, or else using exact weights.

SIMFIT has numerous dedicated programs to fit specialized models and also contains more advanced general-purpose programs to fit models from a library. However, in this paper we shall only discuss results obtained using the two curve fitting programs MMFIT and EXFIT.

Program MMFIT fits the sequence of sums of Michaelis–Menten models:

$$f_k(x) = \sum_{i=1}^{k} \frac{V_i x}{K_i + x},$$

to a data set supplied. This procedure is trivial for $k = 1$ but requires much more effort for success when fitting higher order cases. To do this, the data are first normalized so that $0 \leqslant x \leqslant 1$ and $0 \leqslant y \leqslant 1$, in internal coordinates in order to stabilize the calculations. Then, the initial intercept and slope are estimated by fitting a line to the early (first three distinct) design points, while the final asymptote is estimated by fitting a Michaelis–Menten model to the final group (last three distinct) design points. The idea is that the sum of V_i will be of the same order of magnitude as the asymptote, V_1, from fitting one function, and the K_i values will also be of the order of the best fit single K_1. On account of the initial scaling, these should then be of order unity. From these estimates, an initial set of possible parameter estimates is calculated, then a random search of possible parameter windows is carried out to refine these estimates and eventually generate a parameter scaling vector. The aim of this procedure is to make the optimization as well conditioned as possible by trying to create a situation where the internal parameters and condition number of the internal Hessian matrix are of the unity at the solution point. Finally a constrained optimization is performed using the quasi-Newton method. After curve fitting, the covariance matrix is estimated, tables of parameter estimates and goodness of fit statistics are produced and graphs of residuals and best fit curves are displayed.

Program EXFIT is a similar program, except that a modified Gauss–Newton method is used for unconstrained fitting of the sequence of sums of exponential functions:

$$f_k(x) = \sum_{i=1}^{k} A_i \exp(-k_i x),$$

to a data set.

Our experience with these, and related programs that have been used over many years to fit biological models, is that fitting models of order $k = 1$ is trivial, fitting models of order $k = 2$ is very difficult and can often give poorly defined parameters, while fitting models of order $k = 3$ or more is highly problematical and seldom justified statistically, except in very exceptional cases with very dense, high quality data covering a large proportion of the total possible response.

Although we have investigated the distribution of test statistics for many models, data spacings and error structures, we shall illustrate the use of these programs in this paper only for the Michaelis–Menten model and exponential models of order $k = 2$. Here, without loss of generality, we only need to consider the special case where any two parameters can be assigned arbitrarily, corresponding to choosing scaling units. So we shall only discuss the parameter set with:

$$v = 4 \text{ and } V_1 = K_1 = A_1 = k_1 = 1,$$

while

$$V_2 = A_2 = K_2 = k_2 = 4.$$

For all the data sets simulated, we chose a range of geometrically spaced independent variable so that $0.1 \leqslant y \leqslant 0.9$, then generated numbers from a normal distribution with mean and variance given by:

$$\mu_i = 0,$$

and

$$\sigma_i^2 = [0.05 f_k(x_i, \Theta)]^2, \quad i = 1, 2, \ldots, n.$$

We shall refer to this case as 5% constant relative error. Then we generated data files with x, y, σ and x, y, s as follows. For each of the n distinct values of x_i we generated m replicates u_j,

$$\epsilon_j \approx N(\mu_i, \sigma_i^2), \quad j = 1, 2, \ldots, m,$$

$$u_j = f_k(x_i, \Theta) + \epsilon_j, \quad j = 1, 2, \ldots, m,$$

$$y_{ij} = u_j,$$

$$\bar{u} = \frac{1}{m} \sum_{j=1}^{m} u_j,$$

$$s_i^2 = \frac{1}{m-1} \sum_{j=1}^{m} (u_j - \bar{u})^2,$$

so that, finally, there were $N = mn$ data values, that is, n distinct values of x and m replicates at each of these n design points. We shall only discuss the case with $n = 8$ and $m = 4$ that is, eight geometrically spaced distinct values of x and four replicates at each fixed x-value, i.e. 32 observations in all. We regard this simulated data set as a high-quality set but one that could be achieved by a determined investigator. For each data set generated with weights calculated from the data and with exact weights, we fitted the models of order $k = 1$, $k = 2$ and $k = 3$ in sequence and recorded all the statistics of interest.

