Physics 250

Least squares fitting.

Peter Young

(Dated: December 5, 2007)

I. INTRODUCTION

Frequently we are given a set of data points $(x_i, y_i), i = 1, 2, \cdots, N$, through which we would like to fit to a smooth function. The function could be straight line (the simplest case), a higher order polynomial, or a more complicated function. As an example, the data in Fig. 1 seems to follow a linear behavior and we may like to determine the "best" straight line through it. More generally, our fitting function, $y = f(x)$, will have some adjustable parameters and we would like to determine the "best" choice of those parameters.

The definition of "best" is not unique. However, the most useful choice, and the one nearly always taken, is the "least squares" fit, in which one minimizes the sum of the squares of the difference between the observed y-value, y_i, and the fitting function evaluated at x_i, i.e.

$$\boxed{\sum_{i=1}^{N} \left(y_i - f(x_i) \right)^2 .} \tag{1}$$

The simplest cases, and the only ones to be discussed in detail here, are where the fitting

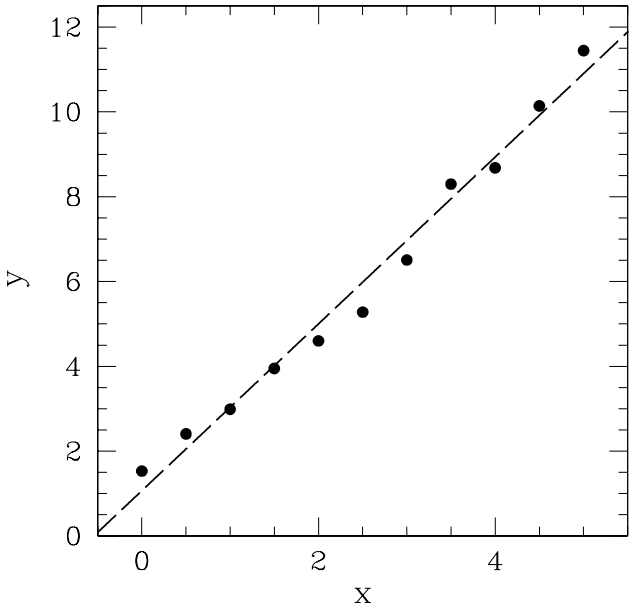

FIG. 1: Data with a straight line fit.

function is a *linear function of the parameters*. We shall call this a *linear model*. Examples are a straight line

$$y = a_0 + a_1 x \tag{2}$$

and an m-th order polynomial

$$y = a_0 + a_1 x + a_2 x^2 + \cdots + a_m x^m = \sum_{\alpha=0}^{m} a_\alpha x^m \,, \tag{3}$$

where the parameters to be adjusted are the a_α. (Note that we are *not* requiring that y is a linear function of x, only of the fit parameters a_α.)

An example where the fitting function depends *non*-linearly on the parameters is

$$y = a_0 x^{a_1} + a_2 \,. \tag{4}$$

Linear models are fairly simply because, as we shall see, the parameters are determined by *linear* equations, which always have a unique solution which can be found by straightforward methods. However, for fitting functions which are non-linear functions of the parameters, the resulting equations are *non-linear* which may have many solutions or non at all, and so are much less straightforward to solve.

Sometimes a non-linear model can be transformed into a linear model by a change of variables. For example if we want to fit to

$$y = a_0 x^{a_1} \,, \tag{5}$$

which has a non-linear dependence on a_1, then taking logs gives

$$\ln y = \ln a_0 + a_1 \ln x \,, \tag{6}$$

which is a *linear* function of the parameters $a_0' = \ln a_0$ and a_1. Fitting a straight line to a log-log plot is a very common procedure in science and engineering.

