# 3. Model Fitting

In this section we apply the statistical tools introduced in Section 2 to explore:

➢ how to estimate model parameters
➢ how to test the goodness of fit of models.

We will consider:

# 3. Model Fitting

In this section we apply the statistical tools introduced in Section 2 to explore:

➢  how to estimate model parameters
➢  how to test the goodness of fit of models.

We will consider:

*But before we do,  we first introduce an important pdf:*
*the  **bivariate normal distribution***

# 3.1 The bivariate normal distribution

Let $x$ and $y$ be RVs with the following joint pdf

$$p(x,y) \;\; = \;\; \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\left[-\frac{1}{2(1-\rho^2)}Q(x,y)\right] \qquad (3.1)$$

where the quadratic form, $Q(x,y)$ is given by

$$Q(x,y) \;\; = \;\; (\frac{x-\mu_x}{\sigma_x})^2 - 2\rho(\frac{x-\mu_x}{\sigma_x})(\frac{y-\mu_y}{\sigma_y}) + (\frac{y-\mu_y}{\sigma_y})^2 \qquad (3.2)$$

# 3.1 The bivariate normal distribution

Let $x$ and $y$ be RVs with the following joint pdf

$$p(x, y) \quad = \quad \frac{1}{2\pi \sigma_{\mathrm{x}} \sigma_{\mathrm{y}} \sqrt{1 - \rho^2}} \exp\left[ -\frac{1}{2(1 - \rho^2)} Q(x, y) \right] \qquad (3.1)$$

where the quadratic form, $Q(x, y)$ is given by

$$Q(x, y) \quad = \quad \left(\frac{x - \mu_{\mathrm{x}}}{\sigma_{\mathrm{x}}}\right)^2 - 2\rho\left(\frac{x - \mu_{\mathrm{x}}}{\sigma_{\mathrm{x}}}\right)\left(\frac{y - \mu_{\mathrm{y}}}{\sigma_{\mathrm{y}}}\right) + \left(\frac{y - \mu_{\mathrm{y}}}{\sigma_{\mathrm{y}}}\right)^2 \qquad (3.2)$$

Then $p(x, y)$ is known as the **bivariate normal pdf** and is specified by the 5 parameters $\mu_{\mathrm{x}}$, $\mu_{\mathrm{y}}$, $\sigma_{\mathrm{x}}$, $\sigma_{\mathrm{y}}$ and $\rho$. This pdf is used often in the physical sciences to model the joint pdf of two random variables.
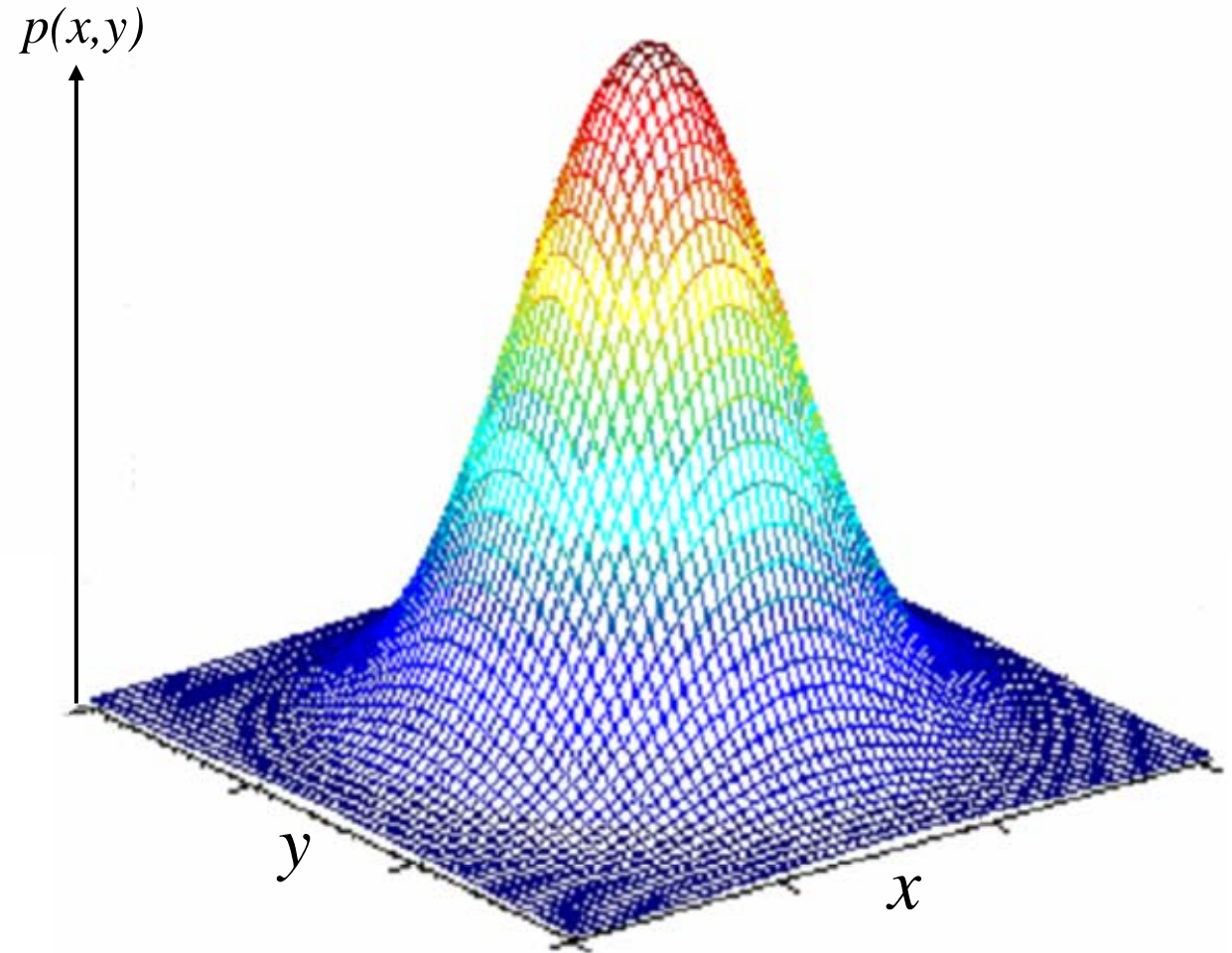
The first 4 parameters of the bivariate normal pdf are, in fact, equal to the following expectation values:-

