

# F test for lack of fit.

## *Using scatter among replicates to assess the fit of a nonlinear model.*

Harvey Motulsky  
President, GraphPad Software  
[hmotulsky@graphpad.com](mailto:hmotulsky@graphpad.com)  
Jan 2005

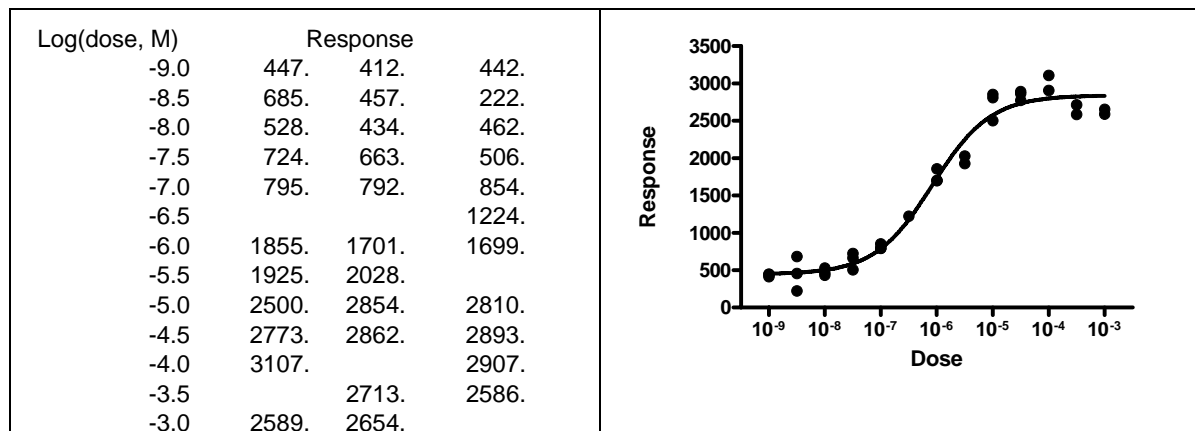
### Introduction

In many experiments, you collect replicate Y values at each X. This article explains how to use the scatter among replicates to assess the scatter of points around a best fit curve in order to determine whether the model adequately explains your data.

This kind of test for lack of fit is often used in linear regression, and discussed in advanced texts of regression (1,2). I have not seen this method used to interpret nonlinear regression results, even though the logic of the tests is identical in linear and nonlinear regression. In the discussion below, I present the math quite differently than do the linear regression texts, but this is simply a rearrangement of the equations with no new concepts.

### Example 1

Below is a dose-response curve, performed in triplicate (with some missing values). The graph shows the fit via nonlinear regression to a sigmoidal (variable-slope) dose response curve.



The response at the last two doses dips down a bit. Is this coincidence? Or evidence of a biphasic response?

One way to approach this problem is to specify an alternative model, and then compare the sums-of-squares of the two fits. In this example, it may not be clear which biphasic model to use as the alternative model. And there probably is no point doing serious investigation of a biphasic model in this example, without first collecting data at higher doses.

Since replicate values were obtained at each dose, the scatter among those replicates lets us assess whether the curve is 'too far' from the points. The calculations are very similar to those used to compare the fit of two models with an F test, and are summarized in this table:

<b>Model</b>	<b>SS</b>	<b>DF</b>
Pure error	274022.2	20
Curve fit	776758.6	29
% increase	183.47%	45.00%
Ratio (F)	4.0770	
P value	0.0043	

The scatter among replicates are pooled into a value called the sum of squares (SS) of pure error. Calculate this by summing the square of the deviation of each replicate from the mean value of all replicates at that value of X. Note that this computation is directly from the replicates, with no need to first fit a model. For this example, the SS of pure error is 274022.2. The number of degrees of freedom for pure error is 20 -- the total number of replicates (33) minus the number of X values (13).

The data were fit to a variable slope dose-response curve using GraphPad Prism, which reported that the SS of the points from the dose-response curve is 776758.6 with 29 df (33 data points minus 4 parameters fit by regression).

