Chapter 11
Systematic Errors and Intrinsic Scatter

Abstract Certain types of uncertainty are difficult to estimate and may not be accounted in the initial error budget. This sometimes leads to a poor goodness-of-fit statistic and the rejection of the model used to fit the data. These missing sources of uncertainty may either be associated with the data themselves or with the model used to describe the data. In both cases, we describe methods to account for these errors and ensure that hypothesis testing is not biased by them.

11.1 What to Do When the Goodness-of-Fit Test Fails

The first step to ensure that a dataset is accurately described by a model is to test that the goodness-of-fit statistic is acceptable. For example, when the data have Gaussian errors, χ^2_{min} can be used as the goodness-of-fit statistic. If the value of χ^2_{min} exceeds a critical value, it is recommended that one rejects the model. At that point, the standard option is to use an alternative model, and repeat the testing procedure.

There are cases when it is reasonable to try a bit harder and investigate further whether the model and the dataset may still be compatible, despite the poor goodness of fit. The general situation when additional effort is warranted is in the case of a model that generally follows the data without severe outliers, yet the best-fit statistic (such as χ^2_{min}) indicates that the model is not acceptable. An example of this situation is that of Fig. 10.1: the best-fit linear model follows the distribution of the data without systematic deviations, yet its high value of $\chi^2_{min} = 60.5$ for 23 degrees of freedom cannot be formally accepted at any level of confidence.

In this chapter we describe two types of analysis that can be performed when the fit of a dataset to a model is poor. The first method assumes that the model itself has a degree of uncertainty that results in an intrinsic scatter above and beyond the variance of the data (Sect. 11.2). The second investigates whether there are additional sources of error in the data that may not have been properly accounted (Sect. 11.3). The two methods are conceptually different but result in similar modifications to the analysis.

M. Bonamente, *Statistics and Analysis of Scientific Data*, Graduate Texts in Physics, DOI 10.1007/978-1-4939-6572-4_11

11.2 Intrinsic Scatter and Debiased Variance

When fitting a dataset to a model we assume that the data are drawn from a parent model that is described by a number of parameters. As such, we surmise that there are exact model parameters that describe the parent distribution of the data, although we don't know their precise values. We use the data to estimate them, typically through a maximum likelihood method that consists of finding model parameters that maximize the likelihood of the data being drawn from that model (Chap. 8). For Gaussian data, the maximum likelihood method consists of finding the minimum of the χ^2 statistic.

A possible reason for a poor value of the minimum χ^2 statistic is that the model itself, although generally accurate, may have an *intrinsic scatter* or variance that needs to be accounted in the determination of the fit statistic. In other words, the parent model may not be exact but it may feature an inherent degree of variability. The goal of this section is to provide a method to describe and measure such scatter.

11.2.1 Direct Calculation of the Intrinsic Scatter

Each measurement in a dataset can be described as the sum of two variables,

$$y_i = \eta_i + \epsilon_i, \tag{11.1}$$

where η_i represents the parent value from which the measurement y_i is drawn and ϵ_i is the variable representing the measurement error. Usually, we assume that $\eta_i = y(x_i)$ is a fixed number, estimated by the least-squares (or other) method. Since ϵ_i is a variable of zero mean, and its variance is simply the measurement variance σ_i^2, (11.1) implies that the variance of the measurement y_i is just σ_i^2.

