Least-Square Fitting

James R. Graham
9/11/2009

A straight line fit

Suppose that we have a set of *N* observations (x_i, y_i) where we believe that the measured value, y, depends linearly on x, i.e.,

$$y = mx + c.$$

Given our data, what is the best estimate of *m* and *c*? Assume that x_i (the independent variable) is known exactly, and y_i (the dependent variable) is drawn from a Gaussian probability distribution function with standard deviation, $\sigma_i = const$. Under these circumstances the most likely values of *m* and *c* are those corresponding to the straight line with the total minimum square deviation, i.e., the quantity

$$\chi^2 = \sum_i [y_i - (mx_i + c)]^2$$

is minimized when *m* and *c* have their most likely values. Figure 1 shows a typical deviation.

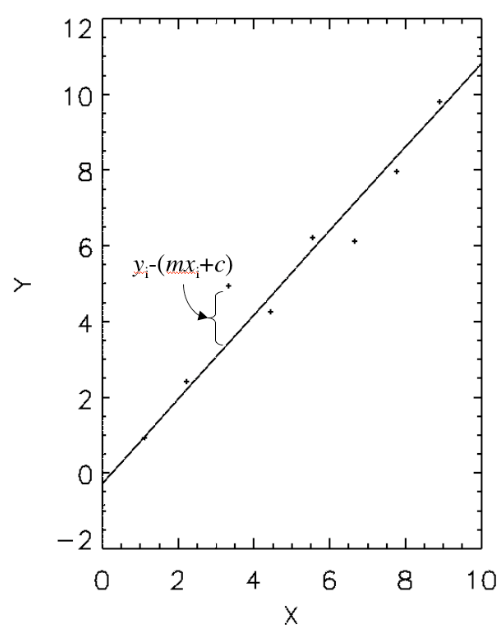

Figure 1: Some data with a least squares fit to a straight line. A typical deviation is illustrated.

Mathematically, the best values of m and c are found by solving the simultaneous equations,

$$\frac{\partial}{\partial m}\chi^2 = 0, \quad \frac{\partial}{\partial c}\chi^2 = 0.$$

Evaluating the derivatives yields

$$\frac{\partial}{\partial m}\chi^2 = \frac{\partial}{\partial m}\sum_i \left[y_i - (mx_i + c)\right]^2 = 2m\sum_i x_i^2 + 2c\sum_i x_i - 2\sum_i x_i y_i = 0$$
$$\frac{\partial}{\partial c}\chi^2 = \frac{\partial}{\partial c}\sum_i \left[y_i - (mx_i + c)\right]^2 = 2m\sum_i x_i + 2cN - 2\sum_i y_i = 0.$$

Which can conveniently be expressed in matrix form,

$$\begin{pmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & N \end{pmatrix}\begin{pmatrix} m \\ c \end{pmatrix} = \begin{pmatrix} \sum x_i y_i \\ \sum y_i \end{pmatrix}$$

and solved by multiplying both sides by the inverse,

$$\begin{pmatrix} m \\ c \end{pmatrix} = \begin{pmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & N \end{pmatrix}^{-1}\begin{pmatrix} \sum x_i y_i \\ \sum y_i \end{pmatrix}.$$

The inverse can be computed analytically, or in IDL it is trivial to compute the inverse numerically, as follows.

Example IDL

```
; Test least squares fitting by simulating some data.
nx = 20            ; Number of data points
m = 1.0            ; Gradient
c = 0.0            ; Intercept

x = findgen(nx)    ; Compute the independent variable

y = m*x + c + 1.0*randomn(iseed,nx) ; Compute the
                                    ; dependent
                                    ; variable and
                                    ; add Gaussian
                                    ; noise
plot,x,y,ps=1

; Construct the matrices
ma = [ [total(x^2), total(x)],[total(x), nx ] ]
mc = [ [total(x*y)],[total(y)]]

; Compute the gradient and intercept
md = invert(ma) ## mc
```

```
; Overplot the best fit
oplot, x, md[0,0]*x + md[0,1]
end
```

See Figure 2 for the output of this program.

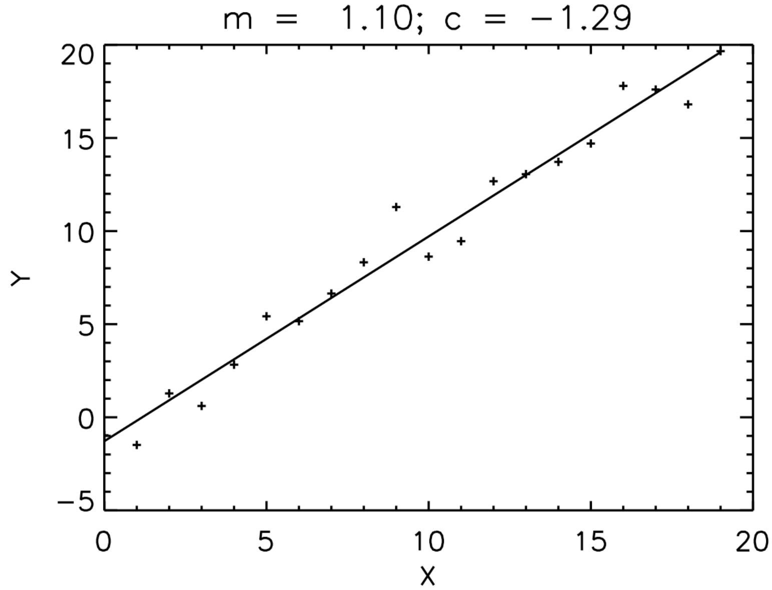

Figure 2—Least squares straight line fit. The true values are m = 1 and c = 0.