2.3 Principle of Maximum Likelihood

Suppose we have a random sample, $\{X_1, ..., X_n\}$, drawn from the population with pdf $p(x;\theta)$. We define the **likelihood function**, $L(\theta)$, as the sampling distribution, $g(x_1, ..., x_n; \theta)$, of $\{X_1, ..., X_n\}$, but now considered as a function of θ. In other words we are now thinking of θ not as a fixed parameter, but as a *variable*. Thus,

$$L(\theta) \quad = \quad g(x_1, ..., x_n; \theta)$$

The principle of maximum likelihood essentially states that forming the likelihood function is a useful way to define a 'good' estimator of the parameter θ. The **maximum likelihood** estimator of θ, denoted by $\hat{\theta}_{\text{ML}}$, is the value of θ which maximises $L(\theta)$. Thus, $\hat{\theta}_{\text{ML}}$ satisfies

$$\frac{\partial L}{\partial \theta} = 0 \qquad \text{when} \quad \theta = \hat{\theta}_{\text{ML}}$$

We can think of this definition in the following way. Suppose the particular values observed in our random sample are $\{x_1, ..., x_n\}$. If we were to vary the parameter, θ, we would generate a family of different pdfs. $\hat{\theta}_{\text{ML}}$ is the value of θ corresponding to the pdf from which it is 'most likely' that the actual sample was drawn.

Note that if the $\{X_i\}$ are iid, then

$$L(\theta) \quad = \quad p(x_1;\theta)\, p(x_2;\theta) \ldots p(x_n;\theta)$$

We extend to the case where the pdf is a function of several unknown parameters in the obvious way

$$\frac{\partial L(\theta_1, ..., \theta_k)}{\partial \theta_j} = 0 \qquad \text{when} \quad \theta_j = \hat{\theta}_j \quad (j = 1, ..., k)$$

For an iid random sample, $\{X_1, ..., X_n\}$, from a normal pdf, the maximum likelihood estimators of the mean, μ, and variance, σ^2, are

$$\hat{\mu}_{\text{ML}} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \hat{\sigma}^2_{\text{ML}} = \frac{1}{n}\sum_{i=1}^{n}[x_i - \hat{\mu}_{\text{ML}}]^2$$

i.e. simply the sample mean and variance. These results are derived on the accompanying handout and in the lectures. We already know from the preceding section that the sample mean is an unbiased estimator. What about $\hat{\sigma}^2_{\text{ML}}$? After a great deal of rather tedious (and non-examinable! But see the handout anyway if you want to follow the details of the derivation) algebra, we can show that

$$E[\hat{\sigma}^2_{\text{ML}}] = \frac{n-1}{n}\sigma^2$$

i.e. the sample variance is a **biased** estimator of σ^2. In fact, for *any* pdf with finite variance, we have:-

$$E\left[\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{i=1}^{n} x_i\right)^2\right] = \frac{n-1}{n}\sigma^2$$

but we can easily define an unbiased estimator of σ^2 by multiplying the sample variance by $n/(n-1)$, i.e.

$$\hat{\sigma}^2_{\text{corr}} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu}_{\text{ML}})^2 \quad = \quad \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 \; - \; n\hat{\mu}^2_{\text{ML}}\right]$$

satisfies

$$E\left[\hat{\sigma}^2_{\text{corr}}\right] \quad = \quad \sigma^2$$

Why is $\hat{\sigma}^2_{\text{corr}}$ biased? If the mean, μ, were known *a priori*, then one can show that

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2\right] = \sigma^2$$

i.e. in this case the sample variance *is* unbiased. It is because in practice we also have to estimate μ that principle of maximum likelihood gives a biased estimator of σ^2.

2.4 : Least Squares Estimators

We now turn to another useful method for estimating parameters – the principle of least squares – which is particularly useful in astronomy where we often try to fit a simple functional relationship between two or more sets of observational data. To fix our ideas we will develop the theory of least squares in the context of a specific astronomical example: the period-luminosity (PL) relation for Cepheid variables.