The model η_i may, however, be considered a variable with non-zero variance. This is to describe the fact that the model is not known exactly, but has an intrinsic degree of variability measured by its variance $\sigma_{int}^2 = Var(\eta_i)$. For simplicity, we assume that this model variance is constant for all points along the model. Under the assumption that the measurement error and the model are independent, variances of the variables on the right-hand side of (11.1) add and this yields to

$$\sigma_{int}^2 = Var(y_i) - \sigma_i^2. \tag{11.2}$$

The equation means that the intrinsic variance is obtained as the difference of the data variance minus the variance due to measurement errors. In keeping up with the definitions of (11.1), $Var(y_i)$ refers to the total variance of the i-th variable at location x_i. It is meaningful to calculate the average variance for all the y_i's assuming that each measurement is drawn from a parent mean of $\hat{y}_i$, the best-fit value of the model $y(x_i)$. In so doing, we make use of the fact that the model is not constant but it varies

at different positions. As a result, (11.2) can be used to calculate the intrinsic scatter or variance of the model σ_{int}^2 as

$$\sigma_{int}^2 = \frac{1}{N-m}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 - \frac{1}{N}\sum_{i=1}^{N}\sigma_i^2. \tag{11.3}$$

where m is the number of model parameters. The intrinsic variance can also be referred to as the *debiased variance*, because of the subtraction of the expected scatter (due to measurement errors) from the total sample variance. Equation (11.3) can be considered a generalization of (2.11) in two ways. First, the presence of errors in the measurements of y_i leads to the addition of the last term on the right-hand side. Second, the total variance of the data are calculated *not* relative to the data mean $\overline{y}$ but to the parent mean of each measurement. It is possible that the second term in the right-hand side of (11.3) is larger than the first term, leading to a negative value for the intrinsic variance. This is an indication that, within the statistical errors σ_i, there is no evidence for an intrinsic scatter of the model. This method to estimate the intrinsic scatter is derived from [2] and [24].

It is important to remember that in calculating the intrinsic scatter we have made the assumption that the model *is* an accurate representation of the data. This means that we can no longer test for the null hypothesis that the model represents the parent distribution—we have already assumed this to be the case.

When the model is constant, with $\hat{y}_i = \overline{y}$ being the sample mean, the intrinsic scatter is calculated as

$$\sigma_{int}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{y})^2 - \frac{1}{N}\sum_{i=1}^{N}\sigma_i^2. \tag{11.4}$$

In this case, (11.4) is an unbiased estimate of the variance of Y.

11.2.2 Alternative Method to Estimate the Intrinsic Scatter

An alternative method to measure the amount of extra variance in a fit makes use of the fact that, for a Gaussian dataset, the expected value of the reduced χ_{min}^2 is one. A large value of the minimum χ^2 can be reduced by increasing the size of the errors until $\chi_{red}^2 \simeq 1$, or

$$\chi_{min}^2 = \sum_{i=1}^{N}\frac{(y_i - \hat{y}_i)^2}{\sigma_i^2 + \sigma_{int}^2} \simeq N - m \tag{11.5}$$

where m is the number of free model parameters, and σ_{int} is the intrinsic scatter that makes the reduced χ^2 unity. In (11.5) we have made the following substitution

relative to the standard use of the χ^2_{min} method:

$$\sigma_i^2 \rightarrow \sigma_i^2 + \sigma_{int}^2. \tag{11.6}$$

This method is only approximate, in that an acceptable model need not yield exactly a value of $\chi^2_{red} = 1$. This method to estimate the intrinsic scatter is nonetheless useful as an estimate of the level of scatter present in the data. Like in the earlier method, the analyst is making the assumption that the model fits the data and that the extra variance is attributed to an intrinsic variability of the model (σ^2_{int}).

Example 11.1 The example shown in Fig. 10.1 illustrates a case in which the data do not show systematic deviations from a best-fit model, and yet the χ^2 test would require a rejection of the model. The quantities Energy 1 (independent variable) and Energy 2 were fit to a linear model, the best-fit linear model yielded a fit statistic of $\chi^2_{min} = 60.5$ for 23 degrees of freedom and the model was therefore not acceptable.

