Statistical Interpretation of Least Squares

We consider experimental data with errors that are distributed according to the normal distribution. From the theory of probability and statistics we have

Central Limit Theorem. Let X_n be a sequence of independent identically distributed random variables with mean c and variance $\sigma^2 < \infty$. Then

$$\frac{X_1 + X_2 + \cdots + X_n - nc}{\sigma\sqrt{n}}$$

converges in distribution to a normal random variable with mean 0 and variance 1.

This theorem motivates the use of the normal distribution to model measurement errors which are supposed to result from the sum of many little errors.

Mathematically, we have a set of data points y_i that are equal to $f(x_i)$ plus an independent normally distributed error with mean 0 and variance σ_i^2. Physically, f represents the phenomenon we are trying to observe, mean 0 implies there are no systematic errors, different values of σ_i^2 indicate that some measurements are more precise than others, and independence implies that one measurement doesn't affect the next. Under these assumptions, our measurements y_i may be represented by the normally distributed random variables $Y_i = f(x_i) + E_i$ with mean $f(x_i)$ and variance σ_i^2.

Recall that if E is a normally distributed random variable with norm 0 and variance σ^2, then the probability that $E \in (a, b)$ is

$$P\big\{E \in (a,b)\big\} = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(\frac{-t^2}{2\sigma^2}\Big)\, dt$$

Therefore,

$$P\big\{Y_i \in (a,b)\big\} = \int_a^b \frac{1}{\sigma_i\sqrt{2\pi}} \exp\Big(\frac{-(t - f(x_i))^2}{2\sigma_i^2}\Big)\, dt$$

Generally, we suppose there is a parametrized family of functions F_c such that $f = F_c$ for some parameter c. Our task is to determine c from the measurements y_i. In particular, we will consider the case where F_c is the linear model

$$F_c(x) = \sum_{j=1}^{m} c_j \phi_j(x)$$

with functions $\phi_i : \mathbf{R} \to \mathbf{R}$ fixed.

Since there are errors in the measurements, we can not hope to find the exact value of c for which $f = F_c$. In fact, for any choice of c there is chance that the data points y_i are a specific realization of the random variables $Y_i = F_c(x_i) + E_i$. To overcome this problem, we look for the value of c for which the data points y_i were most likely to have come from the model $F_c(x_i) + E_i$. In statistics this is called the maximum likelihood estimator for c.

Since the Y_i's are independent, their joint probability density is the product of the probability density functions for each Y_i. Therefore

$$\begin{aligned} P\{Y \in A\} &= \int_A \prod_{i=1}^{n} \left(\frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(\frac{-(t_i - F_c(x_i))^2}{2\sigma_i^2} \right) \right) d^n t \\ &= \int_A \frac{1}{(\Pi_i \sigma_i)(\sqrt{2\pi})^n} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} \left(\frac{t_i - F_c(x_i)}{\sigma_i} \right)^2 \right) d^n t \end{aligned}$$

where $A \subseteq R^n$ and

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \qquad \text{and} \qquad t = \begin{bmatrix} t_1 \\ \vdots \\ t_n \end{bmatrix}.$$

Thus, maximizing the probability that Y_i is in a small neighborhood of y_i is the same as maximizing the probability density function

$$\frac{1}{(\Pi_i \sigma_i)(\sqrt{2\pi})^n} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} \left(\frac{y_i - F_c(x_i)}{\sigma_i} \right)^2 \right),$$

or equivalently minimizing

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{y_i - F_c(x_i)}{\sigma_i} \right)^2.$$

Denoting

$$\tilde{A} = \begin{bmatrix} \phi_1(x_1)/\sigma_1 & \phi_2(x_1)/\sigma_1 & \dots & \phi_m(x_1)/\sigma_1 \\ \phi_1(x_2)/\sigma_2 & \phi_2(x_2)/\sigma_2 & \dots & \phi_m(x_2)/\sigma_2 \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(x_n)/\sigma_n & \phi_2(x_n)/\sigma_n & \dots & \phi_m(x_n)/\sigma_n \end{bmatrix} \qquad \text{and} \qquad \tilde{y} = \begin{bmatrix} y_1/\sigma_1 \\ y_2/\sigma_2 \\ \vdots \\ y_n/\sigma_n \end{bmatrix},$$

we find that $\chi^2 = \|\tilde{A}c - \tilde{y}\|_2^2$ where $\tilde{A}c = \tilde{y}$ is a over determined system of linear equations with the unknown c. Numerically, we use the reduced QR-decomposition $\tilde{A} = \tilde{Q}_1 \tilde{R}_1$ to find the c which minimizes $\|\tilde{A}c - \tilde{y}\|_2$ by solving $\tilde{R}_1 c = \tilde{Q}_1^T \tilde{y}$.

In Matlab solving a least squares problem is even easier. Suppose $\tilde{A}$ is an $n \times m$ matrix and $\tilde{y}$ is an n vector where $n > m$. If $\tilde{A}$ and $\tilde{y}$ are already in memory, then

Matlab Example 19a

```
>> c=Atilde\ytilde
c =
    7.43796398057002
  -13.81315998713759
   -4.87320678928050
    1.95347713954574
```

gives the least squares solution to the over determined problem $Ac = y$.