It is impossible for any curve fit to have a SS lower than the SS of 'pure error' computed from the replicates. So it is not surprising that the SS from the curve is greater than the SS of pure error. The df let us predict how much greater we expect the sum of squares to be. The df of the curve fit is 45% higher than the df of pure error. If the model fit the data well, the SS of the curve fit ought to also be about 45% higher than the SS of pure error. In fact, the SS of the curve fit is 183% higher than the SS of pure error. The ratio 183/45 equals 4.08. In other words, the scatter of points from the curve is a bit more than four times higher than you'd expect to see if the model fit the data well. How unlikely is it to find such a high ratio? This ratio follows the F distribution (if the model fits the data well), and so can be used to compute a P value. Set  $F = 4.08$ , the numerator df to 9 (df of curve fit minus df of pure error), and the denominator df to 20 (df of pure error). Then determine the P value using Excel (3) or GraphPad's free QuickCalc web calculator (4).

The P value is small (0.0043). This means that if the dose-response model was correct, there is less than half a percent chance that the points would be so much further from the curve than from one another. This strongly suggests that the model is inadequate, and prompts a search for a different model.

### Recasting the calculations using an ANOVA approach

These calculations can also be presented as an ANOVA table. If you frequently deal with ANOVA, you may find this approach easier to follow. This table shows the results of the first example, presented as an ANOVA table.

Source of variation	SS	DF	MS	F	P
Lack of fit	502736.4	9	55859.6	4.08	0.0043
Pure error	274022.2	20	13701.1		
Total (residuals)	776758.6	29			

The bottom row (total SS and DF) has the values reported by the nonlinear regression results, so quantify the scatter of replicates around the best fit curve. In other words, it is the SS of the residuals. The top two lines of the table show how this is split into two components. The 'pure error' values were computed from the replicates, as described earlier, and quantify scatter among replicates. The 'Lack of fit' values were computed by subtraction (Total minus Pure Error), and quantifies the additional deviation of the points from the curve. The F ratio then quantifies the variation attributed to 'lack of fit' compared to that predicted from 'pure error'.

### Relationship to the runs test

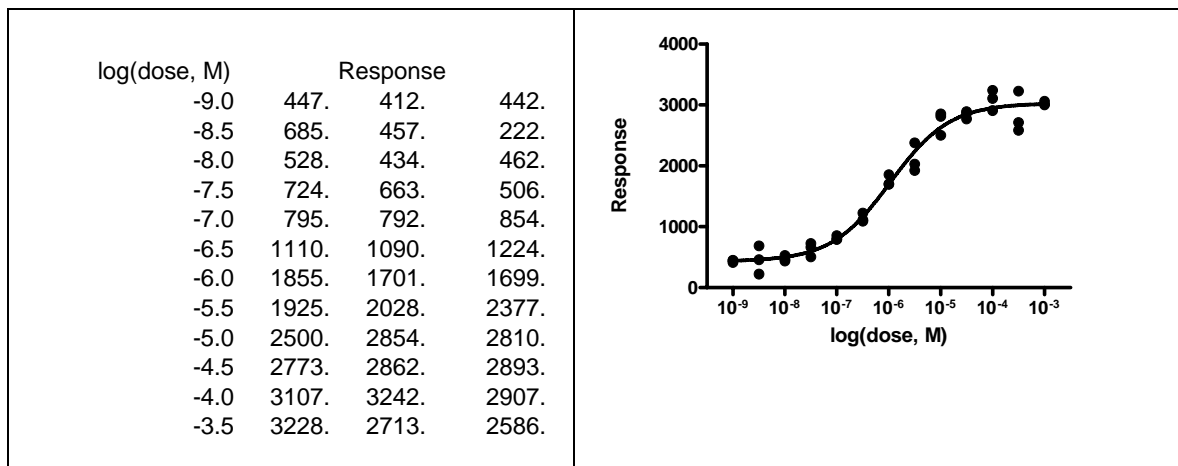
The runs test has the same goal as the F test for lack of fit described here. Both ask if the data deviate significantly from the chosen model, and neither requires an alternative model. A 'run' is a series of sequential points (ordered by X) on the same side of the curve. If there are many fewer runs than expected, points are clustered above and below the curve, which will happen when the model doesn't fit the data very well.

The runs test cannot be used directly with the data of Example 1, because it makes no sense to use individual replicate values with a runs test. But the runs test can be computed using the mean values at each X (Prism does this automatically). The data in Example 1 has data at 13 X values, so the runs test 'sees' 13 points. Of these, are 7 above the curve and 6 are below. With this many points, you expect (on average) 7.5 runs if the curve fits the data well, so the distribution of points above and below the curve is entirely random. In fact, there are only 5 runs. The P value is 0.1212. If the curve fit the data well, so each point was randomly above or below the curve, you'd see 5 or fewer runs in 12% of experiments. This P value is not low enough to conclude that the deviation of the curve from the points is statistically significant.