3. THE CHI-SQUARE TEST

After fitting a model it is customary to assess the goodness of fit of the model to the data by a χ^2 test on the sum of weighted squared residuals at

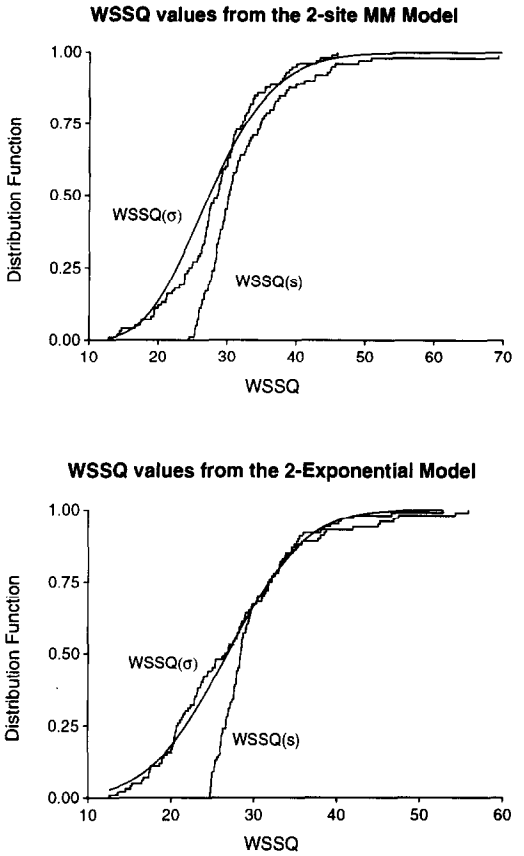

Fig. 1. The distribution of $WSSQ(\sigma)$ and $WSSQ(s)$. An illustration of the program CHISQD. This figure demonstrates the sample cumulative distribution functions for the weighted sum of squared residuals at the solution point and the theoretical χ^2 distribution function from fitting the models of order $k = 2$ as described in the text. The step functions represent 100 simulations, using either the exact standard deviation (σ) or the sample standard deviations (s) from four replicates for weighting. Using exact standard deviations gives objective functions that are indistinguisable from the theoretical χ^2 distribution, but using sample standard deviations results in values skewed to right of the theoretical χ^2 distribution unless more than four replicates are used.

the solution point. To investigate this statistic we used program CHISQD. This program calculates χ^2 PDFs, CDFs and percentage points, as well as doing χ^2 and Fisher exact tests on contingency tables. It also takes in a vector of putative χ^2 random variables, orders them, does an inverse probability transform using the supposed degrees of freedom, then performs a χ^2 and Kolmogorov–Smirnov test on the transforms, and finally plots the cumulative step function on top of the reference CDF curve.

The results from a typical simulation are shown in Fig. 1. For all values of m, the number of replicates at each design point, the values for $WSSQ(\sigma)$ using exact weighting factors were indistinguishable from χ^2 variables, whereas $WSSQ(s)$, using sample estimates for calculating weighting factors, were biased to the right of the reference χ^2 distribution. As m increased from 2 to 16 the ratio defined by:

$$R = \frac{WSSQ(s)}{WSSQ(\sigma)},$$

decreased monotonically towards 1. If the weights were estimated from samples with more than 8 replicates, the distribution of $WSSQ(s)$ could not be differentiated from a χ^2 distribution, but with fewer than 8 replicates at each design point, the $WSSQ(s)$ value was systematically greater than the corresponding $WSSQ(\sigma)$ value.

It is difficult to interpret this result analytically since the parameters estimated using the different weighing schemes were, of course, different, and R is a random variable, so that sometimes it can take a value less than 1. However we can see why such a result can be anticipated when it is remembered that, although s^2 is an unbiased estimator of σ^2, nevertheless:

$$\begin{aligned} P\left(\frac{s^2}{\sigma^2} \leqslant 1\right) &= P\left[\frac{(m-1)s^2}{\sigma^2} \leqslant m-1\right] \\ &= P(\chi^2_{m-1} \leqslant m-1) \\ &> 0.5 \quad , \end{aligned}$$

hence s will more often than not underestimate σ. So residuals weighted by s will tend to exceed residuals weighted by σ and usually $WSSQ(s) > WSSQ(\sigma)$ will occur and be especially noticeable at small values of m. For instance:

$$P(\chi^2_1 < 1) = 0.683,$$

$$P(\chi^2_3 < 1) = 0.608,$$

$$P(\chi^2_7 < 1) = 0.571,$$

$$P(\chi^2_{15} < 1) = 0.549,$$

for the values of $m = 2, 4, 8, 16$ investigated in this study.