2.4.1 : Preamble – The Cepheid PL relation

Cepheids are highly luminous pulsating stars whose pulsation period has been found to be related to their luminosity by a power law, i.e.

$$L \quad = \quad A\,P^b$$

where A and b are constants. The relation is usually considered in terms of magnitudes, i.e.

$$M \quad = \quad a \, + \, b\log P$$

The usefulness of Cepheids derives from the fact that their periods can be measured directly, thus allowing us to infer their absolute magnitude, and hence their distance via the familiar equation:-

$$m \quad = \quad M \, + \, 5\log r \, + \, 25$$

It is the step of inferring the absolute magnitude from the measured period which concerns us here, and which requires the application of statistical techniques. This is because, in practice, any group of Cepheids will not satisfy exactly the above linear relation between M and $\log P$. If we plotted the 'observed' values of $\{M_i, \log P_i\}$ for a sample of Cepheids we would expect the points to be scattered on the plane, due to a combination of observational errors in the measurement of M_i (which is, in any case, not measured directly but would itself have to be inferred from the measured *apparent* magnitude of the Cepheid combined with some independent estimate of its distance) and $\log P_i$, and intrinsic errors due to the inadequacy of the linear relation which we are assuming holds between these two quantities (recall the discussion in the introduction). Figure 11 shows the Cepheid PL relations derived for calibrating data in the LMC and SMC at a series of wavelengths from B to K. (In fact these plots show the apparent magnitude of the Cepheids, which **is** directly observed, but since the LMC and SMC Cepheids can all be assumed equidistant these apparent magnitudes are equivalent to absolute magnitudes, as is easily seen from the distance modulus formula above.)

As can be seen, these data clearly display a linear relationship but there is indeed a non-negligible scatter in the relation, so that – at a given period, there is a range, or distribution, of absolute magnitudes consistent with that period. But in order to use the PL relation to estimate the distance of a more remote Cepheid, we want to assign a **single** value of M to the star. In other words we want to fit a straight line (or more generally, a curve) through the $\{M_i, \log P_i\}$ scatterplot so that we have a one-to-one relationship between the observed (log) period and the inferred absolute magnitude.

We want this straight line to be the one which, in some sense, is the 'best fit' to the data – i.e. we want the observed data points (which we refer to as our 'calibrating data') to lie 'closest' to the best fit line. The principle of least squares provides us with a definition of what we mean by 'closest' in this context. We also want a means of quantifying whether the scatter of the data about this best fit straight line (what we call the **residuals** of the best fit) is consistent with our assumption of a straight line model in the first place. If a plot of our PL calibrating data looked like Figure 12, for example, then common sense would tell us that a straight line model was inappropriate. Statistics provides us with a means of quantifying this degree of 'inappropriateness' – what we call the **goodness of fit**.

Figure 11: PL relations for Cepheids in the LMC and SMC.