II. FITTING TO A STRAIGHT LINE

To see how least squares fitting works, consider the simplest case of a straight line fit, Eq. (2), for which we have to minimize

$$F(a_0, a_1) = \sum_{i=1}^{N} \left(y_i - a_0 - a_1 x_i \right)^2 \,, \tag{7}$$

with respect to a_0 and a_1. Differentiating F with respect to these parameters and setting the results to zero gives

$$\sum_{i=1}^{N}(a_0 + a_1 x_i) = \sum_{i=1}^{N} y_i, \tag{8a}$$

$$\sum_{i=1}^{N} x_i\,(a_0 + a_1 x_i) = \sum_{i=1}^{N} x_i y_i. \tag{8b}$$

We write this as

$$U_{00}\,a_0 + U_{01}\,a_1 = v_0, \tag{9a}$$

$$U_{10}\,a_0 + U_{11}\,a_1 = v_1, \tag{9b}$$

where

$$\boxed{U_{\alpha\beta} = \sum_{i=1}^{N} x_i^{\alpha+\beta},} \quad \text{and} \tag{10}$$

$$\boxed{v_\alpha = \sum_{i=1}^{N} y_i\,x_i^{\alpha}\,.} \tag{11}$$

Equations (9) are two linear equations in two unknowns. These can be solved by eliminating one variable, which immediately gives an equation for the second one. The solution can also be determined from

$$\boxed{a_\alpha = \sum_{\beta=0}^{m} \left(U^{-1}\right)_{\alpha\beta}\,v_\beta,} \tag{12}$$

where $m = 1$ here, and the inverse of the 2×2 matrix U is given, according to standard rules, by

$$U^{-1} = \frac{1}{\Delta}\begin{pmatrix} U_{11} & -U_{01} \\ -U_{01} & U_{00} \end{pmatrix} \tag{13}$$

where

$$\boxed{\Delta = U_{00}U_{11} - U_{01}^2,} \tag{14}$$

and we have noted that U is symmetric so $U_{01} = U_{10}$. The solution for a_0 and a_1 is therefore given by

$$\boxed{a_0 = \frac{U_{11}\,v_0 - U_{01}\,v_1}{\Delta},} \tag{15a}$$

$$\boxed{a_1 = \frac{-U_{01}\,v_0 + U_{00}\,v_1}{\Delta}.} \tag{15b}$$

We see that it is straightforward to determine the slope, a_1, and the intercept, a_0, of the fit from Eqs. (10), (11), (14) and (15) using the N data points (x_i, y_i).

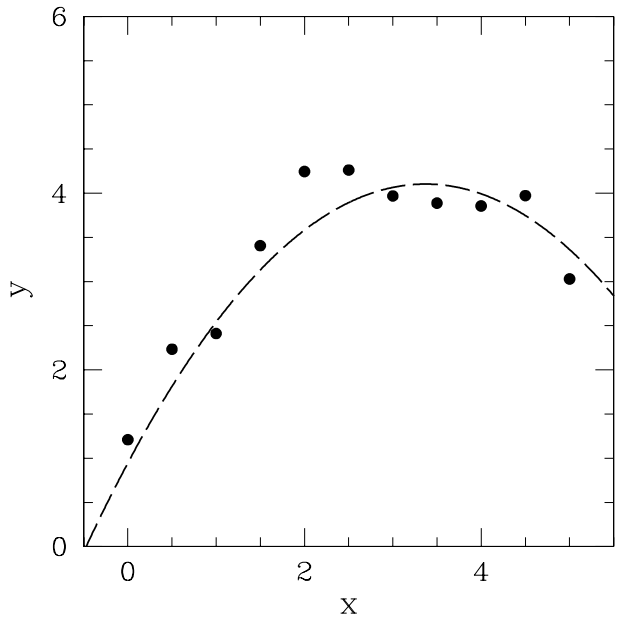

FIG. 2: Data with a parabolic fit.

III. FITTING TO A POLYNOMIAL

Frequently it may be better to fit to a higher order polynomial than a straight line, as for example in Fig. 2 where the fit is a parabola.

Using the notation for an m-th order polynomial in Eq. (3), we need to minimize

$$F(a_0, a_1, \cdots, a_m) = \sum_{i=1}^{N} \left(y_i - \sum_{\alpha=0}^{m} a_\alpha x_i^\alpha \right)^2 \tag{16}$$

with respect to the $M = m + 1$ parameters a_α. Setting to zero the derivative of F with respect to a_α gives

$$\sum_{i=1}^{N} x_i^\alpha \left(y_i - \sum_{\beta=0}^{m} a_\beta x_i^\beta \right) = 0, \tag{17}$$

which we write as

$$\boxed{\sum_{\beta=0}^{m} U_{\alpha\beta}\, a_\beta = v_\alpha,} \tag{18}$$

where $U_{\alpha\beta}$ and v_α are defined in Eqs. (10) and (11). Eq. (18) represents $M = m + 1$ *linear* equations, one for each value of α. Their solution is given formally by Eq. (12).