In order to obtain an approximate value for the ratio R as a function of the number of replicates m, we can suppose that the parameter estimates using exact weights would not be very different from the parameter estimates using sample variances and investigate the behaviour of s as a function of m. Now $(m-1)s_i^2/\sigma_i^2$ is a χ^2 variable with $m-1$ degrees of freedom and the expectation of its reciprocal is $1/(m-3)$ for $m > 3$ (the expectation does not exist for $m \leqslant 3$). Since we have:

$$E\left(\frac{\sigma_i^2}{s_i^2}\right) = \frac{m-1}{m-3}, \quad m > 3,$$

$$\text{while } R = \frac{WSSQ(s)}{WSSQ(\sigma)},$$

$$\approx \frac{\sigma_i^2}{s_i^2}, \quad \text{if } \hat{\Theta}(\sigma) \approx \hat{\Theta}(s),$$

we might anticipate values for R around $(m-1)/(m-3)$. In fact further analysis of results from simulations like those shown in Fig. 1 show that he values of R encountered are not so extreme as this and actually $9(m-1)/(9m-16)$ is often a much better approximation.

Cumulative distribution functions can be constructed like Fig. 1 for any model and choice of design points, replicates and weighting and then used to estimate probability levels for $WSSQ(s)$ as a goodness of fit statistic. However, in practice such a procedure can be very tedious. We conclude that, if fewer than 4 replicates are used to calculate weighting factors, then $WSSQ(s)$ will be systematically skewed to the right of a χ^2 distribution. If as many as 8 replicates are used then this effect becomes minimal. In any given case, the value of R is, of course, a random variable which will not always exceed unity. However, from extensive studies with these and other models, including inspecting plots of R as a function of m and fitting empirical models, we introduce a suggested procedure for correcting the χ^2 test statistic. When sample standard deviations for weighting are calculated from sets of m replicates at each distinct design point, it is not an unreasonable rule of thumb to multiply $WSSQ(s)$ by a correction factor of $(9m-16)/[9(m-1)]$ before rejecting a fit to data by the χ^2 test.

This conclusion has been reached by storing values of $R(m)$ with $(m=2,\ m=4,\ m=8,\ m=16)$, for various models and designs, and then fitting the above theoretical function $\phi(m)=(m-1)/(m-3)$ and also the empirical function $\psi(m)=(m-1)/(m-16/9)$. The function $\psi(m)$ always fits better than $\phi(m)$, and the usual outcome is that the case $m=2$ shows a large discrepancy, the case $(m=4)$ shows that $R(m)$ is converging asymptotically to unity and, by the time $m=8$, the values of $R(m)$ are indistinguishable from unity. Of course, in any given case, R is a random variable and can be greater or less than one. However, from these and other studies, we suggest that if there are less than four replicates at each design point, it is unwise to use sample variances for weighting. If there are four or five replicates sample variances may be used, and the above correction factor applied with care. With six or more replicates, it is probably safe to use sample variances and dispense with the correction factor.

4. THE F-TEST

In order to determine which model in a sequence of models gives a satisfactory fit with the smallest number of statistically significant parameters, it is customary to fit the possible models, then investigate the relative magnitudes of the objective functions from the model fitting by an F-test, as demonstrated for biochemical models by Burguillo *et al.* (1983), Bardsley *et al.* (1986) and Bardsley & McGinlay (1987).

To investigate the distribution of such test statistics for excess variance, we wrote a program called FTEST. This program calculates PDF, CDF and percentage points for the F distribution, given the numerator and denominator degrees of freedom. Also, it takes a vector of random numbers, performs an inverse probability transform using the supposed F distribution, then uses the Kolmogorov–Smirnov and χ^2 tests to see if the vector of transforms is consistent with a sample from the uniform distribution on (0, 1).

To illustrate the use of this program, Figs 2 and 3 show the results of a typical simulation. Data were simulated for the Michaelis–Menten and exponential models where the correct model had $k=2$, then the data were fitted with the models with $k=1$, $k=2$ then $k=3$ in sequence and the following test statistics