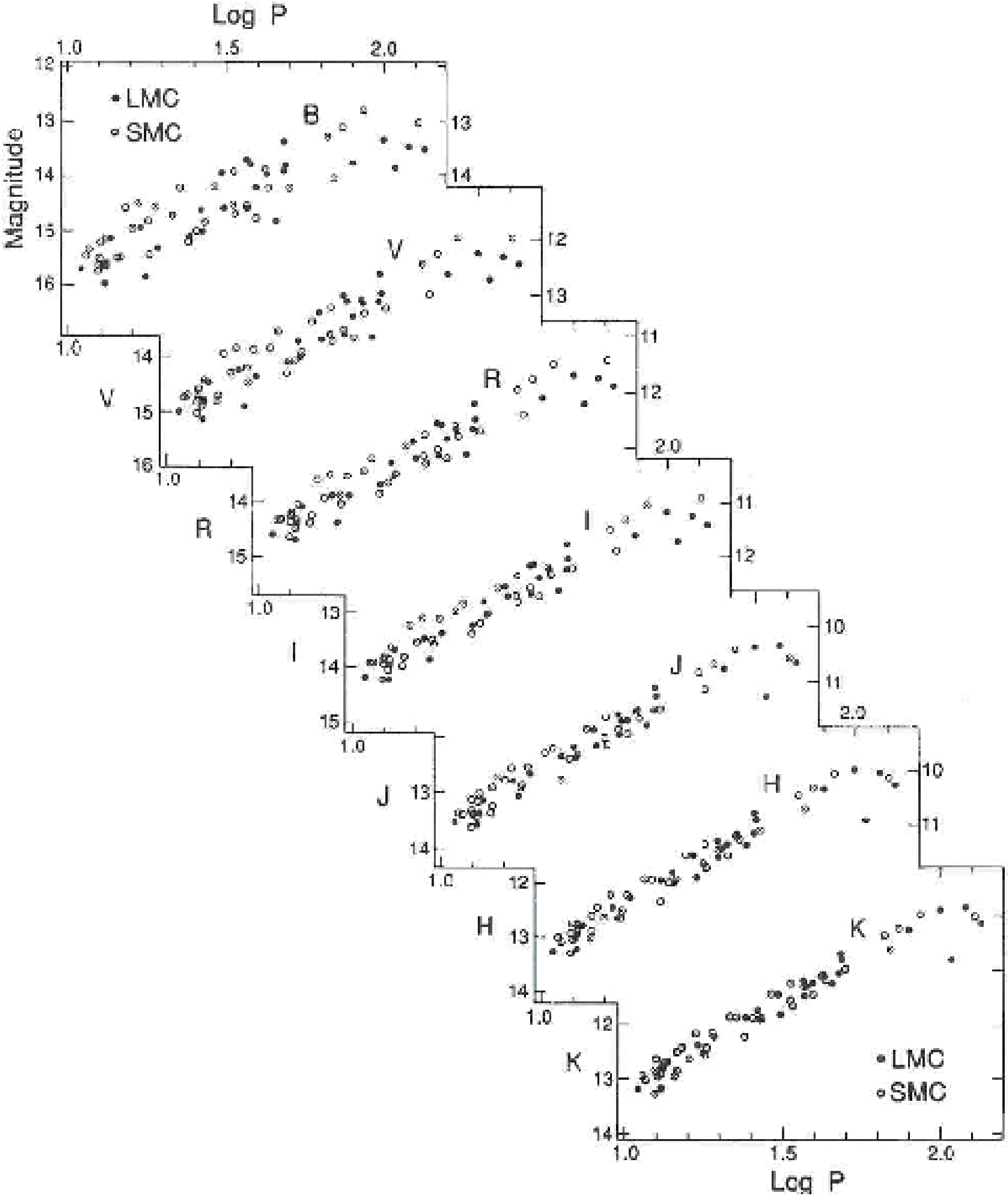
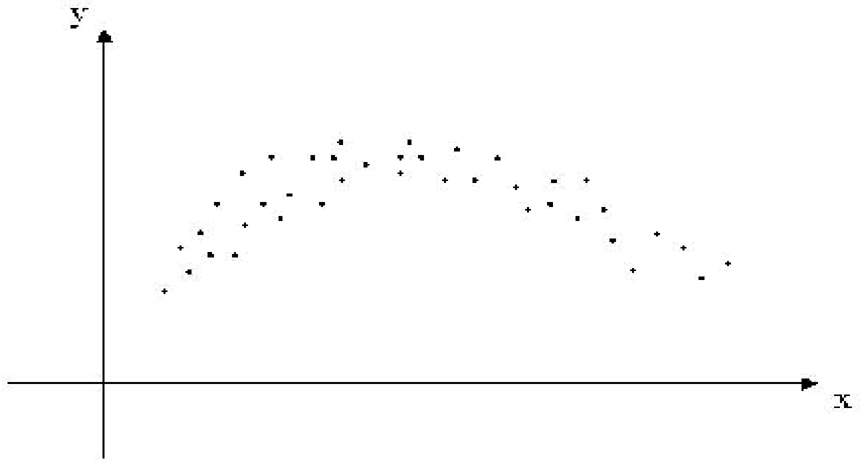


Figure 12: Data for which a straight line model is not appropriate.



2.4.2 : Ordinary Linear Least Squares

Suppose that the scatter in our plot of $\{M_i, \log P_i\}$ is assumed to arise from errors in only one of the two variables. This case is called **Ordinary Least Squares**. In the context of the PL relation, it is probably reasonable to assume that there is no error on the measured period of a Cepheid, or at least that this error is very small compared with the uncertainty on the absolute magnitude. We then call log period the **independent variable**, and absolute magnitude the **dependent variable**. Thus we suppose that we can write, for each Cepheid:-

$$M_i \quad = \quad a + b \log P_i + \epsilon_i$$

where ϵ_i is known as the **residual** of the i^{th} Cepheid – i.e. the difference between the observed value of M_i, and the value predicted by the best-fit straight line (see Figure 13).