1. $E(x) = \mu_x$

2. $E(y) = \mu_y$

3. $\text{var}(x) = \sigma_x^2$

4. $\text{var}(y) = \sigma_y^2$

$p(x,y)$

$y$

$x$

The parameter $\rho$ is known as the **correlation coefficient** and satisfies

$$E[(x - \mu_x)(y - \mu_y)] = \rho \sigma_x \sigma_y \qquad (3.3)$$

Note that if $\rho = 0$ then $x$ and $y$ are independent.

$E[(x - \mu_x)(y - \mu_y)]$ is known as the **covariance** of $x$ and $y$ and is often denoted by $\text{cov}(x, y)$.

The parameter $\rho$ is known as the **correlation coefficient** and satisfies

$$E\big[(x - \mu_x)(y - \mu_y)\big] \quad = \quad \rho\sigma_x\sigma_y \qquad (3.3)$$

Note that if $\rho = 0$ then $x$ and $y$ are independent.

$E\big[(x - \mu_x)(y - \mu_y)\big]$ is known as the **covariance** of $x$ and $y$ and is often denoted by $\mathrm{cov}(x, y)$.

In fact, for *any* two variables $x$ and $y$, we define

$$\mathrm{cov}(x, y) \quad = \quad E\big[(x - E(x))(y - E(y))\big] \qquad (3.4)$$

