Chapter 8
Maximum Likelihood Methods for Two-Variable Datasets

Abstract One of the most common tasks in the analysis of scientific data is to establish a relationship between two quantities. Many experiments feature the measurement of a quantity of interest as function of another control quantity that is varied as the experiment is performed. In this chapter we use the maximum likelihood method to determine whether a certain relationship between the two quantities is consistent with the available measurements and the best-fit parameters of the relationship. The method has a simple analytic solution for a linear function but can also be applied to more complex analytic functions.

8.1 Measurement of Pairs of Variables

A general problem in data analysis is to establish a relationship $y = y(x)$ between two random variables X and Y for which we have available a set of N measurements (x_i, y_i). The random variable X is considered to be the independent variable and it will be treated as having uncertainties that are much smaller than those in the dependent variable, i.e., $\sigma_x \ll \sigma_y$. This may not always be the case and there are some instances in which both errors need to be considered. The case of datasets with errors in both variables is presented in Chap. 12.

The starting point of the analysis of a two-dimensional dataset is an analytic form for $y(x)$, e.g., $y(x) = a + bx$. The function $f(x)$ has a given number of adjustable parameters a_k, $k = 1, \ldots, m$ that are to be constrained according to the measurements. When the independent variable X is assumed to be known exactly, then the two-variable data set can be described as a sequence of random variables $Y(X_i)$. For these variables we typically have a measurement of the standard error such that the two-variable data are of the form

$$(x_i, y_i \pm \sigma_i) \qquad i = 1, \ldots, N.$$

An example of this situation may be a dataset in which the size of an object is measured at different time intervals. In this example the time of measurement t_i is the independent variable, assumed to be known exactly, and $r_i \pm \sigma_i$ is the measurement of the size at that time interval. Although we call $y(x_i) = r_i \pm \sigma_i$ a

© Springer Science+Busines Media New York 2017

M. Bonamente, *Statistics and Analysis of Scientific Data*, Graduate Texts in Physics, DOI 10.1007/978-1-4939-6572-4_8

"measurement," it really may itself be obtained from a number of measurements from which one infers the mean and the variance of that random variable, as described in the earlier chapters. It is therefore reasonable to expect that the measurement provides also an estimate of the standard error.

Before describing the mathematical properties of the method used to estimate the best-fit parameters we need to understand the framework for the analysis. Consider as an example the case of a linear function between X and Y illustrated in Fig. 8.1. The main assumption of the method is that the function $y = y(x)$ is the correct description of the relationship between the two variables. This means that each random variable $y(x_i)$ is a Gaussian with the following parameters:

$$\begin{cases} \mu_i = y(x_i) & \text{the parent mean is determined by } y(x) \\ \sigma_i^2 & \text{variance is estimated from the data.} \end{cases} \tag{8.1}$$

Notice how this framework is somewhat of a hybrid: the parent mean is determined by the parent model $y(x)$ while the variance is estimated from the data. It should not be viewed as a surprise that the model $y = y(x)$ typically cannot determine by itself the variance of the variable. In fact, we know that the variance depends on the quality of the measurements made and therefore it is reasonable to expect that σ_i is estimated from the data themselves. In Sect. 8.2 we will use the assumption that Y has a Gaussian distribution, but this need not be the only possibility. In fact, in Sect. 8.8 we will show how data can be fit in alternative cases, such as when the variable has a Poisson distribution.

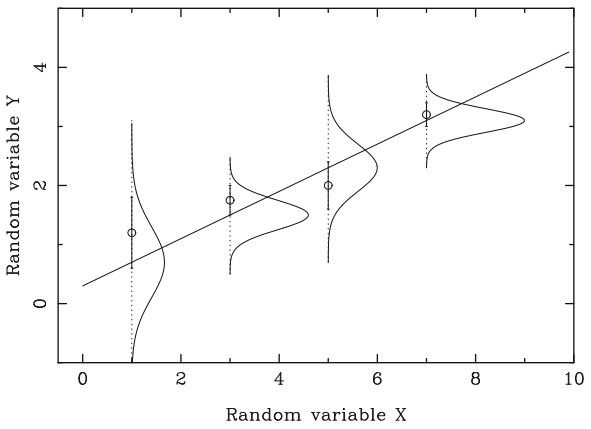

Fig. 8.1 In the fit of two-variable data to a linear function, measurements of the dependent variable Y are made for few selected points of the variable X (in this example $x_1 = 1, x_2 = 3, x_3 = 5$ and $x_4 = 7$). Each datapoint is marked by the *circle with error bars*. The independent variable X is assumed to be known exactly and the size of the *error bar* determines the value of the variance of $y(x_i)$

8.2 Maximum Likelihood Method for Gaussian Data

In many cases the variables $Y(X_i)$ have a Gaussian distribution, as illustrated in Fig. 8.1. The data are represented by points with an error bar and the model for each data point is a Gaussian centered at the value of the parent model $y(x_i)$. The model $y(x)$ can be any function and, as described in the previous section, the standard deviation σ_i is estimated from the data themselves.