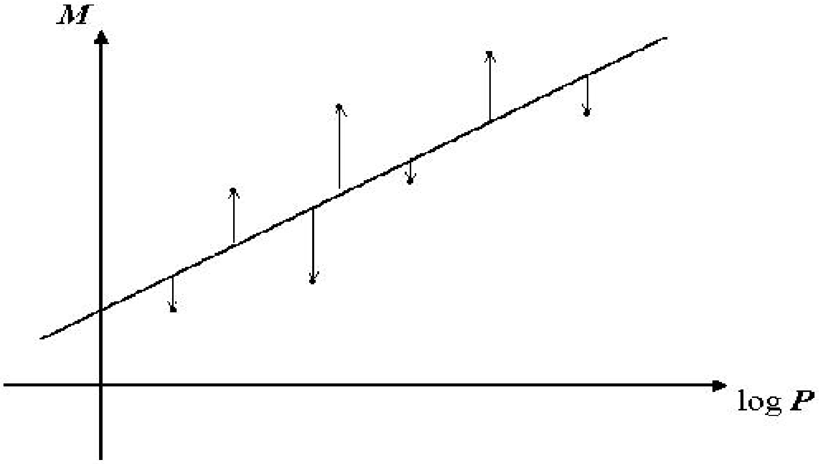
We assume that the $\{\epsilon_i\}$ are an iid random sample from some underlying pdf with mean zero and variance σ^2 – i.e. the residuals are equally likely to be positive or negative and all have equal variance.

The **principle of least squares** says that one should adopt as the best fit estimators of a and b the values which minimise the sum of the squared residuals, $S = \sum \epsilon^2$. Thus

$$S \quad = \quad \sum_{i=1}^{n} \left[M_i - (a + b \log P_i) \right]^2$$

and $\hat{a}$ and $\hat{b}$ are obtained by differentiating S with respect to a and b, setting the resulting equations (called the **normal equations**) equal to zero, and solving for a and b.

Figure 13: Schematic diagram indicating residuals of data points in the $\{M_i, \log P_i\}$ plane.



In general, if we write the linear relation as

$$Y_i \quad = \quad a + bX_i + \epsilon_i$$

where X_i is the independent variable and Y_i as the dependent variable, the **least squares estimators** of a and b minimise

$$S \quad = \quad \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

and $\hat{a}_{\mathrm{LS}}$ and $\hat{b}_{\mathrm{LS}}$ satisfy

$$\frac{\partial S}{\partial a} = 0 \quad \text{when} \quad a = \hat{a}_{\mathrm{LS}} \qquad \frac{\partial S}{\partial b} = 0 \quad \text{when} \quad b = \hat{b}_{\mathrm{LS}}$$

Solving these equations, $\hat{a}_{\mathrm{LS}}$ and $\hat{b}_{\mathrm{LS}}$ are given by

$$\hat{a}_{\mathrm{LS}} \quad = \quad \frac{\sum y_i \sum x_i^2 \ - \ \sum y_i x_i \sum x_i}{n \sum x_i^2 \ - \ (\sum x_i)^2}$$

$$\hat{b}_{\mathrm{LS}} \quad = \quad \frac{n \sum y_i x_i \ - \ \sum y_i \sum x_i}{n \sum x_i^2 \ - \ (\sum x_i)^2}$$

where n denotes the sample size and all summations are for $i = 1, ..., n$. If the residuals are drawn from a normal pdf then it is straightforward to show that the least squares estimators are also maximum likelihood estimators (see lectures).

It can also be shown that $\hat{a}_{\rm LS}$ and $\hat{a}_{\rm LS}$ are **unbiased** estimators of a and b respectively. The variance of $\hat{a}_{\rm LS}$ and $\hat{b}_{\rm LS}$ is given by

$$\begin{aligned}
\mathrm{var}(\hat{a}_{\rm LS}) &= \frac{\sigma^2 \sum x_i^2}{n \sum x_i^2 \; - \; (\sum x_i)^2} \\
\mathrm{var}(\hat{b}_{\rm LS}) &= \frac{\sigma^2 \, n}{n \sum x_i^2 \; - \; (\sum x_i)^2}
\end{aligned}$$

We can use these formulae to assign an error (i.e. by taking the square root of the variance) to the least squares fitted slope and intercept. In general, $\hat{a}_{\rm LS}$ and $\hat{b}_{\rm LS}$ will not be statistically independent. This means that they have non-zero **covariance**. (Recall that we defined in Section 1.9 the covariance of two random variables, X and Y, as $\mathrm{cov}(X, Y) = E[(X - \overline{x})(Y - \overline{y})]$, and it follows that $\mathrm{cov}(X, Y) = 0$ if X and Y are independent). In fact,

$$\mathrm{cov}(\hat{a}_{\rm LS}, \hat{b}_{\rm LS}) = \frac{-\sigma^2 \sum x_i}{n \sum x_i^2 \; - \; (\sum x_i)^2}$$