Making use of the methods developed in this section, we can estimate the intrinsic scatter that makes the model consistent with the data. Using (11.3), the intrinsic scatter is estimated to be $\sigma_{int} = 2.5$. This means that the model has a typical uniform variability of 2.5 units (the units are those of the y axis, in this case used to measure energy). Using (11.5), a value of $\sigma_{int} = 1.6$ is needed to obtain a reduced χ^2_{min} of unity. The two methods were not expected to provide the same answer since they are based on different assumptions. ◇

11.3 Systematic Errors

The errors described so far in this book are usually referred to as *random errors*, since they describe the uncertainties in the random variables of interest. There are many sources of random error. A common source of randome error is the Poisson or counting error which derives from measuring N counts in an experiment and results in an error of $\sqrt{N}$. Another source of error is due to the presence of a background that needs to be subtracted from the measured signal. In general, any instrument used to record data will have sources of error that causes the measurements to fluctuate randomly around its mean value.

One of the main tasks of a data analyst is to find all the important sources of error that contribute to the variance of the random variable of interest. A typical case is the measurement of a total signal T in the presence of a background B, where the random variable of interest is the background-subtracted signal S,

$$S = T - B. \tag{11.7}$$

If the background is measured independently from the signal T, then the variance of the source is

$$\sigma_S^2 = \sigma_T^2 + \sigma_B^2. \tag{11.8}$$

The lesson to learn is that the variance of the random variable of interest *S increases* when the background is subtracted. If one assumes that there is no background, or that the background is constant ($\sigma_B^2 = 0$), the random error associated with S may be erroneously underestimated.

The term *statistical error* is often used as a synonym of random error. Sometimes, however, it is used to designate the leading source of random error, such as the Poisson uncertainty in a counting experiment, not including other sources of random error that are equally statistical or random in nature. Such use is not accurate, but the reader should be aware that there is no universally accepted meaning for the term "statistical error."

The term *systematic error* designates sources of error that systematically shift the signal of interest either too high or too low. Sources of systematic errors need to be identified to correct the erroneous offset. A typical example is an instrument that is miscalibrated and systematically reports measurements that have an erroneous offset. Even after the correction for the offset, it is however quite likely that there still remains a source of error, for example associated with the fact that such correction may not be uniform for all datapoints. If the systematic error is additive in nature, i.e., it shifts the random variable X according to $X' = X \pm E$, then the variance of the data is to be modified according to

$$\sigma_i^{'\,2} = \sigma_i^2 + \sigma_E^2. \tag{11.9}$$

The term σ_E^2 denotes the variance of the systematic error E. If E is known exactly, then it would ideally have zero variance. But in all practical cases, there will be an additional source of variance from the correction of a systematic error that needs to be accounted. The modification of the error σ_i due to the presence of a source of systematic error is therefore identical in form to the presence of intrinsic error [compare (11.6) and (11.9)].

If the systematic error is multiplicative in nature, i.e., $X' = E \cdot X$, it may be convenient to use the logarithms, $\log X' = \log X + \log E$ and then proceed as in the case of a linear offset.

Example 11.2 Continuing with the example shown in Fig. 10.1, we can use the results provided in Example 11.1 to say that an additional error of $\sigma_E = 1.6$ would yield a fit statistic of $\chi^2_{min,red} = 1$. This means that a possible interpretation for the large value of χ^2_{min} is that we had neglected an additional source of error σ_E. This additional source of error would be in place of the intrinsic scatter, since either correction to the calculation of χ^2_{min} is sufficient to bring the data in agreement with the model.