The goal of fitting data to a model is twofold: to determine whether the model $y(x)$ is an accurate representation of the data and, at the same time, to determine what values of the adjustable parameters are compatible with the data. The two goals are necessarily addressed together. The starting point is the calculation of the likelihood $\mathscr{L}$ of the data with the model as

$$\mathscr{L} = P(\text{data/model}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(y_i - y(x_i))^2}{2\sigma_i^2}} = \left(\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2}} \right) e^{-\sum_{i=1}^{N} \frac{(y_i - y(x_i))^2}{2\sigma_i^2}} \tag{8.2}$$

In the previous equation we have assumed that the measurements $y_i \pm \sigma_i$ are independent of one other, so that the Gaussian probabilities can be simply multiplied. Independence between measurements is a critical assumption in the use of the maximum likelihood method.

The core of the maximum likelihood method is the requirement that the unknown parameters a_k of the model $y = y(x)$ are those that maximize the likelihood of the data. This is the same logic used in the estimate of parameters for a single variable presented in Chap. 5. The method of maximum likelihood results in the condition that the following function has to be minimized:

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - y(x_i)}{\sigma_i} \right)^2 \tag{8.3}$$

In fact, the factor in (8.2) containing the product of the sample variances is constant with respect to the adjustable parameters and maximization of the likelihood is obtained by minimization of the exponential term.

Equation (8.3) defines the goodness of fit statistic χ^2_{min}, which bears its name from the fact that it is distributed like a χ^2 variable. The number of degrees of freedom associated with this variable depends on the number of free parameters of the model $y(x)$, as will be explained in detail in Chap. 10. The simplest case is that of a model that has no free parameters. In that case, we know already that

the minimum χ^2 has exactly N degrees of freedom. Given the form of (8.3), the maximum likelihood method, when applied to Gaussian distribution, is also known as the *least squares* method.

8.3 Least-Squares Fit to a Straight Line, or Linear Regression

When the fitting function is

$$y(x) = a + bx \tag{8.4}$$

the problem of minimizing the χ^2 defined in (8.3) can be solved analytically. The conditions of minimum χ^2 are written as partial derivatives with respect to the two unknown parameters:

$$\begin{cases} \dfrac{\partial}{\partial a}\chi^2 = -2\sum \dfrac{1}{\sigma_i^2}(y_i - a - bx_i) = 0 \\ \dfrac{\partial}{\partial b}\chi^2 = -2\sum \dfrac{x_i}{\sigma_i^2}(y_i - a - bx_i) = 0 \end{cases} \tag{8.5}$$

$$\Rightarrow \begin{cases} \sum \dfrac{y_i}{\sigma_i^2} = a\sum \dfrac{1}{\sigma_i^2} + b\sum \dfrac{x_i}{\sigma_i^2} \\ \sum \dfrac{x_i y_i}{\sigma_i^2} = a\sum \dfrac{x_i}{\sigma_i^2} + b\sum \dfrac{x_i^2}{\sigma_i^2} \end{cases} \tag{8.6}$$

which is a system of two equations in two unknowns. The solution is

$$\begin{cases} a = \dfrac{1}{\Delta}\begin{vmatrix} \sum \dfrac{y_i}{\sigma_i^2} & \sum \dfrac{x_i}{\sigma_i^2} \\ \sum \dfrac{x_i y_i}{\sigma_i^2} & \sum \dfrac{x_i^2}{\sigma_i^2} \end{vmatrix}; \\ b = \dfrac{1}{\Delta}\begin{vmatrix} \sum \dfrac{1}{\sigma_i^2} & \sum \dfrac{y_i}{\sigma_i^2} \\ \sum \dfrac{x_i}{\sigma_i^2} & \sum \dfrac{x_i y_i}{\sigma_i^2} \end{vmatrix}. \end{cases} \tag{8.7}$$

where

$$\Delta = \sum \frac{1}{\sigma_i^2}\sum \frac{x_i^2}{\sigma_i^2} - \sum \frac{x_i}{\sigma_i^2}\sum \frac{x_i}{\sigma_i^2}. \tag{8.8}$$

Equation (8.7) provides the solution for the best-fit parameters of the linear model. The determination of the parameters of the linear model is known as *linear regression*.