Hence polynomial least squares fits, being linear in the parameters, are also quite straightforward. We just have to solve a set of linear equations, Eq. (18), to determine the fit parameters.

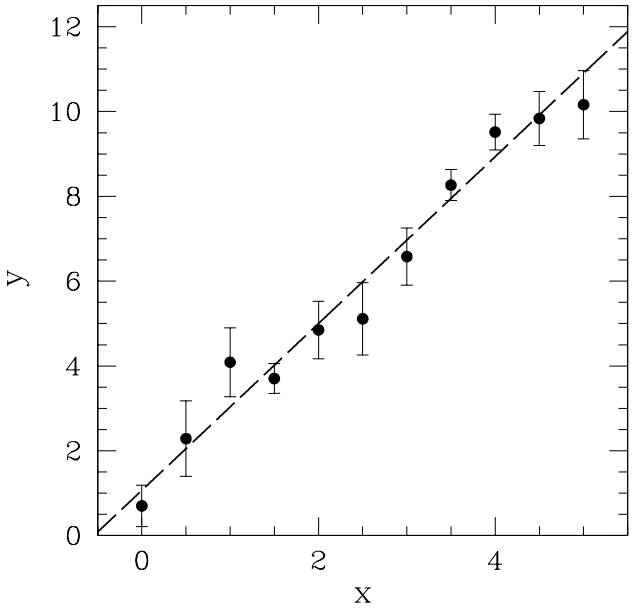

FIG. 3: Straight line fit with error bars.

IV. FITTING TO DATA WITH ERROR BARS

Frequently we have an estimate of the uncertainty in the data points, the "error bar". A fit would be considered satisfactory if it goes through the points within the error bars. An example of data with error bars and a straight line fit is shown in the figure below.

If some points have smaller error bars than other we would like to force the fit to be closer to those points. A suitable quantity to minimize, therefore, is

$$\boxed{\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - f(x_i)}{\sigma_i} \right)^2}, \tag{19}$$

called the "chi-squared" function, in which σ_i is the error bar for point i. Assuming a polynomial fit, we proceed exactly as before, and again find that the best parameters are given by the solution of Eq. (18), i.e.

$$\sum_{\beta=0}^{m} U_{\alpha\beta}\, a_\beta = v_\alpha, \tag{20}$$

but where now $U_{\alpha\beta}$ and v_α are given by

$$\boxed{U_{\alpha\beta} = \sum_{i=1}^{N} \frac{x_i^{\alpha+\beta}}{\sigma_i^2},} \quad \text{and} \tag{21}$$

$$\boxed{v_\alpha = \sum_{i=1}^{N} \frac{y_i\, x_i^\alpha}{\sigma_i^2}\,.} \tag{22}$$

The solution of Eqs. (20) can be obtained from the inverse of the matrix U, as in Eq. (12). Interestingly the matrix U^{-1} contains additional information. We would like to know the *range* of values of the a_α which provide a suitable fit. It turns out, see Numerical Recipes Secs. 15.4 and 15.6, that the square of the uncertainty in a_α is just the appropriate diagonal element of U^{-1}, so

$$\boxed{\delta a_\alpha^2 = \left(U^{-1}\right)_{\alpha\alpha}\,.} \tag{23}$$

For the case of a straight line fit, the inverse of U is given explicitly in Eq. (13). Using this information, and the values of (x_i, y_i, σ_i) for the data in the above figure, I find that the fit parameters (assuming a straight line fit) are

$$a_0 = 0.840 \pm 0.321, \tag{24}$$

$$a_1 = 2.054 \pm 0.109. \tag{25}$$

I had generated the data by starting with $y = 1+2x$ and adding some noise with zero mean. Hence the fit should be consistent with $y = 1 + 2x$ within the error bars, and it is.

