Chapter 12
Fitting Two-Variable Datasets with Bivariate Errors

Abstract The maximum likelihood method for the fit of a two-variable dataset described in Chap. 8 assumes that one of the variables (the independent variable X) has negligible errors. There are many applications where this assumption is not applicable and uncertainties in both variables must be taken into account. This chapter expands the treatment of Chap. 8 to the fit of a two-variable dataset with errors in both variables.

12.1 Two-Variable Datasets with Bivariate Errors

Throughout Chaps. 8 and 10 we have assumed a simple error model where the independent variable X is known without error, and all sources of uncertainty in the fit are due to the dependent variable Y. The two-variable dataset (X, Y) was effectively treated as a sequence of random variables of values $y_i \pm \sigma_i$ at a fixed location x_i with a parent model $y(x_i)$.

There are many applications, however, in which both variables have comparable uncertainties ($\sigma_x \simeq \sigma_y$) and there is no reason to treat one variable as independent. In general, a two-variable dataset is described by the datapoints

$$(x_i \pm \sigma_{xi}, y_i \pm \sigma_{yi})$$

and the covariance σ^2_{xyi} between the two measurements. One example is the two measurements of energy in the data in Table 6.1, where it would be appropriate to account for errors in both measurements. There is in fact no particular reason why one measurement should be considered as the independent variable and the other the dependent variable.

There are several methods to deal with two-variable datasets with bivariate error. Given the complexity of the statistical model, there is not a uniquely accepted solution to the general problem of fitting data with bivariate errors. This chapter presents two methods for the linear fit to data with two-variable errors. The first method (Sect. 12.2) applies to a linear fit and it is an extension of the least-squares method of Sect. 8.3. The second method (Sect. 12.3) is based on an alternative definition of χ^2 and it applies to any type of fit function. Although this method

M. Bonamente, *Statistics and Analysis of Scientific Data*, Graduate Texts in Physics, DOI 10.1007/978-1-4939-6572-4_12

does not have an analytic solution, it can be easily implemented using numerical methods such as Monte Carlo Markov chains described later in this book.

12.2 Generalized Least-Squares Linear Fit to Bivariate Data

In the case of identical measurement errors on the dependent variable Y and no error on the independent variable X, the least-squares method described in Sect. 8.3 estimated the parameters of the linear model as

$$\begin{cases} b = \dfrac{Cov(X,Y)}{Var(X)} = \dfrac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{N}(x_i - \overline{x})^2} \\ a = E(Y) - bE(X) = \dfrac{1}{N}\sum_{i=1}^{N} y_i - b\dfrac{1}{N}\sum_{i=1}^{N} x_i. \end{cases} \tag{12.1}$$

A generalization of this least-squares method accounts for the presence of measurement errors in the estimate of the variances and the covariance in (12.1). The methods of analysis presented in this section were developed by Akritas and Bershady [2] and others [22, 24]. Those references can be used as source of additional information on these methods for bivariate data.

Measurements of the X and Y variables can be described by

$$\begin{cases} x_i = \eta_{xi} + \epsilon_{xi} \\ y_i = \eta_{yi} + \epsilon_{yi}, \end{cases} \tag{12.2}$$

each the sum of a *parent* quantity and a measurement error, as in (11.1). Accordingly, the variances of the parent variables are given by

$$\begin{cases} Var(\eta_{xi}) = Var(x_i) - \sigma_{xi}^2 \\ Var(\eta_{yi}) = Var(y_i) - \sigma_{yi}^2. \end{cases} \tag{12.3}$$

This means that in (12.1) one must replace the sample covariance and variance by a *debiased* or *intrinsic* covariance and variance, i.e., quantities that take into account the presence of measurement errors.

The method of analysis that led to (12.1) assumes that the variable Y depends on X. In other words, we assumed that X is the independent variable. In this case, we talk of a fit of Y-given-X, or Y/X, and we write the linear model as

$$y = a_{Y/X} + b_{Y/X}x. \tag{12.4}$$

Modification of (12.1) with (12.3) (and an equivalent formula for the covariance) leads to the following estimator for the slope and intercept of the linear Y/X model:

$$\begin{cases} b_{Y/X} = \dfrac{Cov(X,Y) - \overline{\sigma_{xy}^2}}{Var(X) - \overline{\sigma_x^2}} = \dfrac{\sum_{i=1}^N (x_i - \overline{x})(y_i - \overline{y}) - \sum_{i=1}^N \sigma_{xyi}^2}{\sum_{i=1}^N (x_i - \overline{x})^2 - \sum_{i=1}^N \sigma_{xi}^2} \\ a_{Y/X} = \overline{y} - b_{Y/X}\overline{x}. \end{cases} \tag{12.5}$$

In this equation the sample variance and covariance of (12.1) were replaced with the corresponding intrinsic quantities, and the subscript Y/X indicates that X was considered as the independent variable.