When all errors are identical, $\sigma_i = \sigma$, it is easy to show that the best-fit parameters estimated by the least-squares method are equivalent to

$$\begin{cases} b = \dfrac{Cov(X,Y)}{Var(X)} \\ a = E(Y) - bE(X) \end{cases} \tag{8.9}$$

[see Problem (8.9)]. This means that, in the absence of correlation between the two variables, the best-fit slope will be zero and the value of a is simply the linear average of the measurements.

8.4 Multiple Linear Regression

The method outlined above in Sect. 8.3 can be generalized to a fitting function of the form

$$y(x) = \sum_{k=1}^{m} a_k f_k(x). \tag{8.10}$$

Equation (8.10) describes a function that is linear in the m parameters. In this case one speaks of *multiple linear regression*, or simply multiple regression. The functions $f_k(x)$ can have any analytical form. The linear regression described in the previous section has only two such function, $f_1(x) = 1$ and $f_2(x) = x$. A common case is when the functions are polynomials,

$$f_k(x) = x^k. \tag{8.11}$$

The important feature to notice is that the functions $f_k(x)$ do not depend on the parameters a_k.

We want to find an analytic solution to the minimization of the χ^2 with the fitting function in the form of (8.10). As we have seen, this includes the simple linear regression as a special case. In the process of χ^2 minimization we will also determine the variance and the covariances on the fitted parameters a_k, since no fitting is complete without an estimate of the errors and of the correlation between the coefficients. As a special case we will therefore also find the variances and covariance between the fit parameters a and b for the linear regression.

8.4.1 Best-Fit Parameters for Multiple Regression

Minimization of χ^2 with respect to the m parameters a_k is obtaining by taking partial derivatives over the m unknown parameters a_k.

This yields the following m equations:

$$\frac{\partial}{\partial a_l} \sum_{i=1}^{N} \left(\frac{(y_i - \sum_{k=1}^{m} a_k f_k(x_i))^2}{\sigma_i^2} \right) = 0$$

or

$$-2 \sum_{i=1}^{N} \left(\frac{y_i - \sum_{k=1}^{m} a_k f_k(x_i)}{\sigma_i^2} \right) f_l(x_i) = 0.$$

These equations can be written as

$$\sum_{i=1}^{N} \frac{f_l(x_i)}{\sigma_i^2} \left(y_i - \sum_{k=1}^{m} a_k f_k(x_i) \right) = 0 \tag{8.12}$$

leading to

$$\sum_{i=1}^{N} \frac{f_l(x_i) y_i}{\sigma_i^2} = \sum_{k=1}^{m} a_k \sum_{i=1}^{N} \frac{f_k(x_i) f_l(x_i)}{\sigma_i^2} \qquad l = 1, \ldots, m. \tag{8.13}$$

Equation (8.13) are m coupled equations in the parameters a_k, which can be solved using matrix algebra, as described below. Notice that the term $f_l(x_i)$ is the lth model component (thus the index l is not summed over), and the index i runs from 1 to N, where N is the number of data points.

The best-fit parameters are therefore obtained by defining the row vectors $\boldsymbol{\beta}$ and $\boldsymbol{a}$ and the $m \times m$ symmetric matrix $\boldsymbol{A}$ as

$$\begin{cases} \boldsymbol{\beta} & = (\beta_1, \ldots, \beta_m) & \text{in which } \beta_k = \sum_{i=1}^{N} f_k(x_i) y_i / \sigma_i^2 \\ \boldsymbol{a} & = (a_1, \ldots, a_m) & \text{(model parameters)} \\ A_{lk} & = \sum_{i=1}^{N} \frac{f_l(x_i) f_k(x_i)}{\sigma_i^2} & (l, k \text{ component of the m}\times\text{m matrix } \boldsymbol{A}) \end{cases}$$

With these definitions, (8.13) can be rewritten in matrix form as

$$\beta = aA, \tag{8.14}$$

and therefore the task of estimating the best-fit parameters is that of inverting the matrix $\boldsymbol{A}$, which can be done numerically. The m best-fit parameters a_k are placed in a row vector $\boldsymbol{a}$ (of dimensions $1 \times m$) and are given by

$$a = \beta A^{-1}. \tag{8.15}$$

The $1 \times m$ row vector β and the $m \times m$ matrix A can be calculated from the data and the fit functions $f_k(x)$.

8.4.2 Parameter Errors and Covariances for Multiple Regression

To calculate errors in the best-fit parameters, we treat parameters a_k as functions of the measurements, $a_k = a_k(y_i)$. Therefore we can use the error propagation method to calculate variances and covariances between parameters as:

$$\begin{cases} \sigma_{a_k}^2 = \sum_{i=1}^{N} \left(\dfrac{\partial a_k}{\partial y_i} \right)^2 \sigma_i^2 \\ \sigma_{a_l\, a_j}^2 = \sum_{i=1}^{N} \dfrac{\partial a_l}{\partial y_i} \dfrac{\partial a_j}{\partial y_i} \sigma_i^2. \end{cases} \tag{8.16}$$

We have used the fact that the error in each measurement y_i is given by σ_i and that the measurements are independent.