# Isoprobability contours for the bivariate normal pdf

$\rho > 0$ :    positive correlation

$y$ tends to increase as $x$ increases

$\rho < 0$ :    negative correlation

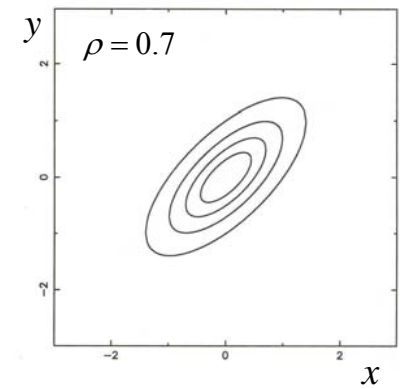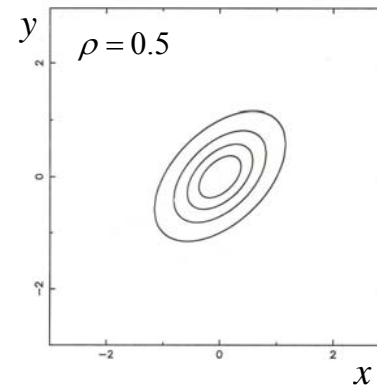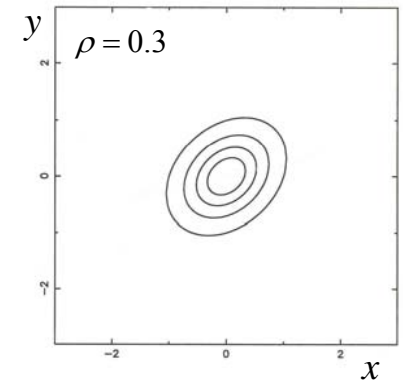$y$ tends to decrease as $x$ increases
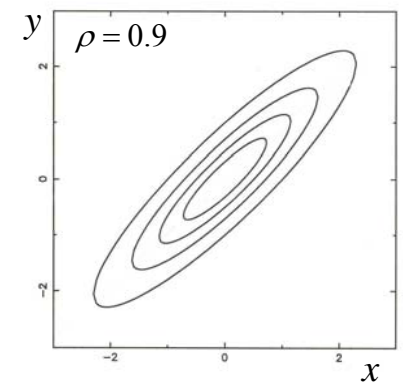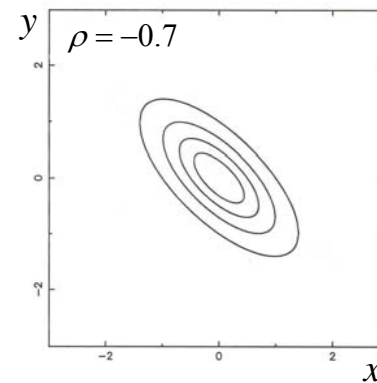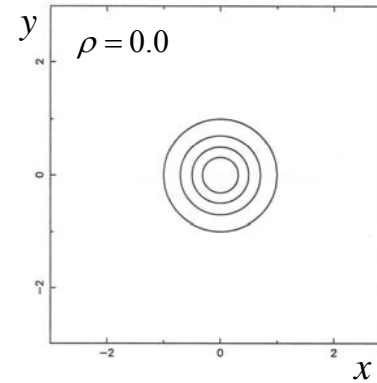
# Isoprobability contours for the bivariate normal pdf

$\rho > 0$ :   positive correlation

$y$ tends to increase as $x$ increases

$\rho < 0$ :   negative correlation

$y$ tends to decrease as $x$ increases

Contours become narrower and steeper as $|\rho| \to 1$

$\Rightarrow$   stronger (anti) correlation between $x$ and $y$.

  i.e.   *Given* value of $x$, value of $y$ is tightly constrained.



$\rho = 0.0$   $\rho = 0.3$   $\rho = 0.5$   $\rho = 0.7$   $\rho = -0.7$   $\rho = 0.9$
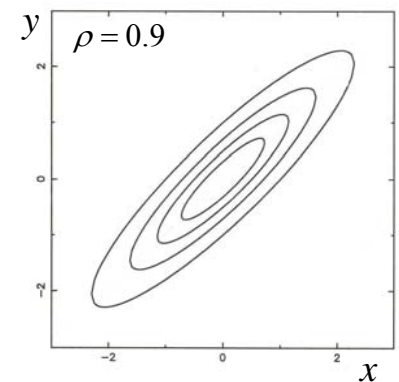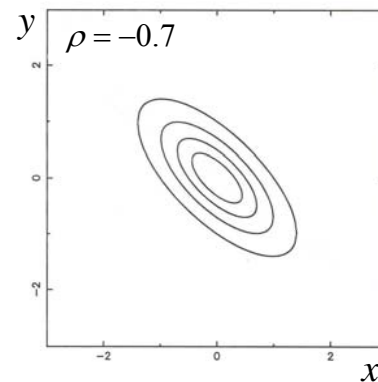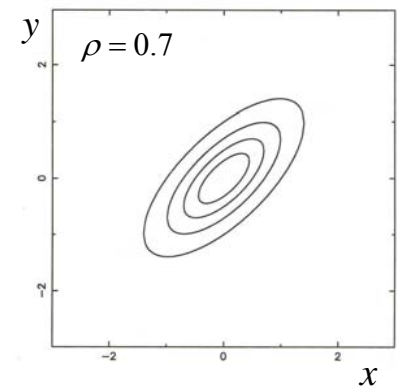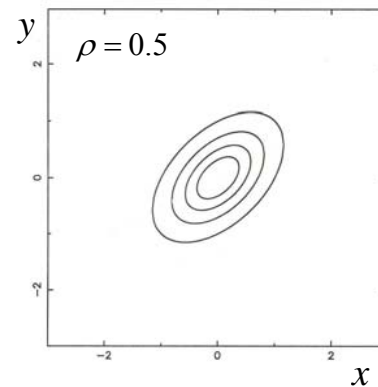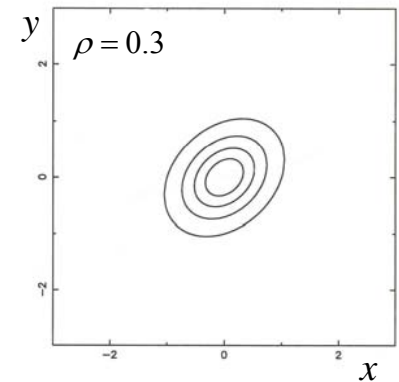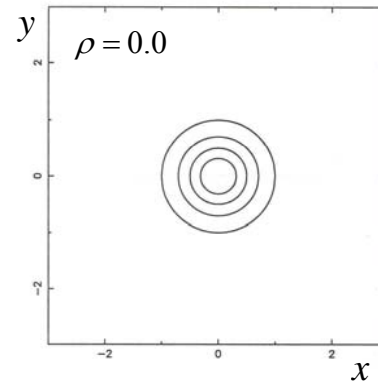
# 3.2 The method of Least Squares

