Environmental Data Laboratory
Professor Kirchner

Laboratory 6: Fitting functions to data

Many data analysis problems are centrally concerned with selecting a function that expresses the relationship between two (or more) variables, and estimating this function's coefficients (also called "parameters") from a data set. In some cases, the algebraic form of the function is dictated by theory. However, in many environmental data analysis problems, there is no solid theoretical argument for preferring one functional form over another. Linear functions and polynomials are the most commonly used general functional forms, perhaps because they are familiar and because they are readily available in many software packages:

linear	$y = a + b\,x$
quadratic	$y = a + b\,x + c\,x^2$
cubic	$y = a + b\,x + c\,x^2 + d\,x^3$
quartic	$y = a + b\,x + c\,x^2 + d\,x^3 + e\,x^4$

where a, b, c, d, and e are constants fitted by regression. If you aren't familiar with the computational details of regression, don't worry. For the moment it is sufficient for you to know that in regression, the constants (a, b, c, and so forth) are adjusted until the function matches the data as closely as possible, where "as closely as possible" has a precise meaning: the sum of squared deviations between the predicted and observed values of y (given the observed values of x) is as small as possible.

Because the software takes care of the tedious details of estimating the coefficients, the only practical problem remaining for the analyst is to choose which functional forms to use. Analysts often choose whatever functions "look right" to them, or choose the functions that visually "fit the data best". In other cases, analysts gauge how well the function (or "model") fits the data using r^2, which is the fraction of the variance in the data that is accounted for or explained by the model. If the model fits the observed data exactly, $r^2=1$. On the other hand, if the model does not capture any of the variation in the data (and "models" the data simply as its mean), then $r^2=0$. Because r^2 is a quantitative, unambiguous measure of goodness-of-fit, it is widely used to gauge the performance of alternative models. However, for several reasons it is dangerous to use r^2 to decide which model is "better". Nonetheless, it is widely assumed that models that give a closer fit to the data will be more reliable predictors of the future. Here you will test this assumption using small, simple data sets. While the "models" at hand are the simple polynomials shown above, the general principles illustrated here apply equally well when you are fitting more complex models, including computer simulation models, to data.

As before, anything requiring a response is indicated by "•".

Venice sea-level data

Using JMP, open the data file, "Venice sea level.jmp". This file contains a 51-year record of the maximum sea levels recorded each year in Venice, Italy.

• (1) Plot sea level as a function of time (using "fit y by x") for the full 51-year record, and briefly describe whatever relationships you see in the data.

• (2) Using the "Fit line" option in the "Bivariate fit" pull-down triangle, fit the data with a linear function. Record the estimated annual trend (the slope of the relationship between sea level and year, shown in JMP as the "parameter estimate" for the variable "year"), and the regression coefficient r^2. Compare the linear fit to the data. Are there any obvious discrepancies?

When fitting a curve to data, it is often useful to look at the *residuals*, that is, the differences between the predicted and observed values at each point. Since this has the effect of subtracting out whatever aspects of the data that are accounted for by the model, any remaining systematic discrepancies between the model and the data will often be clearer in the residual plots than in an ordinary x-y plot. For example, if the model is linear but the data are nonlinear, the curvature will probably be clear in a residual plot. If there is nothing left in the residuals but patternless random noise, then the model is the best that can be hoped for; there isn't any more systematic "signal" left to be modeled in the data (and trying to fit the remaining "noise" will only lead you astray). Residuals are usually plotted either as a function of x, or as a function of the predicted y values.

Using the "Linear Fit" triangle (the one right at the bottom of the graph), save the predicteds (the predicted values of y for each value of x), and the residuals (the differences between the predicted and measured values of y). These options are near the middle of the scroll list that appears when you click and hold on the "linear fit" triangle.

• (3) Plot the residuals as a function of year, and as a function of the predicted y's. Is there any apparent unexplained systematic pattern left in the residuals?

• (4) Using the "Bivariate fit" triangle on your original plot from step 1, fit the data with linear, quadratic, cubic, and quartic functions (you can overlay all four on the same plot). Are there visibly obvious differences among the fitted curves for these various functions? Why would you expect the fitted quadratic, cubic, and quartic functions to be almost exactly linear, given what you observed in the residual plots?