We show that the variance $\sigma_{a_l\, a_j}^2$ is given by the l, j term of the inverse of the matrix $\boldsymbol{A}$, which we define as the *error matrix*

$$\varepsilon = A^{-1}. \tag{8.17}$$

The error matrix ε is a symmetric matrix, of which the diagonal terms contain the variances of the fitted parameters and the off-diagonal terms contain the covariances.

Proof Use the matrix equation $a = \beta\varepsilon$ to write

$$a_l = \sum_{k=1}^{m} \beta_k \varepsilon_{kl} = \sum_{k=1}^{m} \varepsilon_{kl} \sum_{i=1}^{N} \frac{y_i f_k(x_i)}{\sigma_i^2} \Rightarrow \frac{\partial a_l}{\partial y_i} = \sum_{k=1}^{m} \varepsilon_{kl} \frac{f_k(x_i)}{\sigma_i^2}.$$

The equation above can be used into (8.16) to show that

$$\sigma^2_{a_l\, a_j} = \sum_{i=1}^{N} \left[\sigma_i^2 \sum_{k=1}^{m} \left(\varepsilon_{jk} \frac{f_k(x_i)}{\sigma_i^2} \right) \times \sum_{p=1}^{m} \left(\varepsilon_{lp} \frac{f_p(x_i)}{\sigma_i^2} \right) \right]$$

in which the indices k and p indicate the m model parameters, and the index i is used for the sum over the N measurements.

$$\Rightarrow \sigma^2_{a_l\, a_j} = \sum_{k=1}^{m} \varepsilon_{jk} \sum_{p=1}^{m} \varepsilon_{lp} \sum_{i=1}^{n} \frac{f_k(x_i) f_p(x_i)}{\sigma_i^2} = \sum_{k=1}^{m} \varepsilon_{jk} \sum_{p=1}^{m} \varepsilon_{lp} A_{pk}.$$

Now recall that A is the inverse of ε, and therefore the expression above can be simplified to

$$\sigma^2_{a_l\, a_j} = \sum_{k} \varepsilon_{jk} 1_{kl} = \varepsilon_{jl}. \tag{8.18}$$

□

8.4.3 Errors and Covariance for Linear Regression

The results of Sect. 8.4.2 apply also to the case of linear regression as a special case. We therefore use these results to estimate the errors in the linear regression parameters a and b and their covariance. In this case, the functions $f_l(x_i)$ are given, respectively, by $f_1(x) = 1$ and $f_2(x) = x$ and therefore the matrix A is a 2×2 symmetric matrix with the following elements:

$$\begin{cases} A_{11} = \sum_{i=1}^{N} 1/\sigma_i^2 \\ A_{12} = A_{21} = \sum_{i=1}^{N} x_i/\sigma_i^2 \\ A_{22} = \sum_{i=1}^{N} x_i^2/\sigma_i^2. \end{cases} \tag{8.19}$$

The inverse matrix $A^{-1} = \varepsilon$ is given by

$$\begin{cases} \varepsilon_{11} = A_{22}/\Delta \\ \varepsilon_{12} = \varepsilon_{21} = -A_{12}/\Delta \\ \varepsilon_{22} = A_{11}/\Delta \end{cases} \tag{8.20}$$

in which Δ is the determinant of A. Using (8.14) we calculate β:

$$\begin{cases} \beta_1 = \sum y_i/\sigma_i^2 \\ \beta_2 = \sum y_i x_i/\sigma_i^2 \end{cases} \tag{8.21}$$

and thus proceed to calculating the best-fit parameters and their errors. The best-fit parameters, already found in Sect. 8.3, are given by

$$(a, b) = (\beta_1, \beta_2) \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} \\ \varepsilon_{21} & \varepsilon_{22} \end{bmatrix}$$

which give the same results as previously found in (8.7). We are now in a position to estimate the errors in the best-fit parameters:

$$\begin{cases} \sigma_a^2 = \varepsilon_{11} = \dfrac{1}{\Delta} \displaystyle\sum_{i=1}^{N} x_i^2/\sigma_i^2 \\ \sigma_b^2 = \varepsilon_{22} = \dfrac{1}{\Delta} \displaystyle\sum_{i=1}^{N} 1/\sigma_i^2 \\ \sigma_{ab}^2 = \varepsilon_{12} = -\dfrac{1}{\Delta} \displaystyle\sum_{i=1}^{N} x_i/\sigma_i^2. \end{cases} \tag{8.22}$$

The importance of (8.22) is that the errors in the parameters a and b and their covariance can be computed analytically from the N measurements. This simple solution make the linear regression very simple to implement.