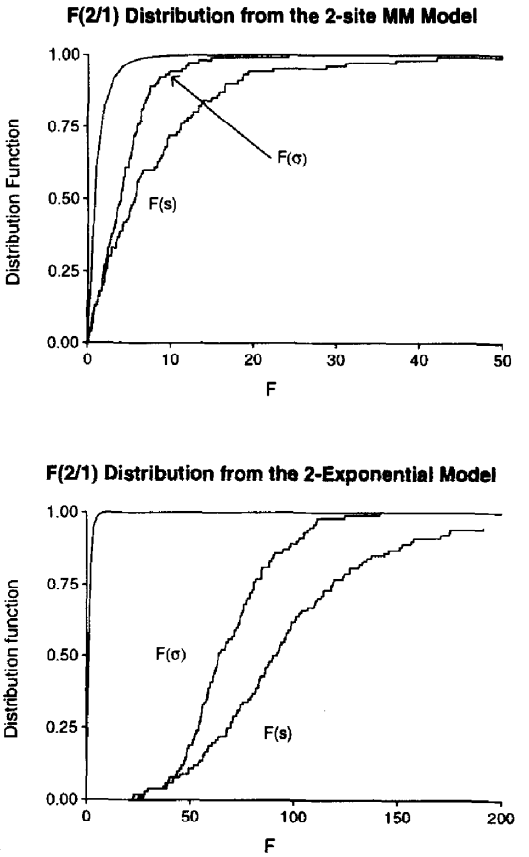

Fig. 2. The distribution of $F(\sigma)$ and $F(s)$. An illustration of the use of program FTEST. This figure demonstrates the sample cumulative distribution functions for the F-statistics and the theoretical F-distribution function, from fitting the models of order $k=1$ then $k=2$ as described in the text. The step functions represent 100 simulations, using either the exact standard deviation (σ) or the sample standard deviations (s) from four replicates for weighting. The test statistics are skewed to the right of the theoretical F distribution, reflecting the substantial improvement in goodness of fit when the correct model is fitted after the deficient model.

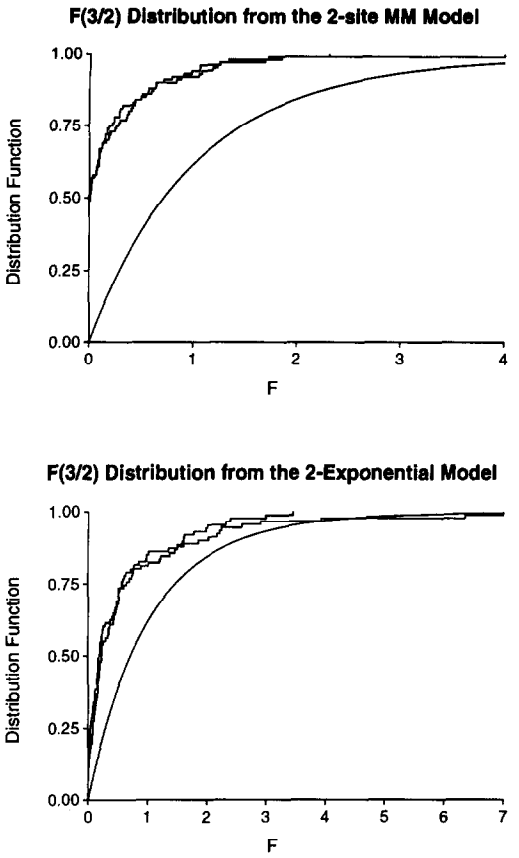

Fig. 3. The distribution of $F(\sigma)$ and $F(s)$. An illustration of the use of program FTEST. This figure shows the sample cumulative distribution functions for the F statistics and theoretical F distribution function, from fitting models of order $k = 2$ then $k = 3$ as described in the text. The step functions represent 100 simulations, using either the exact standard deviation (σ) (step function closest to the smooth curve), or the sample standard deviations (s) from four replicates for weighting. The test statistics were substantially smaller than the theoretical F distribution, reflecting the large number of times that fitting the overdetermined model with redundant parameters gave no appreciable improvement in goodness of fit.

were accumulated:

$$F(2/1)(\sigma) = \frac{[WSSQ_1(\sigma) - WSSQ_2(\sigma)]/2}{WSSQ_2(\sigma)/(N-4)},$$

$$F(2/1)(s) = \frac{[WSSQ_1(s) - WSSQ_2(s)]/2}{WSSQ_2(s)/(N-4)},$$

$$F(3/2)(\sigma) = \frac{[WSSQ_2(\sigma) - WSSQ_3(\sigma)]/2}{WSSQ_3(\sigma)/(N-6)},$$

$$F(3/2)(s) = \frac{[WSSQ_2(s) - WSSQ_3(s)]/2}{WSSQ_3(s)/(N-6)}.$$

The results from this investigation are very clear. From Fig. 2 it is evident that the test statistics were heavily biased to the right of the reference distribution, correctly reflecting the excess variance from fitting the deficient models when models with $k = 1$ were fitted to data generated by order $k = 2$. From Fig. 3 it can be seen that the test statistics were heavily skewed to the left of the appropriate F distribution. This reflects the large number of times when fitting the overdetermined models with $k = 3$ gave no improvement in fit to the correct model with $k = 2$. The conclusion must be that this test is very useful in differentiating models of order 1 from 2 when order 1 or 2 is the correct model, but that differentiating models of order 3 from order 2 is unlikely to be successful. If the F-test indicates that a model with $k = 1$ should be rejected in favour of the corresponding model with $k = 2$ this result should be taken seriously. It should not, however, be expected that this test can be used to differentiate models with $k = 3$ from those with $k = 2$ unless the data are very extensive and accurate.