V. CHI-SQUARED DISTRIBUTION

It is all very well to get the fit parameters and their error bars, but these results don't mean much unless the fit actually describes the data. Roughly speaking, this means that it goes through the data within the error bars. To see if this is the case, we look at the value χ^2 defined by Eq. (19) with the optimal parameters. Intuitively, we expect that if the fit goes through the data within about one error bar, then χ^2 should be about N. This indeed is correct, and, in fact, one can get much more information, including the distribution of χ^2, if we assume that the data points y_i have a Gaussian distribution. (This may not be the case, but the results obtained are often still a useful guide.)

Let us first assume that, apart from the (Gaussian) noise, the data exactly fits a polynomial, $y(x_i) = \sum_{\alpha=0}^{m} a_\alpha^{(0)} x_i^\alpha$ for some parameters $a_\alpha^{(0)}$. We first calculate the distribution of χ^2 where we

put in the *exact* values for the parameters, $a_\alpha^{(0)}$, rather than those obtained my minimizing χ^2 with respect to the parameters a_α. (Afterwords we will consider the effect of minimizing with respect to the a_α.) We therefore consider the distribution of

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - \sum_{\alpha=0}^{m} a_\alpha^{(0)} x_i^\alpha}{\sigma_i} \right)^2 . \tag{26}$$

Each of the N terms is the (square of) a Gaussian random variable with mean 0 and standard deviation unity. Denoting these variables by t_i then

$$\chi^2 = \sum_{i=1}^{N} t_i^2 \, . \tag{27}$$

Firstly, suppose that $N = 1$. The single variable t has the distribution

$$P_t(t) = \frac{1}{\sqrt{2\pi}} \, e^{-t^2/2} \, . \tag{28}$$

We actually want the distribution of t^2. Let us call $u = t^2$. To get the distribution of u, which we call $P_u(u)$, from the distribution of t, where u is a function of t, we note that the probability that t lies between t and $t + dt$ is the probability that u lies between u and $u + du$ where du/dt is the derivative, i.e.

$$P_t(t)dt = P_u(u)du, \text{ or } P_u(u) = P_t(t) \left| \frac{dt}{du} \right| , \tag{29}$$

where we noted that if the derivative is negative we must take the absolute value (since we are just equating the weight of the distribution of t over an interval dt to the weight of the distribution of u over an interval du). In this case we get

$$P_u(u) = \frac{1}{\sqrt{2\pi}} \, e^{-u/2} \, \frac{1}{2u^{1/2}} \, (\times 2) = \boxed{\frac{1}{\sqrt{2\pi u}} \, e^{-u/2} \, ,} \tag{30}$$

where we also multiplied by 2 since there are two solutions for t of $t^2 = u$. Remember Eq. (30) just corresponds to the Gaussian distribution in Eq. (28) but with $u = t^2$. It is instructive to check that the distribution in Eq. (30) is normalized.

$$\frac{1}{\sqrt{2\pi}} \int_0^\infty u^{-1/2} \, e^{-u/2} \, du = \frac{1}{\sqrt{\pi}} \int_0^\infty w^{-1/2} \, e^{-w} \, dw = \frac{1}{\sqrt{\pi}} \, \Gamma(1/2) = 1 \, , \tag{31}$$

where we made the substitution $w = u/2$ and used the result, discussed in class, that $\Gamma(1/2) = \sqrt{\pi}$.

As we have discussed, to obtain the distribution of a sum of random variables, such as in Eq. (27), it is useful to Fourier transform the distribution of the individual variables. We therefore take the Fourier transform of $P_u(u)$,

$$\widetilde{P}_u(k) = \frac{1}{\sqrt{2\pi}} \int_0^\infty u^{-1/2} \, e^{-u/2} \, e^{iku} \, du. \tag{32}$$

With the substitution $w = (1-2ik)(u/2)$ we get

$$\widetilde{P}_u(k) = \frac{1}{\sqrt{2\pi}} \sqrt{\frac{2}{1-2ik}} \int_0^\infty w^{-1/2}\, e^{-w}\, dw = \boxed{(1-2ik)^{-1/2}\,.} \tag{33}$$