8.5 Special Cases: Identical Errors or No Errors Available

It is common to have a dataset where all measurements have the same error. When all errors in the dependent variable are identical ($\sigma_i = \sigma$) (8.7) and (8.22) for the linear regression are simplified to

$$\begin{cases} a = \dfrac{1}{\Delta} \dfrac{1}{\sigma^4} (\displaystyle\sum_{i=1}^{N} y_i \sum_{i=1}^{N} x_i^2 - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} x_i y_i) \\ b = \dfrac{1}{\Delta} \dfrac{1}{\sigma^4} (N \displaystyle\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} y_i \sum_{i=1}^{N} x_i) \\ \Delta = \dfrac{1}{\sigma^4} (N \displaystyle\sum_{i=1}^{N} x_i^2 - (\sum_{i=1}^{N} x_i)^2) \end{cases} \tag{8.23}$$

and

$$\begin{cases} \sigma_a^2 = \dfrac{1}{\Delta}\dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{N} x_i^2 \\ \sigma_b^2 = \dfrac{1}{\Delta}\dfrac{N}{\sigma^2} \\ \sigma_{ab}^2 = -\dfrac{1}{\Delta}\dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{N} x_i. \end{cases} \tag{8.24}$$

The important feature is that the best-fit parameters are *independent* of the value σ of the error.

For dataset that do not have errors available it is often reasonable to assume that all datapoints have the same error and calculate the best-fit parameters without the need to specify the value of σ. The variances, which depend on the error, cannot however be estimated. The absence of errors therefore limits the applicability of the linear regression method. It is in general not possible to reconstruct the errors σ_i a posteriori. In fact, the errors are the result of the experimental procedure that led to the measurement of the variables. A typical example is the case in which each of the variables $y(x_i)$ was measured via repeated experiments which led to the measurement of $y(x_i)$ as the mean of the measurements and its error as the square root of the sample variance. In the absence of the "raw" data that permit the calculation of the sample variance, it is simply not possible to determine the error in σ_i.

Another possibility to use a dataset that does not report the errors in the measurements is based on the assumption that the fitting function $y = f(x)$ *is* the correct description for the data. Under this assumption, one can estimate the errors, assumed to be identical for all variables in the dataset, via a *model sample variance* defined as

$$\sigma^2 = \frac{1}{N-m}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{8.25}$$

where $\hat{y}_i$ is the value of the fitting function $f(x_i)$ evaluated with the best-fit parameters, which must be first obtained by a fit assuming identical errors. The underlying assumption behind the use of (8.25) is to treat each measurement y_i as drawn from a parent distribution $f(x_i)$, $i = 1, \ldots N$, e.g., assuming that the model is the correct description for the data. In the case of a linear regression, $m = 2$, since two parameters (a and b) are estimated from the data. It will become clear in Sect. 10.1 that this procedure comes at the expenses of the ability to determine whether the dataset is in fact well fit by the function $y = f(x)$, since that is the working assumption.

In the case of no errors reported, it may not be clear which variable is to be treated as independent. We have shown in (8.9) that, when no errors are reported,

the best-fit parameters can be written as

$$\begin{cases} b = \dfrac{Cov(X,Y)}{Var(X)} \\ a = E(Y) - bE(X). \end{cases}$$

This equation clearly shows that the best-fit linear regression model is *dependent* on the choice of which between x and y is considered the independent variable. In fact, if y is regarded as the independent variable, and the data fit to the model

$$x = a' + b'y \tag{8.26}$$

the least-squares method gives the best-fit slope of

$$b' = \frac{Cov(X,Y)}{Var(Y)}.$$

When the model is rewritten in the usual form

$$y = a_{X/Y} + b_{X/Y}x$$

in which the notation X/Y means "X given Y," the best-fit model parameters are

$$\begin{cases} b_{X/Y} = -\dfrac{1}{b'} = \dfrac{Var(Y)}{Cov(X,Y)} \\ a_{X/Y} = E(Y) - b_{X/Y}E(X) \end{cases}$$

and therefore the two linear models assuming x or y as independent variable will be different from one another. It is up to the data analyst to determine which of the two variables is to be considered as independent when there is a dataset of (x_i, y_i) measurements with no errors reported in either variable. Normally the issue is resolved by knowing how the experiment was performed, e.g., which variable had to be assumed or calculated first in order to calculate or measure the second. Additional considerations for the fit of two-variable datasets are presented in Chap. 12.