The runs test only takes into account whether the mean Y value at each X is above or below the curve. The F test for lack of fit takes into account the distance of each individual replicate from the curve. Since the runs test uses so much less information, it is not surprising that it has less power to detect deviations of the curve from the points.

When you have collected independent replicate values, there is little point in using a runs test, as the F test for lack of fit has more power. If you don't have replicate values, the F test for lack of fit cannot be used, and it makes sense to use the runs test.

### Example 2



Model	SS	DF
Pure error	650707.5	26
Curve fit	869738	35
% increase	33.66%	34.62%
Ratio (F)	0.9724	
P value	0.4844	

This example has no missing values, so the df values are larger than in the first example. The df of the curve fit is 35% higher than the df of pure error. So we'd expect the SS of the curve fit to be about 35% higher than the SS of pure error. In fact, the SS of the curve is about 34% higher than the SS of pure error. Accordingly, the F ratio is near 1.0, and the P value is high. The scatter of the points around the curve is entirely consistent with the amount of scatter among the replicates. The model fits the data well, and these data give you no reason to consider an alternative model.

## Limitations

It only makes sense to use scatter among replicates as a way to assess goodness-of-fit to a curve when each replicate is truly independent. In our examples, the replicates would be independent if the dose-response curve data were collected *in vitro*, with triplicate tubes (or wells) at each concentration. In this case, all sources of experimental error are included in every manipulation of each replicate. In contrast, the replicates would not contribute independent information if the dose-response curve by giving each dose to a single animal which is then assessed three times. In this case the triplicate measurements would only quantify scatter among repeated measurements, but not differences between animals, so the SS of 'pure error' would be misleading.

It only makes sense to compare the SS of pure error with SS from regression when the nonlinear regression program treated each replicate as an individual point (this is Prism's default). If the nonlinear regression program fits the mean Y value at each value of X, the SS will be much smaller and cannot be compared in the same way.

These calculations only work when each point is given equal weight in the regression. If the values are given different weights, this F test for lack of fit will not be useful.

In the first example, there was only one value at the seventh dose (with two missing values). This point contributes nothing to the assessment of the pure error goodness of fit, since the distance of that value from the mean at that concentration is zero. But it also contributes nothing to the assessment of the pure error degrees of freedom as it adds one to the count of data points, but also adds one to the count of X values. The number of df equals the number of data points minus the number of X values, so the contribution of that point cancels out. It doesn't help or hurt to leave singlet values in the computation of pure error. They don't help you assess scatter, but also don't mess up the calculations.

## Interpreting the results

If the P value is low, it means that the scatter of the points around the best-fit curve is much greater than predicted by the scatter among the replicates. There are three explanations for a low P value:

- The scatter among replicates doesn't truly assess all aspects of scatter in the experiment, as mentioned above (each dose is a separate animal). You can rule out this possibility if the experiment is well designed.
- The model is correct and the extra scatter is due to coincidence. The P value tells you how rare such a coincidence would be.
- The model is not adequate to fit the data. Consider using a different model, or loosening any constraints you applied.

If the P value is high, it means the scatter of points from the curve is entirely consistent with the scatter among the replicates. There is no reason (based on this test) to question the validity of the model. You can't conclude that you have chosen

the best possible model. But you can conclude that the fit of the model is entirely consistent with expectations from the scatter among replicates within the range of X values you chose to use.

## Summary

Many biological experiments are performed with replicate values at each value of X. When these data are fit using linear or nonlinear regression, the F test for lack of fit compares the scatter of points around the curve with the variation of replicates from one another. If the scatter around the curve is much higher than predicted from the variation among replicates, you have evidence that the model doesn't fully describe your data.

## References

1. *Applied Regression Analysis*, Norman Draper and Harry Smith, Wiley Intersciences, 3rd edition, 1998 page 47-56.
2. *Applied Linear Statistical Models* by Mike Kutner, Chris Nachtsheim, John Neter, William Li, Irwin/McGraw-Hill; 5th edition (September 26, 2004), pages 119-127
3. To compute the P value from the F ratio using Excel, use this formula:  $Fdist(4.08, 9, 20)$ . The first argument is F; the second is the numerator df; the third argument is the denominator df.
4. To compute the P value from the F ratio using GraphPad's free web calculator, go to this URL: <http://www.graphpad.com/quickcalcs/PValue1.cfm>