The errors of the data in Fig. 10.1 accounted for several sources of random error, including Poisson errors in the counting of photons from these sources, the background subtraction and for errors associated with the model used to describe the distribution of energy. The additional error of order $\sigma_E = 1.6$ for each datapoint may therefore be (a) an intrinsic error of the model (as described in Example 11.1), (b) an additional error from the correction of certain systematic errors that were

performed in the process of the analysis or (c) an additional random error that were not already included in the original error budget. The magnitude of possible errors in cases (b) and (c) can be estimated based on the knowledge of the collection of the data and its analysis. If such errors cannot be as large as required to obtain an acceptable fit, the only remaining option is to attribute this error to an intrinsic variance of the model or to conclude that the model is not an accurate description of the data. ◇

11.4 Estimate of Model Parameters with Systematic Errors or Intrinsic Scatter

In Sects. 11.2 and 11.3 we have assumed that intrinsic scatter or additional sources of systematic errors could be estimated using the best-fit values $\hat{y}_i$ obtained from the fit *without* these errors. Systematic errors or intrinsic scatter, however, do have an effect on the estimate of model parameters. The presence of systematic errors or intrinsic scatter, as discussed earlier in this chapter, is accounted with the addition of another source of variance to the data according to

$$\sigma_i^{'\,2} = \sigma_i^2 + \sigma^2. \tag{11.10}$$

The quantity σ is either the systematic error σ_E not accounted in the initial estimate of σ_i, or the intrinsic scatter σ_{int}. Both cases lead to the same effect on the overall error budget and the χ^2 fit statistic to minimize becomes

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - y(x_i))^2}{\sigma_i^2 + \sigma^2}. \tag{11.11}$$

It is clear that repeating the fitting procedure with the larger $\sigma_i^{'}$ errors instead of the original error will lead to new best-fit values and new uncertainties for the model parameters. The effect of the larger errors is to de-weight datapoints that have small values of σ_i and in general to provide larger confidence intervals for the model parameters. An acceptable procedure to obtain truly *best-fit* values of model parameters and their confidence intervals is to first estimate the additional source of error σ (either an intrinsic scatter or additional statistical or systematic errors) and then repeat the fit.

Example 11.3 The linear fit to the data of Table 6.1 for Energy 1 (independent variable) and Energy 2 resulted in a $\chi^2_{min} = 60.5$ for 23 degrees of freedom. The fit was not acceptable at any level of confidence. In Example 11.1 we calculated that an additional variance of $\sigma^2 = 1.6$ yields a $\chi^2_{min} = 23$. We fit the data with the addition of this error to the dependent variable and find the best-fit values of $a = -0.085 \pm 0.48$, $b = 1.05 \pm 0.05$.

For comparison, the fit obtained with the original errors returned values of $a = -0.26 \pm 0.088$, $b = 1.04 \pm 0.27$. These values could not be properly called "best-fit," since the fit was not acceptable. Yet, comparison between these values and those for the $\chi^2_{red} = 1.0$ case shows that best-fit parameters are affected by the additional source of error and that the confidence intervals become larger with the increased errors, as expected. ◇

Summary of Key Concepts for this Chapter

- ☐ *Intrinsic scatter*: An uncertainty of the model that increases the measurement error according to $y_i = \eta_i + \epsilon_i$.
- ☐ *Debiased variance*: A correction to the measured variance that accounts for the presence of measurement errors,

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

  The square root provides a measure of the intrinsic scatter.
- ☐ *Systematic error*: A type of measurement error σ_E that systematically shifts the measurements (as opposed to the *statistical error* σ_i). The two errors typically are added in quadrature, $\sigma_i^{'\,2} = \sigma_i^2 + \sigma^2$.

Problems

11.1 Fit the data from Table 6.1 for the radius vs. ratio using a linear model and calculate the intrinsic scatter using the best-fit linear model.

11.2 Using the same data as in Problem 11.1, provide an additional estimate of the intrinsic scatter using the $\chi^2_{red} \simeq 1$ method.

11.3 Justify the $1/(N-m)$ and $1/(N-1)$ coefficients in (11.3) and (11.4).

11.4 Using the data for the Hubble measurements of page 157, assume that each measurement of $\log v$ has an uncertainty of $\sigma = 0.01$. Estimate the intrinsic scatter in the linear regression of $\log v$ vs. m.

11.5 Using the data of Problem 8.2, estimate the intrinsic scatter in the linear fit of the X, Y data.