8.6 A Classic Experiment: Edwin Hubble's Discovery of the Expansion of the Universe

In the early twentieth century astronomers were debating whether "nebulae," now known to be external galaxies, were in fact part of our own Galaxy, and there was no notion of the Big Bang and the expansion of the universe. Edwin Hubble pioneered the revolution via a seemingly simple observation that a

(continued)

number of "nebulae" moved away from the Earth with a velocity v that is proportional to their distance d, known as *Hubble's law*

$$v = H_0 d. \tag{8.27}$$

The quantity H_0 is the *Hubble constant*, typically measured in the units of $\mathrm{km\,s^{-1}\,Mpc^{-1}}$, where Mpc indicates a distance of 10^6 parsec. The data used by Hubble [21] is summarized in Table 8.1.

The quantity m is the apparent magnitude, related to the distance via the following relationship,

$$\log d = \frac{m - M + 5}{5} \tag{8.28}$$

where $M = -13.8$ is the absolute magnitude, also measured by Hubble as part of the same experiment, and considered as a constant for the purpose of this dataset, and d is measured in parsecs.

The first part of the experiment consisted in fitting the (v, m) dataset to a relationship that is linear in $\log v$,

$$\log v = a + b \cdot m \tag{8.29}$$

where a and b are the adjustable parameters of the linear regression. Instead of performing the linear regression described in Sects. 8.3 and 8.4.3, Hubble reported two different fit results, one in which he determined also the error in a,

$$\log v = (0.202 \pm 0.007) \cdot m + 0.472 \tag{8.30}$$

and one in which he fixed $a = 0.2$, and determined the error in b:

$$\log v = 0.2 \cdot m + 0.507 \pm 0.012. \tag{8.31}$$

Using (8.31) into (8.28), Hubble determined the following relationship between velocity and distance,

$$\log \frac{v}{d} = 0.2M - 0.493 = -3.253 \tag{8.32}$$

and this results in the measurement of his name-sake constant, $H_0 = v/d = 10^{-3.253} = 558 \times 10^{-6}\,\mathrm{km\,s^{-1}\,pc^{-1}}$, or $558\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$.

(continued)

Table 8.1 Data from E. Hubble's measurements

Name of nebula	Mean velocity km s^{-1}	Number of velocities	Mean m
Virgo	890	7	12.5
Pegasus	3810	5	15.5
Pisces	4630	4	15.4
Cancer	4820	2	16.0
Perseus	5230	4	16.4
Coma	7500	3	17.0
Ursa Major	11,800	1	18.0
Leo	19,600	1	19.0
(No name)	2350	16	13.8
(No name)	630	21	11.6

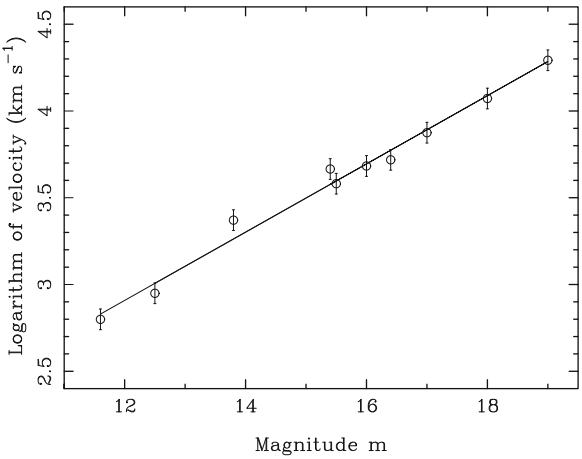

Fig. 8.2 Best-fit linear regression model for the data in Table 8.1

Example 8.1 The data from Hubble's experiment are a typical example of a dataset in which no errors were reported. A linear fit can be initially performed by assuming equal errors, and the best-fit line is reported in red in Fig. 8.2. Using (8.25), the common errors in the dependent variables $\log v(x_i)$ are found to be $\sigma = 0.06$, the best-fit parameters of the models are $a = 0.55 \pm 0.13$, $b = 0.197 \pm 0.0085$, and the covariance is $\sigma^2_{ab} = -1.12 \times 10^{-3}$, for a correlation coefficient of -0.99. The uncertainties and the covariance are measured using the method of (8.23). The best-fit line is shown in Fig. 8.2 as a solid line. ◇

8.7 Maximum Likelihood Method for Non-linear Functions

The method described in Sect. 8.4 assumes that the model is linear in the fitting parameters a_k. This requirement is, however, not necessary to apply the maximum likelihood criterion. We can assume that the relationship $y = f(x)$ has any analytic form and still apply the maximum likelihood criterion for the N measurements [see (8.3)]. The best-fit parameters are still those that minimize the χ^2 statistic. In fact, all considerations leading to (8.3) do not require a specific form for the fitting function $y = f(x)$. The assumption that must still be satisfied is that each variable y_i is Gaussian distributed, in order to obtain the likelihood in the form of (8.2).