A different result is obtained if Y is considered as the independent variable. In that case, the X-given-Y (or X/Y) model is described as

$$x = a' + b'y. \tag{12.6}$$

The same equations above apply by exchanging the two variables X and Y:

$$\begin{cases} b' = \dfrac{\sum_{i=1}^N (x_i - \overline{x})(y_i - \overline{y}) - \sum_{i=1}^N \sigma_{xyi}^2}{\sum_{i=1}^N (y_i - \overline{y})^2 - \sum_{i=1}^N \sigma_{yi}^2} \\ a' = \overline{x} - b'\overline{y}. \end{cases}$$

It is convenient to compare the results of the Y/X and X/Y fits by rewriting the latter in the usual form with x as the independent variable:

$$y = a_{X/Y} + b_{X/Y}x = -\frac{a'}{b'} + \frac{x}{b'}$$

for which we find that the slope and intercept are given by

$$\begin{cases} b_{X/Y} = \dfrac{\sum_{i=1}^N (y_i - \overline{y})^2 - \sum_{i=1}^N \sigma_{yi}^2}{\sum_{i=1}^N (x_i - \overline{x})(y_i - \overline{y}) - \sum_{i=1}^N \sigma_{xyi}^2} \\ a_{X/Y} = \overline{y} - b_{X/Y}\overline{x}. \end{cases} \tag{12.7}$$

In general the two estimators Y/X and X/Y will give different results for the best-fit line. This difference highlights the importance of interpreting the data to determine which variable should be considered the independent quantity.

Uncertainties in the parameters a and b and the covariance between them have been calculated by Akritas and Bershady [2]. For the Y/X estimator they can be

obtained via the following variables:

$$\begin{aligned}\xi_i =& \frac{(x_i - \bar{x})(y_i - b_{Y/X}x_i - a_{Y/X}) + b_{Y/X}\sigma_{xi}^2 - \sigma_{xyi}^2}{\frac{1}{N}\sum(x_i - \bar{x})^2 - \frac{1}{N}\sum\sigma_{xi}^2} \\ \zeta_i =& y_i - b_{Y/X}x_i - \bar{x}\xi_i.\end{aligned} \tag{12.8}$$

With these, the variances of a and b and the covariance is given by

$$\begin{cases}\sigma_{bY/X}^2 = \frac{1}{N}\sum(\xi_i - \bar{\xi})^2 \\ \sigma_{aY/X}^2 = \frac{1}{N}\sum(\zeta_i - \bar{\zeta})^2 \\ \sigma_{ab}^2 = \frac{1}{N}\sum(\xi_i - \bar{\xi})(\zeta_i - \bar{\zeta}).\end{cases} \tag{12.9}$$

For the X/Y estimator there are equivalent formulas for the ξ and ζ variables that need to be used in place of (12.8):

$$\begin{aligned}\xi_i =& \frac{(y_i - \bar{y})(y_i - b_{X/Y}x_i - a_{X/Y}) + b_{X/Y}\sigma_{xyi}^2 - \sigma_{yi}^2}{\frac{1}{N}\sum(x_i - \bar{x})(y_i - \bar{y}) - \frac{1}{N}\sum\sigma_{xyi}^2} \\ \zeta_i =& y_i - b_{X/Y}x_i - \bar{x}\xi_i.\end{aligned} \tag{12.10}$$

These values can then be used to calculate variances and the covariance of the parameters as in the Y/X fit.

Example 12.1 In Fig. 12.1 we illustrate the difference in the best-fit models when X is the independent variable (12.5) or Y is the independent variable (12.7), using the data of Table 6.1. The Y/X parameters are $a_{Y/X} = -0.367$ and $b_{Y/X} = 1.118$ and the X/Y parameters are $a_{X/Y} = -0.521$ and $b_{X/Y} = 1.132$. Unfortunately there is no definitive prescription to decide which variable should be regarded as independent. In this example each variable could be equally treated as the independent variable and the difference between the two best-fit models is relatively small. The difference between the two models for a value of the x axis of 1 is approximately 20 %. Note that the linear model and the data were plotted in a logarithmic scale to provide a more compact figure.

