Final Exam Review - Model vs Data

December 15, 2006

Model vs. Data

In most experiments we can control a set of independent variables $\underline{x}$ and can measure the value of the dependent variable y. For such a system we can propose a model which relates the value of the dependent variable to the values of the independent variables as shown in Equation(1)

$$y = f(\underline{x}; \underline{\Theta}) \tag{1}$$

The aim of the generating a model for the system is to obtain an answer to the following three questions

- For what value of $\underline{\Theta}$ is the deviation between model and data minimum?
- Is the model consistent with the data?
- What are the error bars on the values of parameters?

In the context of models we classify models as linear or non-linear. Linear models depend on the parameters $\underline{\Theta}$ linearly. For example the log of the rate of an arrhenius reaction is linear in the parameters $\log(A)$ and $\frac{E_a}{R}$. This model is linear even tough $\log(A)$ is not linearly dependent on the dependent variable temperature T.

$$\log(k) = \log(A) - \frac{E_a}{RT} \tag{2}$$

Usually in solving these problems we make the following two assumption

1. The dependent variable y, that is being measured is distributed normally (a gaussian distribution) around its mean value. This distribution can be due to many factors which are not in control of the experimentalist.
2. On the other hand the independent variables x are known exactly.

Best values of parameters

Let us assume that the model that we have is correct and that the standard deviation of the random errors in the measurement of the dependent variable y is σ. Then the probability of getting a measured value y_i is given by the Equation(3).

$$p(y_i) \propto \exp\left(-\frac{(y_i - f(\underline{x}_i; \underline{\Theta}))^2}{2\sigma^2}\right) \tag{3}$$

If we perform N different experiments then the probability of obtaining a vector $\underline{y}$ measurements of the dependent variables is given in Equation(4).

$$\begin{aligned} p(\underline{y}) &\propto \prod_{i=1}^{N} \exp\left(-\frac{(y_i - f(\underline{x}_i; \underline{\Theta}))^2}{2\sigma^2}\right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - f(\underline{x}_i; \underline{\Theta}))^2\right) \end{aligned} \tag{4}$$

The probability of getting this vector $\underline{y}$ becomes highest when sum of squares of errors become minimum. Thus the idea of minimizing the sum of squares of errors is based on the assumption that the errors in the measurement are normally distributed. Central limit theorem ensures that this assumption is justified if we assume that each measurement is obtained by performing many repeats.

If the model is linear then we have an analytical solution for the best fit values of the parameters. If the model is non-linear in parameters then there can potentially be multiple local minima and we have to be careful. The linear model can be written as shown in Equation(5).

$$\underline{y} = \underline{\underline{X}}\,\underline{\Theta} \tag{5}$$

The solution to the model is given by the expression in Equation(6).

$$\underline{\Theta} = (X^T X)^{-1} X^T \underline{y} \tag{6}$$

A potentially better method of solving minimization problem is by performing SVD decomposition of X (recall SVD of X gives three matrices U, Σ and V which are related to X as $X = U\Sigma V$). The best fit values of $\underline{\Theta}$ are given in Equation(7).

$$\underline{\Theta} = \sum_{i=1}^{N}\left(\frac{U_i.\underline{y}}{\sigma_i}\right)V_i \tag{7}$$

Model consistency

Let σ be the variance in the measured data. Then the probability that we get a vector $\underline{y}$ of measured values is given by Equation(4). Now we define a parameter χ^2 as

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - f(\underline{\mathrm{x}}_i; \underline{\Theta})}{\sigma} \right)^2$$

The least square method of calculating the parameters is nothing but the same as minimizing the value of χ^2. Also χ^2 is the a sum of N normally distributed quantities with mean 0 and variance 1. This χ^2 is itself a random variable and is distributed as chi-square distribution with $N - \dim(\theta)$ degrees of freedom. This chi-square distribution can be used to quantify the goodness of the fit. The probability of a model being correct is given by the area under the curve of a chi-square pdf between the abscissa χ^2 and inf.

Confidence intervals

If we know the value of σ we can assume that y is distirubted normally around its mean value $\hat{y}$ with a variance σ. We can then go ahead and calculate the approximate probability distribution functions of the parameters. From the probability distribution functions of the parameters we can calculate the 95% confidence intervals for the parameters. If the model is linear in parameters then one would expect that if the pdf of y was normal then the pdf of the parameters would also be normal. This is infact true and for more than one parameter we obtain a higher dimension gaussian. The covariance matrix for the parameters Θ; $cov(\Theta) = \sigma^2 [X^T X \mid_{\Theta_M}]^{-1}$. The 95% confidence intervals for a parameter θ is given in Equation(8).

$$\theta_i = \theta_{M,j} \pm Z_{2.5}\sigma \left([X^T X \mid_{\underline{\theta}_M}]^{-1}_{jj} \right)^{-1/2} \tag{8}$$

An interesting thing to note in the above equation is that the error bars on a parameter θ_i depends on the matrix X and thus by cleverly chosing our experimental conditions we can use a X that minimizes the error bars on the parameter of interest. When the model is not linear we can still use Equation(8) to calculate the error bars on the parameters, only in this case we can generate a linearised design matrix using Equation(9) and then use Equation(7) to calculate the error bars on the matrix.

$$X_{i,j} = \frac{\partial f(\underline{x}_i, \underline{\Theta})}{\partial \theta_j} \tag{9}$$

This is the same way that matlab function `nlinfit` and `nlparci` work. Note that by using this linearized design matrix we lose the information

of the covariance between different parameters. The graphical way to look at any pair of parameters is to plot the χ^2 value for a range of these two parameters. To convert the χ^2 plot to a plot of probability we just calculate the value of $\Delta\chi^2 = \chi^2(\theta_1, \theta_2) - \chi^2_{min}$ and this value of $\Delta\chi^2$ is distributed with a chi-square distribution of 2 degrees of freedom. An example of this was worked out in the homework 9.