The only complication for nonlinear functions is that an analytic solution for the best-fit values and the errors is in general no longer available. This is often not a real limitation, since numerical methods to minimize the χ^2 are available. The most straightforward way to achieve a minimization of the χ^2 as function of all parameters is to construct an m dimensional grid of all possible parameter values, evaluate the χ^2 at each point, and then find the global minimum. The parameter values corresponding to this minimum can be regarded as the best estimate of the model parameters. The direct grid-search method becomes rapidly unfeasible as the number of free parameters increases. In fact, the full grid consists of n^m points, where n is the number of discrete points into which each parameter is investigated. One typically wants a large number of n, so that parameter space is investigated with the necessary resolution, and the time to evaluate the entire space depends on how efficiently a calculation of the likelihood can be obtained. Among the methods that can be used to bypass the calculation of the entire grid, one of the most efficient and popular is the Markov chain Monte Carlo technique, which is discussed in detail in Chap. 16.

To find the uncertainties in the parameters using the grid search method requires a knowledge of the expected variation of the χ^2 around the minimum. This problem will be explained in the next chapter. The Markov chain Monte Carlo also technique provides estimates of the parameter errors and their covariance.

8.8 Linear Regression with Poisson Data

The two main assumptions made so far in the maximum likelihood method are that the random variables $y(x_i)$ are Gaussian and the variance of these variables are estimated from the data as the measured variance σ_i^2. In the following we discuss how the maximum likelihood method can be applied to data without making the assumption of a Gaussian distribution. One case of great practical interest is when variables have Poisson distribution, which is the case in many counting experiments. For simplicity we focus on the case of linear regression, although all considerations can be extended to any type of fitting function.

When $y(x_i)$ is assumed to be Poisson distributed, the dataset takes the form of (x_i, y_i), in which the values y_i are intended as integers resulting from a counting

experiment. In this case, the value $y(x_i) = a + bx_i$ is considered as the parent mean for a given choice of parameters a and b,

$$\mu_i = y(x_i) = a + bx_i. \tag{8.33}$$

The likelihood is calculated using the Poisson distribution and, under the hypothesis of independent measurements, it is

$$\mathscr{L} = \prod_{i=1}^{N} \frac{y(x_i)^{y_i} e^{-y(x_i)}}{y_i!}. \tag{8.34}$$

Once we remove the Gaussian assumption, there is no χ^2 function to minimize, but the whole likelihood must be taken into account. It is convenient to minimize the logarithm of the likelihood,

$$\ln \mathscr{L} = \sum_{i=1}^{N} y_i \ln y(x_i) - \sum_{i=1}^{N} y(x_i) + \mathrm{A} \tag{8.35}$$

where $A = -\sum \ln y_i!$ does not depend on the model parameters but only on the fixed values of the datapoints. Minimization of the logarithm of the likelihood is equivalent to a minimization of the likelihood, since the logarithm is a monotonic function of its argument. The principle of maximum likelihood requires that

$$\begin{cases} \dfrac{\partial}{\partial a} \ln \mathscr{L} = 0 \\ \dfrac{\partial}{\partial b} \ln \mathscr{L} = 0 \end{cases} \Rightarrow \begin{cases} N = \sum \dfrac{y_i}{a + bx_i} \\ \sum x_i = \sum \dfrac{x_i y_i}{a + bx_i}. \end{cases} \tag{8.36}$$

The fact that the minimization was done with respect to $\ln \mathscr{L}$ instead of χ^2 is a significant difference relative to the case of Gaussian data. For Poisson data we define the fit statistic C as

$$C = -2 \ln \mathscr{L} + B, \tag{8.37}$$

where B is a constant term. This is called the *Cash statistic*, after a paper by Cash in 1979 [9]. This statistic will be discussed in detail in Sect. 10.2 and it will be shown to have the property of being distributed like a χ^2 distribution with $N - m$ degrees of freedom in the limit of large N. This result is extremely important, as it allows to proceed with the Poisson fitting in exactly the same way as in the more common Gaussian case in order to determine the goodness of fit.

