Notes on Regression Analysis

Physics 4L - David Kleinfeld

We often record data that we believe may be described by a certain functional relation. How do we fit that function? Consider the case of N data points which consist of a pair of variable, (x_i, y_i), where the index i lables th data point. The set of values $[x_i|i = 1, n]$ are the independent variables, i.e., the things we change in the lab, while the set of values $[y_i|i = 1, n]$ are the dependent variables, i.e., the things we measure in the lab.

Suppose that we have a model $f(x)$ that describes the points. We can fit the model to the data according to a variety of minimization criteria. One, that is useful when the measured values $[y_i|i = 1, n]$ have gaussian distributed additive noise about their "true" value, is to minimize against the square of the differences between the measured and model values. A popular alternate criteria, that we will not explore, is to minimize the largest error.

We seek to minimize an error that is defined by the square of the distance, i.e., effectively the magnitude of the distance, between the model and measured points. The error is then defined in terms of the variance, σ_y^2, between the model and the data, i.e.,:

$$Error \equiv \sigma_y^2 = \frac{1}{N-M} \sum_{i=1}^{N} [f(x_i) - y_i]^2 \tag{1}$$

where M is the number of free parameters in the model.

We now need a specific model for $f(x)$. For the sake of simplicity, we take a straight line that intercepts the axes at $(0, 0)$. This has a single free parameter, A, i.e.,

$$f(x) = Ax \tag{2}$$

so that $M = 1$. The variance is thus evaluated as:

$$\begin{aligned} \sigma_y^2 &= \frac{1}{N-1} \sum_{i=1}^{N} (Ax_i - y_i)^2 \qquad (3) \\ &= \frac{1}{N-1} \sum_{i=1}^{N} \left(Ax_i^2 - 2Ax_i y_i + y_i^2 \right) \\ &= \frac{A^2}{N-1} \sum_{i=1}^{N} x_i^2 - \frac{2A}{N-1} \sum_{i=1}^{N} x_i y_i + \frac{1}{N-1} \sum_{i=1}^{N} y_i^2 \end{aligned}$$

The error is in the form of a quadratic equation in the parameter A. The sums are just terms that can be calculated directly from the data. We can find the optimal value of the parameter A by minimizing the variance with respect to A, i.e., when $\frac{\partial \sigma_y^2}{\partial A} = 0$. We find

$$\frac{\partial \sigma_y^2}{\partial A} = \frac{2A}{N-1}\sum_{i=1}^{N} x_i^2 - \frac{2}{N-1}\sum_{i=1}^{N} x_i y_i \tag{4}$$

so that

$$A = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2} \tag{5}$$

To insure that this is really the minimum, we check if $\frac{\partial^2 \sigma_y^2}{\partial A^2} > 0$. We find

$$\frac{\partial^2 \sigma_y^2}{\partial A^2} = \frac{2}{N-1}\sum_{i=1}^{N} x_i^2 \tag{6}$$

which consists of solely positive terms. So we have indeed found the relation for A that insures that the variance between the measured and fitted values is a minimum.

The parameter A is useful only if we also know the variance of A, denoted σ_A^2. What is this? We use the "sum-of-squares" rule to write:

$$\begin{aligned} \sigma_A^2 &= \left(\frac{\partial A}{\partial y_1}\right)^2 \sigma_{y_1}^2 + \left(\frac{\partial A}{\partial y_2}\right)^2 \sigma_{y_2}^2 + \cdots \\ &= \sum_{j=1}^{N} \left(\frac{\partial A}{\partial y_j}\right)^2 \sigma_{y_i}^2 \end{aligned} \tag{7}$$

We can simplify this expression by noting that all of the $\sigma_{y_i}^2$ may be taken as equal, so that

$$\sigma_A^2 = \sigma_y^2 \sum_{j=1}^{N} \left(\frac{\partial A}{\partial y_j}\right)^2 \tag{8}$$

The two terms in the expression for σ_A^2 can be readily evaluated using our expressions for σ_y^2 and A. We have

$$\sigma_y^2 = \frac{1}{N-1}\left[A^2 \sum_{i=1}^{N} x_i^2 - 2A \sum_{i=1}^{N} x_i y_i + \sum_{i=1}^{N} y_i^2\right] \tag{9}$$

