MATLAB for SMS

Lesson 4: Curve Fitting

Concepts

Background

When I use the phrase, "curve fitting," I mean to say something more along the lines of "regression analysis by the method of least squares."

The method of least squares (LS) was originally formulated around 1795 by Carl Friedrich Gauss. So often as the story goes with important scientific developments, A.M. Legendre and Gauss, as well as the oft-ignored American Robert Adrain, formulated the same method independently at around the same time. Legendre was the first to publish the method in 1805, but presented little proof or support of his method. Adrain published his formulation, which is apparently very similar to Gauss's, in 1808 (but he was in America, so no one noticed. And it's possible that he had Legendre's book on hand anyway). Gauss's publication followed in 1809, and was accompanied by a thorough formulation. Ultimately, the detail of the formulation fueled Gauss's claim that he had first used the method in 1795, and he is now widely accepted as the inventor of LS. If you like, see Stigler, Stephen M. *Gauss and the Invention of Least Squares*. The Annals of Statistics. 9(3):465-474 (1981) for a decent historical account and a fairly ingenious argument, albeit nearly 200 years after the fact, that Gauss did, in fact, invent LS in 1795.

Also interesting fare is the fact that, until very recently (2009), we thought Adrien-Marie Legendre was this guy:

That guy was just some French politician. Rather unfortunately for Legendre, this caricature is now the only known portrait of him (this isn't *just* a joke):

Be aware that the terminology associated with LS varies occasionally. Words carrying both specificity and importance are sometimes used differently across different parts of the literature (not to mention the internet…). Most common is the confusing substitution of *error* for *residual*. Here, and *most* everywhere else, *residual* will signify deviation of the LS regression to the observable data, and *error* will refer to, in its various forms, the error (or uncertainty, or noise) associated with the observable data itself.

The Basic Premise

The method of least squares seeks to optimize a parametric model function with respect to some set of observable data. The model function represents some hypothesis about some observable data set, and the LS regression is a test of said hypothesis. The parameters of the model function are independent variables that are adjusted to conform, if possible, to the characteristics of the particular observable.

From a more practical perspective, we have some discrete, 2-D data set $\{\{d_i\},\{x_i\}; i = 1,2,\ldots\}$, where the d_i are the observables, and are a function of the independent variables x_i. We then make some conjecture about our observable, e.g., that the model function m(x,β), where β contains some number of adjustable parameters (such as a mean, a variance, etc.), is an accurate description of the characteristics of our observable. We test our conjecture with the LS regression.

Linear and Nonlinear Model Functions

Suppose our data may be described as a straight line. In this case we specify:

$$m(x,\beta) \;=\; \beta_0 \;+\; \beta_1 x$$

This is a linear model. It is linear not because it is linear in x, but is linear because it is linear in BETA. This will become clearer if we consider another linear model. Suppose our data is described by a quadratic polynomial. We specify:

$$m(x,\beta) \;=\; \beta_0 \;+\; \beta_1 x \;+\; \beta_2 x^2$$

This is also a linear model function because *m(x,β)* is linear in all elements of β. LS regression with a linear model function is appropriately called linear LS regression.

Most often however, a linear model will be inapplicable to the types of data we encounter in SMS. In this case we use a nonlinear LS regression. An example of a nonlinear model function is an exponential function:

$$m(x,\beta) \;=\; \beta_0 e^{\beta_1 x}$$

This particular model can be linearized, but be aware that linearization may lead to poor results. Specifically, for the exponential above, taking the natural logarithm of both sides leads to a linear model, but the error distribution associated with the observable d_i must be lognormally distributed, which will not usually be the situation (see Central Limit Theorem). Linearization in the case of normally distributed error will lead to a different (read: less accurate) LS regression than had the nonlinear model been retained.