There are many cases in which a Poisson dataset can be approximated with a Gaussian dataset, and therefore use χ^2 as fit statistic. When the number of counts in each measurement y_i is approximately larger than 10 or so (see Sect. 3.4), the Poisson distribution is accurately described by a Gaussian of same mean and

variance. When the number of counts is lower, one method to turn a Poisson dataset into a Gaussian one is to *bin* the data into fewer variables of larger count rates. There are, however, many situations in which such binning is not desirable, especially when the dependent variable y has particular behaviors for certain values of the independent variable x. In those cases, binning of the data smears those features, which we would like to retain in the datasets. In those cases, the best option is to use the Poisson fitting method described in this section, and use C as the fit statistic instead.

Example 8.2 Consider a set of $N = 4$ measurements (3,5,4,2) to be fit to a constant model, $y = a$. In this case, (8.36) become

$$a = \frac{1}{N}\sum_{i=1}^{N} y_i$$

which means that the maximum likelihood estimator of a constant model, for a Poisson dataset, is the average of the measurements. The maximum likelihood best-fit parameter is therefore $a = 3.5$. ◇

Summary of Key Concepts for this Chapter

☐ *ML fit to two-dimensional data*: A method to find best-fit parameters of a model fit to x, y data assuming that one variable (typically x) is the independent variable.

☐ *Linear regression*: ML fit to a linear model, best-fit parameters when all errors are identical are

$$\begin{cases} b = \dfrac{Cov(X,Y)}{Var(X)} \\ a = E[Y] - bE[X] \end{cases}$$

(assuming x as independent variable).

☐ *Multiple linear regression*: An extension of the linear regression to models of the type

$$y = \sum a_k f_k(x).$$

☐ *Model sample variance*: When errors in the dependent variable (y) are not known, they can be estimated via the model sample variance

$$\sigma^2 = \frac{1}{N-m}\sum (y_i - \hat{y}_i)^2$$

where m is the number of model parameters.

Problems

8.1 Consider the data from Hubble's experiment in Table 8.1.

(a) Determine the best-fit values of the fit to a linear model for $(m, \log v)$ assuming that the dependent variables have a common value for the error.
(b) Using the best-fit model determined above, estimate the error from the data and the best-fit model, and then estimate the errors in the parameters a and b, and the correlation coefficient between a and b.
(c) Calculate the minimum χ^2 of the linear fit, using the common error as estimated in part (a).

8.2 Consider the following two-dimensional data, in which X is the independent variable, and Y is the dependent variable assumed to be derived from a photon-counting experiment:

x_i	y_i
0.0	25
1.0	36
2.0	47
3.0	64
4.0	81

(a) Determine the errors associated with the dependent variables Y_i.
(b) Find the best-fit parameters a, b of the linear regression curve

$$y(x) = a + bx;$$

also compute the errors in the best-fit parameters and the correlation coefficient between them;
(c) Calculate the minimum χ^2 of the fit, and the corresponding probability to exceed this value.

8.3 Consider the following Gaussian dataset in which the dependent variables are assumed to have the same unknown standard deviation σ,

x_i	y_i
0.0	0.0
1.0	1.5
2.0	1.5
3.0	2.5
4.0	4.5
5.0	5.0

The data are to be fit to a linear model.

(a) Using the maximum likelihood method, find the analytic relationships between $\sum x_i$, $\sum y_i$, $\sum x_i y_i$, $\sum x_i^2$, and the model parameters a and b.
(b) Show that the best-fit values of the model parameters are $a = 0$ and $b = 1$.

8.4 In the case of a maximum likelihood fit to a 2-dimensional dataset with equal errors in the dependent variable, show that the conditions for having best-fit parameters $a = 0$ and $b = 1$ are

$$\begin{cases} \sum_{i=1}^{N} y_i = \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} x_i y_i. \end{cases} \tag{8.38}$$

8.5 Show that the best-fit parameter b of a linear fit to a Gaussian dataset is insensitive to a change of all datapoints by the same amount Δx, or by the same amount Δy. You can show that this property applies in the case of equal errors in the dependent variable, although the same result applies also for the case of different errors.

8.6 The background rate in a measuring apparatus is assumed to be constant with time. N measurements of the background are taken, of which $N/2$ result in a value of $\overline{y}+\Delta$, and $N/2$ in a value $\overline{y}-\Delta$. Determine the sample variance of the background rate.

8.7 Find an analytic solution for the best-fit parameters of a linear model to the following Poisson dataset:

x	y
−2	−1
−1	0
0	1
1	0
2	2

8.8 Use the data provided in Table 6.1 to calculate the best-fit parameters a and b for the fit to the radius vs. pressure ratio data, and the minimum χ^2. For the fit, you can assume that the radius is known exactly, and that the standard deviation of the pressure ratio is obtained as a linear average of the positive and negative errors.

8.9 Show that, when all measurement errors are identical, the least squares estimators of the linear parameters a and b are given by $b = Cov(X, Y)/Var(X)$ and $a = E(Y) - bE(X)$.