5. THE RUN AND SIGN TESTS

There are three nonparametric tests that can be applied to residuals from curve fitting which only use the sign of the residuals, not their magnitude. These are the sign test, the run test given the total number of residuals and the run test given the number of positive and negative signs. Before dealing with these tests, we first point out that the null hypothesis is usually that the order of signs is random, in that successive errors will be distributed above and below the true model due to the normal distribution of errors, and that this order will not change too dramatically if the best fit parameters are close to the true ones. Of course correlations induced by the parameter estimation will mean that the sign pattern in the sequence of residuals will not be the same as the sign pattern in the sequence of errors: the residuals will not be normally distributed even though the errors are. A more serious problem arises when replicates are obtained, or when functions of several variables are fitted. This is because the natural order that exists when single design points are used and residuals are ordered according to the increasing magnitude of the independent variable no longer applies. For example, by rearranging the order of measurements within sets of replicates, different run patterns can be produced. We have to suppose that, when replicates exist, the order within the replicates is maintained, say in the sequence in which the replicates were originally measured. The alternative is to do the tests on means only, with concomitant loss of power. With functions of several variables, run tests only make sense if the sequence of residuals is in some logical order, such as the order in time of the original measurements.

If N numbers are uncorrelated and the probability of either sign is 0.5, then the number of positive values $N_\oplus$ and negative values $N_\ominus$ (so that $N = N_\oplus + N_\ominus$) are binomially distributed with parameters N and $p = 0.5$, where p is the binomial probability parameter. Consider now the distribution of runs of

like sign amongst these N signed numbers. There is certainly at least one run, which leaves $(N-1)$ possible positions for the start of the next run. In other words, there remain $(N-1)$ places where the next run can start and, at each place, either sign is equally probable. It follows that, if there are r runs of like signs amongst the N signed numbers, then the random variable $(r-1)$ is binomially distributed with parameters $(N-1)$ and $p=0.5$. So we can easily calculate the conditional probabilities:

$$P(\text{runs} \leqslant r|N) = P(\text{runs} - 1 \leqslant r - 1|N - 1),$$

using the incomplete β-function. However, the distribution of signs, and this particular conditional run probability are not always very useful in practice for analysing goodness of fit. This is because these particular tests have low power, being based only on the binomial distribution. So it is often preferable to use the conditional probability $P(\text{runs} \leqslant r|N_{\oplus} \text{ and } N_{\ominus})$, which involves more searching assumptions, and is the run test usually employed for residuals analysis. Tables are available for this purpose (Swed and Eisenhart, 1943), but we wrote a program called RSTEST to calculate the exact probability levels for the sign and both of the conditional run test statistics. Program RSTEST can be used to analyze any type of run and sign data, but the numerical routines developed for this program are also used by all the SIMFIT curve fitting programs after fitting has been completed, as part of the analysis of residuals. Program RSTEST also performs the runs up and runs down test for randomness, which tests for correlations in large sets of random numbers, as well as the Kolmogorov–Smirnov one and two sample tests and the Mann–Whitney U-test.

Now the sign and run tests just described are not exact, since the residuals will be correlated even though the errors may not be correlated. It is, of course, easy to simulate results for the sign test and the run test conditional upon N. However, it is very difficult to simulate the run test conditional upon $N_{\oplus}$ and $N_{\ominus}$ due to the large number of possible partitions of N into components $N_{\oplus}$ and $N_{\ominus}$. Nevertheless, if it can be shown that the distribution of signs is indistinguishable from a binomial distribution, and the distribution of $(r-1)$ given $(N-1)$ is indistinguishable from a binomial distribution, then it follows that the distribution of runs given $N_{\oplus}$ and $N_{\ominus}$ follow the distribution calculated by Swed and Eisenart (1943), which is a convolution of these binomial distributions.

Program BINOMIAL was written to study such a distribution of runs and signs. This program calculates probabilities, cumulative probabilities, binomial coefficients, percentage points and binomial parameter estimates, with exact confidence limits from the F distribution and with quadratic confidence limits using the normal approximation. It also takes in samples and performs χ^2 tests using selected binomial parameters, or binomial parameters estimated from the samples. Typical results are shown in Fig. 4. We have found that the distribution of signs is generally indistinguishable from a binomial distribution, and the parameter confidence limits usually include 0.5, but the sample variance systematically underestimates the true variance. The distribution of $(r-1)$ was consistently skewed to the right of the theoretical binomial distribution, but could not be differentiated from a binomial distribution with parameters $(N-1)$ and $\hat{p}$, where the binomial parameter was estimated from the sample. Of course, the binomial parameter p was systematically overestimated, especially when the number of replicates was small.