$$
\begin{aligned}
&= \frac{1}{N-1}\left[\frac{\left(\sum_{i=1}^N x_i y_i\right)^2}{\left(\sum_{i=1}^N x_i^2\right)^2}\sum_{i=1}^N x_i^2 - 2\frac{\sum_{i=1}^N x_i y_i}{\sum_{i=1}^N x_i^2}\sum_{i=1}^N x_i y_i + \sum_{i=1}^N y_i^2\right] \\
&= \frac{1}{N-1}\left[\frac{\left(\sum_{i=1}^N x_i y_i\right)^2}{\sum_{i=1}^N x_i^2} - 2\frac{\left(\sum_{i=1}^N x_i y_i\right)^2}{\sum_{i=1}^N x_i^2} + \frac{\left(\sum_{i=1}^N y_i^2\right)\left(\sum_{i=1}^N x_i^2\right)}{\sum_{i=1}^N x_i^2}\right] \\
&= \frac{1}{N-1}\left[\frac{\left(\sum_{i=1}^N y_i^2\right)\left(\sum_{i=1}^N x_i^2\right) - \left(\sum_{i=1}^N x_i y_i\right)^2}{\sum_{i=1}^N x_i^2}\right]
\end{aligned}
$$

and

$$
\begin{aligned}
\sum_{j=1}^N \left(\frac{\partial A}{\partial y_j}\right)^2 &= \sum_{j=1}^N \left[\frac{\partial}{\partial y_j}\left(\frac{\sum_{i=1}^N x_i y_i}{\sum_{i=1}^N x_i^2}\right)\right]^2 \\
&= \sum_{j=1}^N \left[\frac{1}{\sum_{i=1}^N x_i^2}\sum_{i=1}^N \frac{\partial\left(x_i y_i\right)}{\partial y_j}\right]^2 \\
&= \sum_{j=1}^N \left(\frac{1}{\sum_{i=1}^N x_i^2} x_j\right)^2 \\
&= \frac{1}{\left(\sum_{i=1}^N x_i^2\right)^2}\sum_{j=1}^N x_j^2 \\
&= \frac{1}{\sum_{i=1}^N x_i^2}
\end{aligned}
\tag{10}
$$

Thus

$$
\begin{aligned}
\sigma_A^2 &= \frac{1}{N-1}\left[\frac{\left(\sum_{i=1}^N y_i^2\right)\left(\sum_{i=1}^N x_i^2\right) - \left(\sum_{i=1}^N x_i y_i\right)^2}{\sum_{i=1}^N x_i^2}\right]\frac{1}{\sum_{i=1}^N x_i^2} \\
&= \frac{1}{N-1}\left[\frac{\left(\sum_{i=1}^N y_i^2\right)\left(\sum_{i=1}^N x_i^2\right) - \left(\sum_{i=1}^N x_i y_i\right)^2}{\left(\sum_{i=1}^N x_i^2\right)^2}\right]
\end{aligned}
\tag{11}
$$

so that the standard deviation in A is

$$
\sigma_A = \sqrt{\frac{1}{N-1}\left[\frac{\left(\sum_{i=1}^N y_i^2\right)\left(\sum_{i=1}^N x_i^2\right) - \left(\sum_{i=1}^N x_i y_i\right)^2}{\left(\sum_{i=1}^N x_i^2\right)^2}\right]} \tag{12}
$$

$$= \sqrt{\frac{1}{N-1}} \frac{\sqrt{\left(\sum_{i=1}^{N} y_i^2\right)\left(\sum_{i=1}^{N} x_i^2\right) - \left(\sum_{i=1}^{N} x_i y_i\right)^2}}{\sum_{i=1}^{N} x_i^2}$$

We now have expressions for both A and σ_A.

A final piece of business is to write an expression for the fractional error in A, so that we can gain insight into the nature of error reduction by averaging over the N data points in the fit. We have

$$\frac{\sigma_A}{A} = \sqrt{\frac{1}{N-1}} \frac{\sqrt{\left(\sum_{i=1}^{N} y_i^2\right)\left(\sum_{i=1}^{N} x_i^2\right) - \left(\sum_{i=1}^{N} x_i y_i\right)^2}}{\sum_{i=1}^{N} x_i^2} \cdot \frac{\sum_{i=1}^{N} x_i^2}{\sum_{i=1}^{N} x_i y_i} \quad (13)$$

$$= \sqrt{\frac{1}{N-1}} \sqrt{\frac{\left(\sum_{i=1}^{N} y_i^2\right)\left(\sum_{i=1}^{N} x_i^2\right)}{\left(\sum_{i=1}^{N} x_i y_i\right)^2} - 1}$$

The important result is that the fractional error decreases essentially in proportion to $N^{-1/2}$, just like the decrease in standard deviation one finds from averaging.

We have chosen to demonstrate the procedure of model fitting with a linear fit. In general, models may contain a multitude of parameters, but the basic concept of writing an error and minimizing it with respect to each of the parameters in the model still holds.