Also, the data of Table 6.1 do not report any covariance measurement and therefore the best-fit lines were calculated assuming independence between all measurements ($\sigma_{xyi}^2 = 0$). ◇

The example based on the data of Table 6.1 show that there is not just a single slope for the best-fit linear model, but that the results depend on which variable is assumed to be independent, as in the case of no measurement errors available

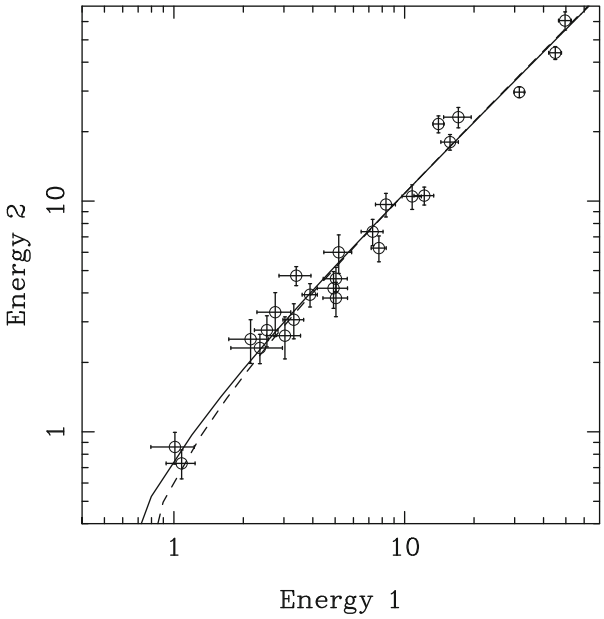

Fig. 12.1 Linear model fits to the data of Table 6.1 using the debiased variance method. The *solid line* is the model that uses Energy 1 as the independent variable X (12.4), the *dashed line* is the model that uses Energy 2 as the independent variable Y (12.6). Note the logarithmic scale for both axes

(Sect. 8.5). In certain cases it may be appropriate to use a model that is intermediate between the two Y/X and X/Y results. This is called the *bisector* model, which consists of the linear model that bisects the two lines obtained from the Y/X and X/Y fits described above. This method is also described by Akritas and Bershady [2] and Isobe and Feigelson [22] and the best-fit bisector line can be obtained from the following formulae:

$$\begin{cases} b_{bis} = \dfrac{b_{Y/X} b_{X/Y} - 1 + \sqrt{(1 + b_{Y/X}^2)(1 + b_{X/Y}^2)}}{b_{Y/X} + b_{X/Y}} \\ a_{bis} = \bar{y} - b_{bis} \bar{x}. \end{cases} \tag{12.11}$$

The uncertainties in the slope and intercept parameters can also be obtained using this definition for the ξ and ζ variables:

$$\begin{aligned}\xi_i =& \frac{(1+b_{X/Y}^2)b_{bis}}{(b_{Y/X}+b_{X/Y})\sqrt{(1+b_{Y/X}^2)(1+b_{X/Y}^2)}}\xi_{Y/X} + \\ & \frac{(1+b_{Y/X}^2)b_{bis}}{(b_{Y/X}+b_{X/Y})\sqrt{(1+b_{Y/X}^2)(1+b_{X/Y}^2)}}\xi_{X/Y} \\ \zeta_i =& y_i - b_{bis}x_i - \bar{x}\xi_i,\end{aligned} \tag{12.12}$$

where $\xi_{Y/X}$ is the ξ variable defined in (12.8) for the Y/X fit and $\xi_{X/Y}$ is the ξ variable defined in (12.10) for the X/Y fit.

Example 12.2 Figure 12.2 shows the fit to the variables Radius (X variable) and Ratio of thermal energies (Y variable) from Table 6.1. The solid line is the Y/X best-fit line with parameters $a = 1.1253$ and $b = -0.0005$, the dashed line is the X/Y best-fit line with parameters $a = 1.4260$ and $b = -0.0018$ and the dot-dash line is the bisector line with parameters $a = 1.2778$ and $b = -0.0011$. Notice how the Y/X and X/Y regressions give significantly different results. This is in part due to the presence of substantial scatter in the data, which results in several datapoints significantly distant from the best-fit regression lines. In the other