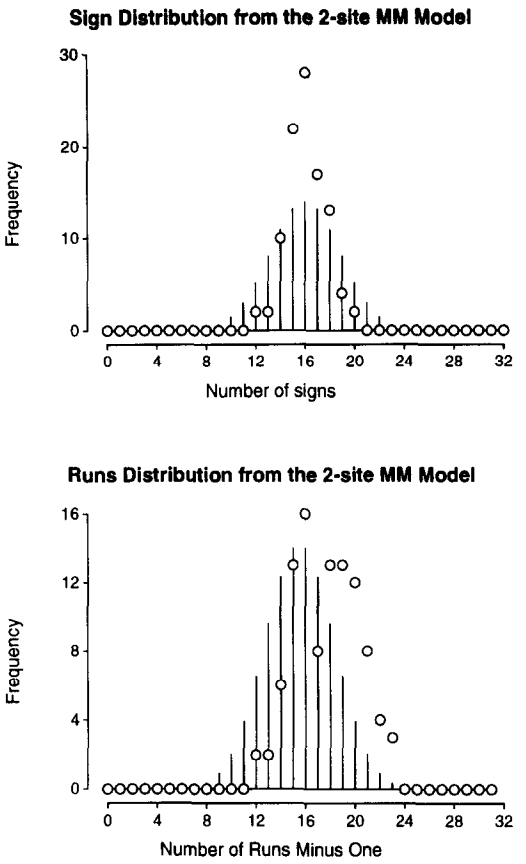

Fig. 4. The distribution of signs and runs. An illustration of the use of program BINOMIAL. This figure demonstrates the sample distribution functions for the sign and run statistics and the theoretical distribution function, from fitting the models of order $k=2$. The histograms compare the results of simulation with the corresponding theoretical binomial distribution. It is clear that the distribution of signs cannot be differentiated from the theoretical distribution. However the number of runs tended to be larger than the theoretical distribution, reflecting the increase in frequency of runs due to correlations resulting from parameter estimation, as discussed in the text.

It is easy to see how correlations induce this type of effect. Suppose that there are just 2 measurements, then the possible error patterns are $++$, $+-$, $-+$, $--$, each with probability 0.25, so that we can have either 1 or 2 runs of like signs, each with probability

0.5. However, after curve fitting, the residuals patterns would be $+-$ or $-+$, each with probability 0.5, leading to precisely 2 runs.

In addition to studying the pattern of runs using program BINOMIAL, we also used program NORMAL, which calculates PDF, CDF and percentage points for the normal distribution, does a Shapiro–Wilks test for a normal distribution, plots normal scores, does an inverse probability transform and tests the transforms for a uniform (0, 1) distribution using the Kolmogorov–Smirnov and χ^2 tests with parameters known or estimated from samples.

From simulations we have concluded that when there are models of order $k = 1$, $k = 2$ or $k = 3$ and only two replicates per design point, i.e. $m = 2$, it is usually possible to detect that the residuals weighted by σ_i or s_i are not normally distributed. However, when there are four or more replicates per design point, the weighted residuals could not be distinguished from a random sample taken from the standard normal distribution, as illustrated in Fig. 5.

Another use for program NORMAL is also illustrated in Fig. 5. According to linear theory, the parameters from a regression should be normally distributed, and this result is supposed to hold asymptotically in the nonlinear case. An example of a rather complicated regression problem where this approximation was found to hold is described by Bardsley *et al.* (1992). The fit of the sample cumulative distribution for the estimated parameters to a normal distribution was poor in general for parameters with estimated weighting and $m = 2$, but was reasonably good with four or more replicates. With noisy data and less well defined parameter spacing, the goodness of fit to a normal distribution rapidly deteriorates.