2.4.3 : Weighted Least Squares

A common situation met in astronomy (and indeed in all the physical sciences) is where one can model the relationship between bivariate data as a straight line, but it is **not** reasonable to assume that the residuals are all drawn from the same pdf. In particular, it is often the case that the residuals each have a different variance. For example, in the case of the Cepheid PL relation, shorter period Cepheids are – on average – less luminous, which could mean that the uncertainty on the measured apparent magnitude would be larger than that for longer period Cepheids. Equally, it could be the case that the *intrinsic scatter* (as opposed to the scatter due to observational errors) about the assumed straight line relation is a function of the independent variable; this situation has recently been suggested for the Tully-Fisher relation, which is a straight line relationship between the absolute magnitude (dependent variable) and log rotation velocity (independent variable) for spiral galaxies. Thus, in such cases, the i^{th} residual, $\{\epsilon_i\}$, is assumed to be drawn from some underlying pdf with mean zero and variance σ_i^2, where the variance is allowed to be different for each residual.

If the residuals are **not** identically distributed, this will affect the best-fit straight line relation derived for a given set of data. One must 'weight' the least squares solution to take account of the different variance on each residual, since the residuals with large variance should have less influence on determining the best-fit parameters. We call such a procedure **weighted least squares**. We can find weighted least squares estimators of a and b in a similar fashion to that for ordinary least squares, but with a modified sum of squares function, S, given by

$$S = \sum_{i=1}^{n} \left[\frac{y_i - (a + bx_i)}{\sigma_i} \right]^2$$

which yields the solution

$$\begin{aligned} \hat{a}_{\mathrm{WLS}} &= \frac{\sum \frac{y_i}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \sum \frac{y_i x_i}{\sigma_i^2} \sum \frac{x_i}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \left(\sum \frac{x_i}{\sigma_i^2} \right)^2} \\ \hat{b}_{\mathrm{WLS}} &= \frac{\sum \frac{1}{\sigma_i^2} \sum \frac{y_i x_i}{\sigma_i^2} - \sum \frac{y_i}{\sigma_i^2} \sum \frac{x_i}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \left(\sum \frac{x_i}{\sigma_i^2} \right)^2} \end{aligned} \tag{2}$$

where as before all summations are for $i = 1, ..., n$. The variance of $\hat{a}_{\mathrm{WLS}}$ and $\hat{b}_{\mathrm{WLS}}$ is given by

$$\begin{aligned} \mathrm{var}(\hat{a}_{\mathrm{WLS}}) &= \frac{\sum \frac{x_i^2}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \left(\sum \frac{x_i}{\sigma_i^2} \right)^2} \\ \mathrm{var}(\hat{b}_{\mathrm{WLS}}) &= \frac{\sum \frac{1}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \left(\sum \frac{x_i}{\sigma_i^2} \right)^2} \end{aligned} \tag{3}$$

In the case where σ_i^2 is constant, for all i, these formulae reduce to those given in Section 2.4.2 for the unweighted case.

2.4.4 : Least Squares and Linear Regression

In the case of a bivariate normal distribution we saw in section 1.9 that the conditional distribution of Y given x, denoted $p(y|x)$, was a normal distribution with mean value which was a linear function of x. In other words if we consider the **conditional expectation value** of Y given x, denoted by $E(Y|x)$, as we vary x, this conditional expectation defines a straight line in the $\{x, y\}$ plane. We call this straight line the **linear regression** or **regression line** of Y on X. It can be shown that this regression line is identical to the best-fit straight line obtained by an ordinary least squares fit to the $\{x_i, y_i\}$ data, so that in this sense least squares and linear regression are equivalent. In fact, this equivalence holds not only for a bivariate normal distribution, but any bivariate distribution for which the conditional expectation of Y given x is a linear function of x.

2.4.5 : Extending Ordinary Least Squares

The simple formulation of ordinary least squares considered in this course can be extended in several different ways. For example, one can express the dependent variable as a linear function of two or more independent variables (e.g. for the Cepheid PL relation we can include a term which depends on the colour of a Cepheid; we call this the PLC relation). This extension is known as **multilinear least squares** or **multilinear regression** and can be formulated quite neatly – and completely generally – in terms of vectors and matrices. We do not consider multilinear least squares in this course, however.

One can also modify the assumptions of ordinary least squares by accounting for errors, or residuals, on *both* variables (e.g. for the Cepheid PL relation one could allow for an uncertainty on the measured period). This means that one has to modify the form of the sum of squares function S, which has to be minimised with respect to the unknown parameters of the best-fit straight line. The details of this generalisation to errors on both variables are quite straightforward in principle, but are algebraically rather messy and we do not attempt them here.