The Sum of Squared Residuals

A LS regression's objective is to find the values of the parameters of the model function that minimize the sum of squared residuals (often called the sum of squared error). The sum of squared residuals is expressed as:

$$S = \sum_i (d_i - m(x_i, \beta))^2$$

This quantity represents what is often called the unexplained variance. This quantity is "unexplained" because it represents variation in the data that isn't described by the model. In conjunction with the sum of total squared error S_{tot} (notice this quantity is proportional to the variance)

$$S_{tot} = \sum_i (d_i - \langle d \rangle)^2,$$

we obtain the coefficient of determination:

$$R^2 = 1 - \frac{S}{S_{tot}}$$

Be conservative with your confidence in R^2. It's a biased measure that trends towards 1 as the number of samples (data points) increases and as the overall variance in the observable data increases, regardless of the quality of the fit. I only mention R^2 here to show how to calculate it and to warn you against believing it has any real or quantitative significance.

Weighted LS

It is important to note that we are working under the assumption that errors in the observable are uncorrelated. In other words, the error is a *random noise*. We'll get a better definition, as well as more than our fill of random noise in the coming weeks. If there are correlated errors in the observable, the entire variance-covariance matrix must be estimated and used to weight each contributor to each residual. We'll skip this entirely, as it is even less manageable than it sounds.

In the special case that error is uncorrelated but error (variance) varies (…) across the observables in a single data set, weighted LS should be used. An example of such behavior would be auto- or cross-correlation curves, in which each observable – i.e., each single correlation amplitude at each single lag time – contains contributions from a number of contributors, and this number may be different from any other observable in the data set. Considering that the variance within each observable is proportional to the number of contributors, long lag times will usually have larger variances than do short lag times. The weighted sum of squared residuals is:

$$S_{weighted} = \sum_i W_{ii}(d_i - m(x_i, \beta))^2$$

$$W_{ii} = \frac{1}{\sigma_{ii}^2}$$

Here, each W_{ii} is inversely proportional to the variance of the corresponding element along the diagonal of the covariance matrix of the observable data. Note that we are not performing weighted fits to correlation curves to reduce the contributions of the longer lag times to the fit. Rather, we are weighting the fit to *correct* the squared residual at each lag time such that its contribution to the sum S more accurately represents the (this is important…) *amount of unexplained variance within the observable data point*. More details on correlations next week.

Miscellany

Confidence intervals are usually calculated with the assumption that the distribution of error in the observable data set is a normally distributed zero mean process. The error in the observables is applied to the model parameters as well, resulting in confidence bounds. More concisely, if the standard deviation of the error (noise) distribution is σ, then for some set of model parameters β, the 68% bound is $\beta \pm \sigma$, and the 95% bound is $\beta \pm 2\sigma$.

Constraints are often used to restrict the range of possible values the model parameters can take. Constraints should only be used if the unconstrained fit becomes problematic, or if there is some relationship between two or more parameters that must be maintained. Constraints should not be used to force a model parameter to take on a value in the range that YOU think it should have. This would be the definition of subjectivity. If the range of a parameter must be restricted, choose the constraints so that the domain is as large as possible, thereby allowing for the highest level of objectivity.

These comments on objectivity raise one final point. Any observable will fit to an inappropriate model function if the model contains enough parameters. e.g., many distributions, regardless of what kind of data they contain, will fit to the sum 3 or 4 or 5 Gaussians. This does not necessarily mean that the observable distribution actually contains 5 Gaussian distributions. It means that 15 parameters can be manipulated to yield a satisfactory result. Take care to select the appropriate model. For example, spectral bands are not usually Gaussian but Lorentzian. Making distinctions such as these will result in better and more reliable fits.

A parting digression

The assumption that error is normally distributed without regard to the distribution of the observable data is otherwise known as the Central Limit Theorem. We will be looking at some substantial implications of this theorem very soon. Feel free to (re)acquaint yourself with it. Said Sir Francis Galton (the father of *standard deviation*…) of the central limit theorem:

"I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the "Law of Frequency of Error". The law would have been personified by the Greeks and deified if they had known of it. It reigns with serenity and in complete self-effacement, amidst the wildest confusion. The huger the mob, and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshaled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along."