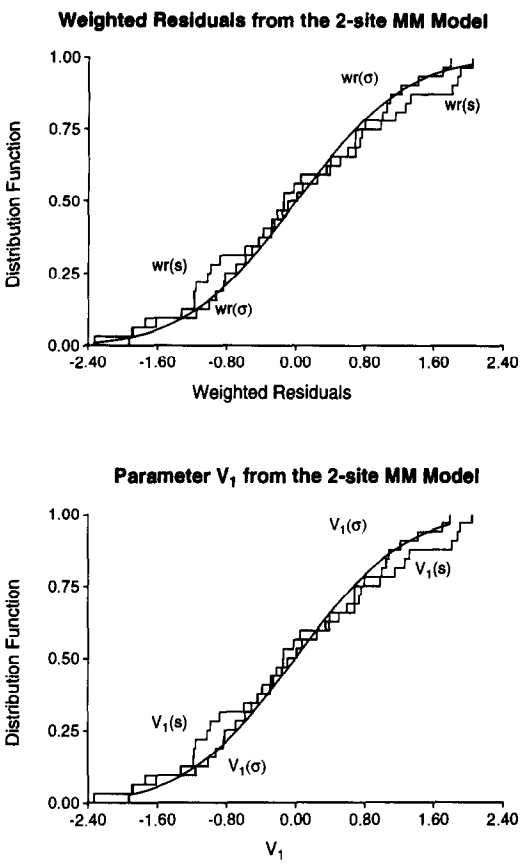

Fig. 5. Weighted residuals and parameter estimates. An illustration of the use of program NORMAL. This figure demonstrates the sample distribution function for a set of weighted residuals using the exact standard deviations (σ) and sample standard deviations based on four replicates (s) for weighting. The theoretical unit normal distribution shown is the theoretical cumulative distribution function. This figure also shows a typical distribution of parameter estimates, in this case the parameter V_1 from fitting a model of order $k = 2$. If the number of replicates falls below four or the number of estimated parameters exceeds four, the fit to a theoretical normal distribution deteriorates.

6. THE t-TEST

When a model has been fitted to some data it is often important to estimate the reliability of the parameter estimates, and the t-test is frequently used as a test for such parameter redundancy. To perform this test, the Jacobian at the solution point is typically used to construct an estimate for the asymptotic variance/covariance matrix, then the square root of the diagonal elements of this matrix are taken as parameter standard errors. For example, if it is wished to test whether a particular parameter estimate is significantly different from a fixed value the following statistic would be calculated:

$$t = \frac{\text{fixed value} - \text{parameter estimate}}{\text{estimated parameter standard error}},$$

and compared to the t distribution with $N - v$ degrees of freedom. Usually, to test whether a parameter is significantly different from zero, the fixed value is set to zero, so that the statistic calculated is simply the absolute value of the parameter estimate divided by the estimated parameter standard error, when a two tailed t-test would be appropriate.

To investigate the distribution of this test statistic we wrote a program called TTEST. This program calculates PDF, CDF and percentage points for the t-distribution, given the number of degrees of freedom, and also does the variance ratio test and paired and unpaired t-tests on samples. It also takes in a vector of random numbers, performs an inverse probability transform on the numbers using the t-distribution with the appropriate degrees of freedom, then tests the transforms for consistency with a uniform distribution on (0, 1) using the Kolmogorov–Smirnov and χ^2 tests.

The results of a typical simulation will be clear from Fig. 6. In general, we have found that when the number of replicates m is at least 4 and the model fitted has only two or three parameters, then the statistics cannot be differentiated from a t distri-

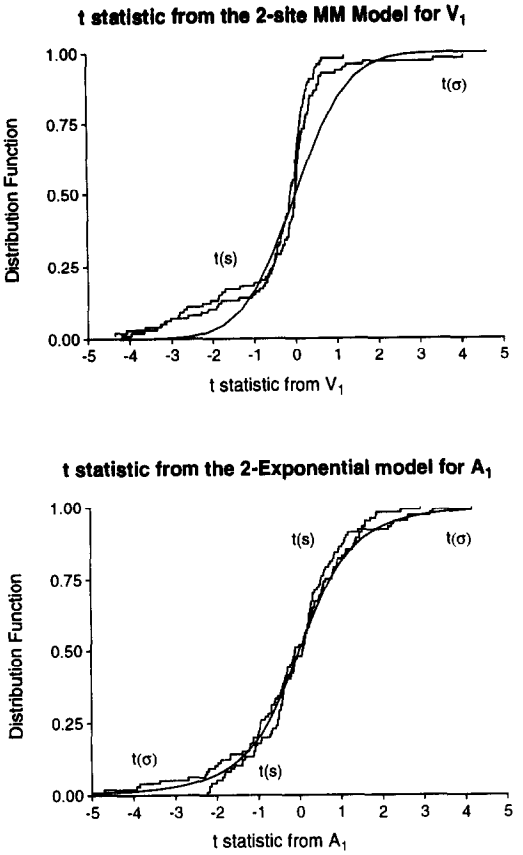

Fig. 6. Distribution of $t(\sigma)$ and $t(s)$. An illustration of the use of program TTEST. This figure demonstrates the sample distribution function for a set of t-statistics using the exact standard deviations (σ) and sample standard deviations based on four replicates (s) for weighting, and the theoretical t distribution. The sample statistics were calculated by subtracting the parameter estimate from the known exact parameter value, then dividing by the estimated parameter standard error.

bution with the appropriate degrees of freedom. However, when models have four or more parameters the distribution of the test statistics has a much greater variance than the t distribution.

