Straight-Line Fit

Dr. Darrel Smith[1]
Physics Department
Embry-Riddle Aeronautical University

(Dated: 26 February 2020)

The purpose of this document is to introduce students to the techniques used to perform a straight line fit to data that exhibit a linear trend. The data fitting techniques presented in this document can be applied to data with instrumental uncertainties, and uncertainties that are not constant between data points. A Monte Carlo data set is used to illustrate the effects of fitting the straight line function $y(x) = a + bx$. The equations used to calculate the fitting parameters a and b, and their respective uncertainties δa and δb are also presented. Comments regarding the χ^2 for the various fitting techniques are also presented.

I. BACKGROUND

The most common data sets encountered in physics labs are those exhibiting a linear trend. In most cases this is a reflection of the linear dependence of the phenomenon under observation. Many of these phenomena are described by simple functions, for example, $F = -(YA/L_0)\,x$, where the observed response to an external force F is a displacement from equilibrium measured by the parameter x. Students in our PS315 Modern Physics Lab will encounter this linear dependence in multiple labs.

Also, there are many computer programs already written to fit this kind of data. Two common languages used in this process are MATLAB and Mathematica. Because there is no accepted standard on our campus for software and hardware, the purpose of this note is to present the equations used in a straight-line fit for various kinds of data, independent of the computer program and computer platform (e.g., PC, Mac, etc.) While many of the programs can calculate the best fit parameters to a function $y(x) = a+bx$, many of them ignore the uncertainties associated with these parameters, δa and δb. In this paper, the techniques used to determine the uncertainties δa and δb are also described.

II. MONTE-CARLO GENERATED DATA SET

A sample data set of 31 points was generated where the data were *smeared* in the y-direction by randomly choosing values from a gaussian distribution with a σ=0.300. This kind of smearing is typical for a data set exhibiting *instrumental* uncertainties. The data set described in this paper contains 31 data points $\{x_i, y_i, \sigma\}$ and this is shown in Fig. 1.

III. WHICH PATH TO CHOOSE

There are two approaches to a linear fit and the choice depends upon the kind of uncertainties contained in the data. Either the data has uncertainties that vary from point-to-point denoted by σ_i, or the uncertainties are constant from point-to-point (i.e., instrumental uncertainties) denoted as σ. These two paths are shown in Fig.2.

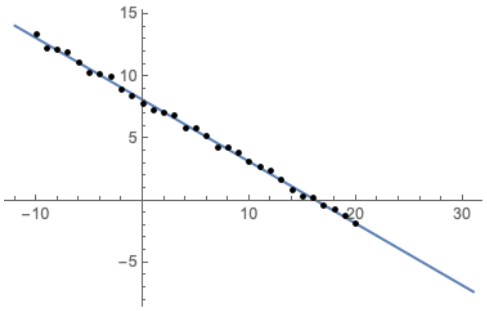

FIG. 1. This figure shows Monte-Carlo generated data $\{x_i, y_i, \sigma\}$ using a straight line function $y(x) = a + bx$ where the y values are smeared according to a gaussian function with $\sigma = 0.30$.

A. Variable uncertainties σ_i

The most general solution to a straight-line fit is where the uncertainties vary from point-to-point $\{x_i, y_i, \sigma_i\}$. The fitting parameters are determined from minimizing the χ^2, more specifically solving the system of two equations $\partial\chi^2/\partial a = 0$ and $\partial\chi^2/\partial b = 0$ where χ^2 is defined by Eq. 1

$$\chi^2 = \sum_{i=1}^{N} \frac{(\mathrm{dev}_i)^2}{\sigma_i^2} \qquad (1)$$
$$\mathrm{dev}_i = y_i - (a + bx_i)$$

The results from solving this system of two equations and two unknowns (a and b) is shown in Eq. 2.

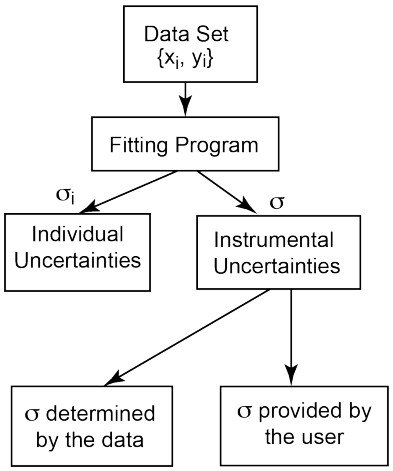

FIG. 2. This figure illustrates the two paths described in this paper. In the first case, the data contain uncertainties that change with each data point σ_i (e.g., random or calculated uncertainties). In the second case, the uncertainty is described by a single instrumental uncertainty σ that is either determined by the data, or provided by the user.