You rarely find a time series as long or as well-behaved as this one. Instead, you often face the problem of trying to infer trends, and changes in trends, from much shorter time series where the variable of interest is a less well defined function of time. To explore some of the interesting problems that can arise when analyzing short time series, let's look at just the first decade of the Venice sea level data. Using the exclude command (Rows >> Exclude/Unexclude), exclude all the rows except the first ten. Now, plot sea level as a function of time again; you should see only the first ten points. Remember, in the real world these first ten years might well be all the data you have. In the real world, you wouldn't have the benefit of peeking at the following 40-some years of data to discover how things turned out over the long haul.

• (5) Using the "Bivariate Fit" triangle, fit the first ten years of data with linear, quadratic, cubic, and quartic functions. Record the r^2 for each of the fits. Which functions seem, by eye, to best match the behavior of the data? Why do some functions appear to fit the data much better than others in the 10-year data set, whereas the differences between the various functions are much smaller in the full 51-year data set?

• (6) Keep the plots generated in (5) available. Using the exclude command, remove the exclusion on rows 11 through 20, so that now you can fit to the first 20 years of data. Repeat the analysis in step (5), answering the same questions for your new 20-year fits.

Now, how well would you have done if you had tried to predict the future on the basis of your 10 or 20 years of data? What you need to do is to superimpose your 10-year and 20-year function fits on the full 51-year data set. The clever way to do this is the following: you remove any exclusions of rows (using "clear row states" under the "rows" menu), then you "fit y by x" where x is year, y is sea level, and weight is either the "weight (10 years)" column, or the "weight (20 years)" column. These columns tell JMP how much to weight to put on each of the individual data points in the regressions. Because all the points after 10 or 20 years have a weight of zero, they won't influence the regression calculations, although they will be plotted.

• (7) Plot the full 51 years of data, selecting the "weight 10 years" column for "weight" in the "fit y by x" menu. Then, as before, use the "Bivariate Fit" triangle to fit linear, quadratic, cubic, and quartic functions to the data. Note that the curves are exactly the same as you got in step (6), except this time the scales of the axes are different (to fit the rest of the data). Keep this window up, and repeat the same analysis in a new window with the first 20 years (using the "weight 20 years" column). Do any of the functions estimated from 10 or 20 years of data give plausible values for sea level over the whole period of record? Which ones? Do any of the functions predict implausibly high or low sea level values, shortly after the end of the 10 or 20 year time span to which they were fitted? Which ones?

• (8) Does "goodness of fit", as measured either by r^2 or by the visual match between the function and the data, make a reliable basis for judging whether one function or another is likely to be a better predictor of future behavior? That is, did the curves that fit the 10-year and 20-year data "better"--as measured by r^2--work better, or worse, at predicting the future? How can some functions fit the short-term data so well, but fit the long-term data so poorly? Can you think of other criteria that one might want to use in selecting functions to fit data?

Monocacy River rainfall/runoff data

Open the "Monocacy storms.jmp" data file. This file contains data for 25 storms that hit the watershed of the Monocacy River in Maryland. The two pieces of data are the total rainfall in each storm (measured in units of depth), and the total runoff from each storm (that is, the sum of runoff from the time the storm began until the stream returned to base flow after the storm ended). The runoff data have been converted to units of depth (by dividing by the watershed area), so that runoff can be directly compared to the rainfall figures. You can visualize this process as collecting all the runoff in a huge tank, then spreading it like peanut butter in an even layer over the watershed. Rainfall/runoff relationships can be interpreted at various levels of hydrological sophistication. For this example, it is sufficient for you to know that water falling as rain on the watershed can do at least three different things: it can evaporate, it can run off relatively quickly (as storm runoff), or it can recharge long-term groundwater storage, leaving the watershed as base flow over a long period.

•(9) Plot runoff as a function of precipitation, and fit the data with a linear relationship. Write a brief description of what this relationship suggests about what fraction of rainfall leaves the watershed as storm flow, for different sizes of storms. In particular, what does this relationship suggest will happen for very small storms, as precipitation approaches zero? (You will want to reset the x-axis range so that the origin is 0,0 to make this clear. Double click on the x-axis to do this.)

•(10) Fit the same data with a quadratic relationship. What does this relationship suggest will happen to runoff during small storms (as precipitation approaches zero)? Is there any clear reason to prefer either a quadratic or a linear fit to the available data? If not, is there a way to choose among the conflicting interpretations that are suggested by these two different functional relationships?