7. CONCLUSIONS

We have described the following programs from the SIMFIT package:

- MAKDAT: make data using a library of models
- ADDERR: add errors to exact data to simulate experiments
- MMFIT: fit a sequence of sums of Michaelis–Menten models
- EXFIT: fit a sequence of sums of exponential models
- CHISQD: χ^2 distribution and tests
- RSTEST: run, sign and other nonparametric tests
- BINOMIAL: binomial distribution and tests
- NORMAL: normal distribution and tests
- FTEST: F-distribution and tests
- TTEST: t-distribution and tests

These programs have been in routine use for many years, as mainframe programs for simulation studies and analyzing data, and they are now freely available for PC users. In this paper we have attempted to summarize our experience by presenting an analysis of two well-known tests cases, differentiating two binding sites from one, and confirming the need for two exponential terms. However, we believe that the findings are of more general applicability. The use of these programs in numerous simulation and fitting studies has led us to the following conclusions.

- When models with two or three parameters are fitted, there are few computational problems and the statistical tests for model discrimination and goodness of fit are quite useful. With more than four parameters, serious problems are encountered in curve fitting and also the test statistics become much less reliable.
- If more than four replicates are obtained at each distinct design point and these are used to calculate standard errors for weighting, the resulting test statistics have a distribution that is similar to that when exact weights are used. In other words, if only three or four replicates are available, then it is probably unwise to use these to calculate weights unless smoothing is used. If there are rather more than four replicates at each distinct design point, the difference between using exact weights and weights calculated from the samples will be small, and so there seems little point in smoothing.
- The χ^2 test statistic for goodness of fit will tend to be skewed to the right of a χ^2 distribution if four or fewer replicates are used to calculate weights, and a correction formula has been suggested to compensate for this effect before rejecting a fit by the χ^2 test.
- The F-test for excess variance seems to be very reliable in differentiating models of order $k = 2$ from those with $k = 1$ when $k = 2$ is the correct model. However, there are serious computational difficulties fitting the cases $k > 2$ and the F-test is not likely to be able to differentiate $k = 2$ from $k = 3$ when $k = 3$ is the correct model.
- The sign test and run test given the total number of residuals are not always useful, but the run test conditional upon the number of positive and negative residuals seems to be quite reliable. The complications arising when simulating this test and when there are replicates or functions of several variables have been considered.
- Although the weighted residuals from regression are not actually normally distributed, it will not often be possible to detect such deviations from normality with experimental data.
- The t-test for parameter redundancy behaves much like the F-test for excess variance and

is only useful with low order models ($k \leqslant 2$), where the parameter estimates will tend to be distributed similarly to the appropriate normal distribution.

In addition to their general use in assessing goodness of fit, these programs can be used in sensitivity studies to estimate probability levels for statistics from specific regression problems, using simulation as now described. If the models, number of distinct design points, number of replicates and error structure are assumed, then cumulative distribution functions can be constructed as we have shown in this paper. From these, the approximate probability levels for any given test statistic can then be obtained by reading off the probability level from a sample cumulative distribution function. This can also be done in a more sophisticated and convenient manner, e.g. by spline smoothing and inverse prediction. The SIMFITT package also has a spline smoothing program called CALCURVE which can be used for just such purposes. The data is fitted by adjusting spline knot density until an adequate representation has been achieved, then the spline coefficients can be stored for recall or used directly to calculate the approximate inverse function numerically.

Acknowledgements—We thank E. K. Kyprianou for many helpful discussions during the course of this work. Also, J. A. Cachaza thanks the British Council for a scholarship.

REFERENCES

Bardsley W. G. & McGinlay P. B. (1987) *J. theor. Biol.* **126,** 183–201.

Bardsley W. G., McGinlay P. B. & Wright A. J. (1986) *Biometrika* **73,** 501–508.

Bardsley W. G., McGinlay P. B. & Roig M. G. (1989) *J. theor. Biol.* **139,** 85–102.

Bardsley W. G., Ross Wilson A., Kyprianou E. K. & Melikhova E. K. (1992) *J. Immunol. Meth.* **153,** 235, 247.

Bates D. M. & Watts D. G. (1980) *J. R. Statist. Soc.* **42,** 1–25.

Beale E. M. C. (1960) *J. R. Statist. Soc.* **22,** 41–88.

Burguillo F. J., Wright A. J. & Bardsley W. G. (1989) *Biochem. J.* **211,** 23–34.

Draper N. & Smith H. S. (1981) *Applied Regression Analysis*, 2nd Edn. Wiley, New York.

Endrenyi L. (1981) *Kinetic Data Analysis.* Plenum Press, New York.

Pettersson G. & Pettersson I. (1970) *Acta Chem. Scand.* **24,** 1275–1286.

Reich J. G. (1970) *FEBS Lett.* **9,** 245–251.

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

Swed F. S. & Eisenhart C. (1943) *Ann. Mathl Statist.* **14,** 66–87.