Now χ^2 in Eq. (27) is just the sum of N random variables each of which has the distribution in Eq. (30). Hence the Fourier transform of the distribution of χ^2 is the N-th power of the Fourier transform of the distribution of one variable, which is given by Eq. (33), so

$$\widetilde{P}_{N,\chi^2}(k) = (1-2ik)^{-N/2}\,. \tag{34}$$

We will now verify that the function whose Fourier transform gives this expression is

$$\boxed{P_N(\chi^2) = \frac{1}{2^{N/2}\Gamma(N/2)}\, \left(\chi^2\right)^{\frac{N}{2}-1}\, e^{-\chi^2/2}\,.} \tag{35}$$

(Note that we are determining the distribution of χ^2, not χ; we are following convention in writing the basic variable as the square of something.) The Fourier transform of Eq. (35) is

$$\frac{1}{2^{N/2}\Gamma(N/2)} \int_0^\infty \left(\chi^2\right)^{\frac{N}{2}-1}\, e^{-\chi^2/2} e^{ik\chi^2}\, d\chi^2 = (1-2ik)^{-N/2}\, \frac{1}{\Gamma(N/2)} \int_0^\infty w^{\frac{N}{2}-1}\, e^{-w}\, dw = (1-2ik)^{-N/2} \tag{36}$$

which gives Eq. (34) as desired. We have again made the substitution $w = (1-2ik)(\chi^2/2)$, and we also used the definition of the $\Gamma(N/2)$.

Hence the χ^2 distribution for N variables is given by Eq. (35). From this it is easy to determine the mean and standard deviation of χ^2, if we remember the definition of the Gamma function:

$$\langle \chi^2 \rangle \;=\; 2\,\frac{\Gamma(N/2+1)}{\Gamma(N/2)} = 2\,(N/2) = \boxed{N,} \tag{37}$$

$$\langle \left(\chi^2\right)^2 \rangle \;=\; 2^2\,\frac{\Gamma(N/2+2)\,\Gamma(N/2+1)}{\Gamma(N/2)} = 4\left(\frac{N}{2}+1\right)\,\frac{N}{2} = N(N+2) \tag{38}$$

$$\sigma^2_{\chi^2} \;\equiv\; \langle \left(\chi^2\right)^2 \rangle - \langle \chi^2 \rangle^2 = \boxed{2N}. \tag{39}$$

Hence the average value of χ^2 is N (as we guessed intuitively above) and the standard deviation is $\sqrt{2N}$.

In my opinion, it is often more convenient to consider the distribution of χ^2/N which is called the $\boxed{\chi^2 \textit{ per degree of freedom}}$. This has mean unity (independent of N) and standard deviation $\sqrt{2/N}$.

Figure 4 shows the distribution of χ^2/N, the chi-squared per degree of freedom, for $N = 2, 5$ and 50. Note that for $N = 2$ the distribution of χ^2/N, is an exponential, for $N > 2$ it vanishes at the origin, and for $N \to \infty$ the central limit theorem tells us (as we can verify) that it is a Gaussian of mean 1 and standard deviation $\sqrt{2/N}$.