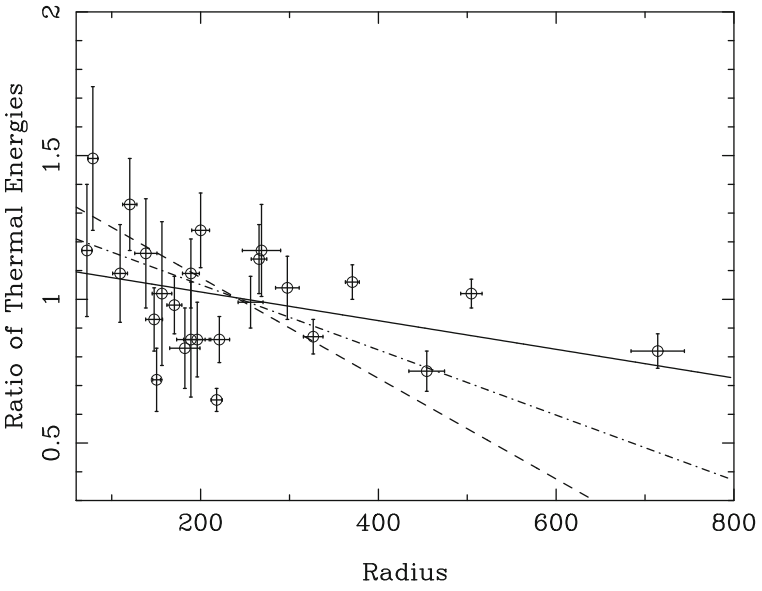

Fig. 12.2 Fit to the data of Table 6.1 using errors in both variables (see Example 12.2)

example of regression with errors in both variables (Fig. 12.1) the Y/X and X/Y best-fit lines were in better agreement. ◇

12.3 Linear Fit Using Bivariate Errors in the χ^2 Statistic

An alternative method to fit a dataset with errors in both variables is to re-define the χ^2 statistic to account for the presence of errors in the X variable. In the case of a linear fit, the square of the deviation of each datapoint y_i from the model is given by

$$(y_i - a - bx_i)^2. \tag{12.13}$$

When there is no error in the X variable, the variance of the variable in (12.13) is simply the variance of Y, σ_{yi}^2. In the presence of a variance σ_{xi}^2 for X, the variance of the linear combination $y_i - a - bx_i$ is given by

$$Var(y_i - a - bx_i) = \sigma_{yi}^2 + b^2\sigma_{xi}^2,$$

where a and b are the parameters of the linear model and the variables X and Y are assumed to be independent. This suggests a new definition of the χ^2 function for this dataset [35, 40], namely

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - a - bx_i)^2}{\sigma_{yi}^2 + b^2\sigma_{xi}^2}. \tag{12.14}$$

Since each term at the denominator is the variance of the term at the numerator, the new χ^2 variable defined in (12.14) is χ^2-distributed with $f = N - 2$ degrees of freedom.

The complication with the minimization of this function is that the unknown parameter b appears both at the numerator and the denominator of the function that needs to be minimized. As a result, an analytic solution to the maximum likelihood method cannot be given in general. Fortunately, the problem of finding the values of a and b that minimize (12.14) can be solved numerically. This method for the linear fit of two-variable data with errors in both coordinates is therefore of common use, and it is further described in [35].

Summary of Key Concepts for this Chapter

☐ *Data with bivariate errors*: A two-variable dataset that has errors in both variables. For these data there is no commonly accepted fit method.

☐ *Generalized least-squares fit to bivariate data*: An extension of the traditional ML fit to two-variable data. When x is the independent variable the best-fit parameters of the linear model are

$$\begin{cases} b_{Y/X} = \dfrac{Cov(X,Y) - \overline{\sigma_{xy}^2}}{Var(X) - \overline{\sigma_x^2}} \\ a_{Y/X} = \overline{y} - b_{Y/X}\overline{x}. \end{cases}$$

☐ *Bisector model*: A best-fit model for bivariate data that bisects the Y/X and X/Y models, intended to provide and intermediate model.

☐ *Use of bivariate errors in* χ^2: The χ^2 statistic can also be redefined to accommodate bivariate errors according to

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - a - bx_i)^2}{\sigma_{yi}^2 + b^2\sigma_{xi}^2}.$$

Problems

12.1 Use the bivariate error data of Energy 1 and Energy 2 from Table 6.1. Calculate the best-fit parameters and errors of the linear model Y/X, where X is Energy 1 and Y is Energy 2.

12.2 Use the bivariate error data of Energy 1 and Energy 2 from Table 6.1. Calculate the best-fit parameters and errors of the linear model X/Y, where X is Energy 1 and Y is Energy 2.

12.3 For the Energy 1 and Energy 2 data of Table 6.1, use the results of Problems 12.1 and 12.2 to calculate the bisector model to the Energy 1 vs. Energy 2 data.

12.4 Repeat Problem 12.1 for the Ratio vs. Radius data of Table 6.1.

12.5 Repeat Problem 12.2 for the Ratio vs. Radius data of Table 6.1.

12.6 Repeat Problem 12.3 for the Ratio vs. Radius data of Table 6.1.