$$\Delta = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} - \left(\sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} \right)^2 \qquad (2)$$

$$a = \frac{1}{\Delta} \left[\sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2} - \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} \sum_{i=1}^{N} \frac{x_i y_i}{\sigma_i^2} \right]$$

$$b = \frac{1}{\Delta} \left[\sum_{i=1}^{N} \frac{1}{\sigma_i^2} \sum_{i=1}^{N} \frac{x_i y_i}{\sigma_i^2} - \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2} \right]$$

$$\delta a \simeq \sqrt{\frac{1}{\Delta} \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2}} \qquad \delta b \simeq \sqrt{\frac{1}{\Delta} \sum_{i=1}^{N} \frac{1}{\sigma_i^2}}$$

When these equations are applied to the data set above, the following parameters are obtained:

$$a = \ \ 7.89 \pm 0.06$$
$$b = -0.486 \pm 0.006$$

The χ^2 for $N - 2 = 29$ degrees of freedom is 21.27. In this fit, we presumed that all the uncertainties were the same $\sigma_i = 0.300$, even though the equations were set up to handle unequal σ_i.

B. Fixed Uncertainties σ

In this case, there is only one uncertainty σ instead of multiple uncertainties σ_i. The sole uncertainty can be (1) determined by the data, or (2) provided by the user.

1. Uncertainty – determined by the data

In the first case where the uncertainty is **determined by the data**, σ is calculated using the following equation:

$$\sigma \simeq \sqrt{\frac{1}{N-2} \sum_{i=1}^{N} \mathrm{dev}_i^2} \qquad (3)$$

The equations for calculating a, b, δa, and δb, are similar to Eqs. 2:

$$\Delta' = N \sum_{i=1}^{N} \left(x_i^2 \right) - \left(\sum_{i=1}^{N} x_i \right)^2 \qquad (4)$$

$$a = \frac{1}{\Delta'} \left[\sum_{i=1}^{N} x_i^2 \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} x_i y_i \right]$$

$$b = \frac{1}{\Delta'} \left[N \sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i \right]$$

$$\delta a \simeq \sigma \sqrt{\frac{\sum x_i^2}{\Delta'}} \qquad \delta b \simeq \sigma \sqrt{\frac{N}{\Delta'}}$$

Using the above equation for σ, the resulting χ^2 must be

$$\chi^2 = \sum_{i=1}^{N} \frac{(\mathrm{dev}_i)^2}{\sigma^2} = N - 2 = 29 \qquad (5)$$

the *number of degrees of freedom*. The parameters obtained using Eqs. 4 are:

$$a = \ \ 7.89 \pm 0.05$$
$$b = -0.487 \pm 0.005$$

While the values of a and b do not show much difference when compared to the previous values of a and b, the uncertainties would show significant difference if the degrees of freedom were smaller (e.g., $N - 2 = 1 \rightarrow 10$), instead of 29.

2. Uncertainty – determined by the user

If the user decides to introduce their own uncertainty (σ') in the data set $\{x_i, y_i, \sigma'\}$, then σ' will replace Eq. 3. The user should proceed with their data analysis and use Eqs. 4 to obtain a, b, δa, and δb. However the χ^2 will not be equal to $N-2$ as shown in Eq. 5. The normal rules of comparing χ^2 results to acceptable confidence levels then applies.

One could also take the modified data set $\{x_i, y_i, \sigma'\}$ and apply Eqs. 2 to their data analysis, and obtain the same results as in the previous paragraph.

IV. IDIOSYNCRASIES WITH GENFIT

When comparing these results with Genfit, the user could run into disagreements in two areas. First, Genfit evaluates the number of significant digits in the final parameters and their uncertainties. This is based on the number of significant digits provided by the user when passing the data set $\{x_i, y_i, \sigma_i\}$ to Genfit. Normally, this is a nice feature to have; however, the returned parameters can be significantly skewed. Some vendors present their data in integer format, or insufficient number of significant digits (e.g., the time markers in the Cavendish experiment). Thus Genfit prematurely reduces the number of significant digits in its calculations. The author of Genfit would rightfully say that the "vendor" should correct this oversight, and "they" should.

The second area of disagreement is how Genfit calculates the instrumental uncertainty shown in Eq. 3. It uses $1/N$ instead of $1/(N-2)$. Since two degrees of freedom are used to determine a and b for dev_i, this should be reflected in the uncertainty (σ) of the *sample* distribution as determined by the data. Once again, if the number of degrees of freedom is small (i.e., between $1 \rightarrow 10$), this will skew the uncertainties δa and δb coming from Genfit compared to standard Linear Regression models.