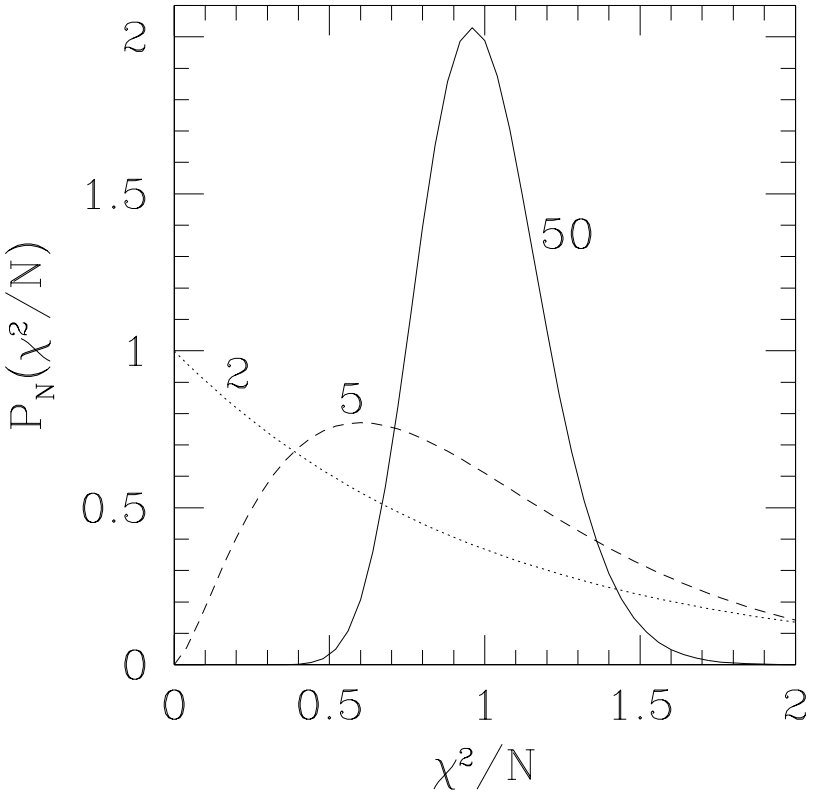

FIG. 4: The chi-squared distribution per degree of freedom for $N = 2, 5$ and 50.

We have obtained the distribution of χ^2 in Eq. (26) assuming that data fits exactly the specified polynomial apart from the Gaussian noise. However, in practice, we do not know the polynomial and we estimate it by *minimizing* χ^2 in Eq. (19) with respect to the parameters a_α. It turns out that the net effect is to decrease the "number of degrees of freedom", which was N before, by the number of parameters in the fit M (so for example $M = m + 1$ for an m-th order polynomial fit). To show this precisely requires some rather heavy math, which I will avoid, but instead indicate intuitively how the result arises.

Consider the simple case of a straight line fit, $y = a_0 + a_1 x$. Suppose we added a constant to all the y_i. Clearly χ^2 would be unchanged *once I've minimize with respect to* a_0 because the modified value of a_0 would exactly compensate for the shift in the y values. Similarly, if I change the y_i by an amount which is proportional to the x_i then again the χ^2 would be unchanged because the new value for a_1 would exactly compensate for the change in the data. Hence only changes in the y_i which are *orthogonal* to $\delta y_i = c_0$, and $\delta y_i = c_1 x_i$ give a change in χ^2 after minimization. There are $N - 2$ such linear combinations. It turns out that each gives a contribution to χ^2 with the same distribution as each of the terms in the unoptimized case, i.e. Eq. (30).

Hence, when we minimize χ^2 to get the best fit, if the fit were perfect apart from the random Gaussian noise in the data, the distribution of χ^2, would be the chi-squared distribution in Eq. (35), but with N replaced by ν, the "number of degrees of freedom", which is equal to **the number of data points N minus the number of fit parameters M**, i.e. the distribution of χ^2 is

$$\boxed{P_\nu(\chi^2) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} \left(\chi^2\right)^{\frac{\nu}{2}-1} e^{-\chi^2/2}\,.} \tag{40}$$

where

$$\boxed{\nu = N - M\,.} \tag{41}$$

Finally, all the discussion of fit parameters and their errors, is irrelevant if the curve does fit the data. It is therefore essential to also calculate a "goodness of fit parameter" Q. This is defined to be the probability that one would get a value for χ^2 greater than or equal to the observed one by chance. In otherwords it is the area under the curve of $P_\nu(\chi^2)$ to the right of the observed value. If χ^2 is less than $\nu - \sqrt{2\nu}$, Q will be close to 1 (so the fit is very likely), whereas if χ^2 is much greater than $\nu + \sqrt{2\nu}$, Q will be close to zero (so the fit is very unlikely).

An expression for Q can be found in terms of tabulated functions, since

$$Q = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} \int_{\chi^2}^{\infty} w^{\frac{\nu}{2}-1} e^{-w/2}\, dw = \frac{1}{\Gamma(\nu/2)} \int_{\chi^2/2}^{\infty} t^{\frac{\nu}{2}-1} e^{-t}\, dt = \boxed{\frac{\Gamma(\nu/2, \chi^2/2)}{\Gamma(\nu/2)}\,,} \tag{42}$$

where

$$\Gamma(a,x) = \int_x^\infty t^{a-1} e^{-t}\, dt \tag{43}$$

is called the incomplete Gamma function. Note that $\Gamma(a,0) = \Gamma(0)$ and $\Gamma(a,\infty) = 0$. See Numerical Recipes, Sec. 6.2 for discussion of these functions.

For further information on least square fitting, the interested student is referred to the books, such as Numerical Recipes, Ch. 15.