o   'workhorse' method for fitting lines and curves to data in the physical sciences

o   method often encountered (as a 'black box'?) in elementary courses

o   useful demonstration of underlying statistical principles

o   simple illustration of fitting straight line to  $(x, y)$ data

# Ordinary Linear Least Squares

Suppose that the scatter in a plot of $\{x_i, y_i\}$ is assumed to arise from errors in only one of the two variables. This case is called **Ordinary Least Squares**. We then call $x$ the **independent variable**, and $y$ the **dependent variable**. Thus we suppose that we can write, for each data point:-

$$y_i \quad = \quad a \; + \; bx_i + \epsilon_i \tag{3.8}$$

where $\epsilon_i$ is known as the **residual** of the $i^{th}$ data point – i.e. the difference between the observed value of $y_i$, and the value predicted by the best-fit straight line, characterised by parameters $a$ and $b$.

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

$$S = \sum_{i=1}^{n} \varepsilon_i^2$$

The **least squares estimators** of $a$ and $b$ minimise

$$S \quad = \quad \chi^2(a,b) \quad = \quad \sum_{i=1}^{n} [y_i - (a + bx_i)]^2 \qquad (3.8)$$

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

$$\frac{\partial S}{\partial a} = 0 \quad \text{when} \quad a = \hat{a}_{\mathrm{LS}} \qquad \frac{\partial S}{\partial b} = 0 \quad \text{when} \quad b = \hat{b}_{\mathrm{LS}} \qquad (3.9)$$

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

$$\hat{a}_{\mathrm{LS}} = \frac{\sum y_i \sum x_i^2 - \sum y_i x_i \sum x_i}{n \sum x_i^2 - \left(\sum x_i\right)^2} \tag{3.10}$$

$$\hat{b}_{\mathrm{LS}} = \frac{n \sum y_i x_i - \sum y_i \sum x_i}{n \sum x_i^2 - \left(\sum x_i\right)^2} \tag{3.11}$$

where $n$ denotes the sample size and all sums are for $i = 1, ..., n$.

We can show that

$$E\left(\hat{a}_{LS}\right) = a_{LS}$$

i.e. LS estimators are **unbiased**.

$$E\left(\hat{b}_{LS}\right) = b_{LS} \qquad (3.12)$$

Also

$$\mathrm{var}(\hat{a}_{\mathrm{LS}}) \quad = \quad \frac{\sigma^2 \sum x_i^2}{n \sum x_i^2 \; - \; (\sum x_i)^2} \qquad (3.13)$$

$$\mathrm{var}(\hat{b}_{\mathrm{LS}}) \quad = \quad \frac{\sigma^2 \, n}{n \sum x_i^2 \; - \; (\sum x_i)^2} \qquad (3.14)$$

and

$$\mathrm{cov}(\hat{a}_{\mathrm{LS}}, \hat{b}_{\mathrm{LS}}) \quad = \quad \frac{-\sigma^2 \sum x_i}{n \sum x_i^2 \; - \; (\sum x_i)^2} \qquad (3.15)$$

We can show that
$$E\left(\hat{a}_{LS}\right) = a_{LS}$$
i.e. LS estimators are **unbiased**.
$$E\left(\hat{b}_{LS}\right) = b_{LS} \qquad (3.12)$$

Also

$$\text{var}(\hat{a}_{LS}) = \frac{\sigma^2 \sum x_i^2}{n \sum x_i^2 - \left(\sum x_i\right)^2} \qquad (3.13)$$

$$\text{var}(\hat{b}_{LS}) = \frac{\sigma^2 n}{n \sum x_i^2 - \left(\sum x_i\right)^2} \qquad (3.14)$$

and

$$\text{cov}(\hat{a}_{LS}, \hat{b}_{LS}) = \frac{-\sigma^2 \sum x_i}{n \sum x_i^2 - \left(\sum x_i\right)^2} \qquad (3.15)$$

Choosing the $\{x_i\}$ so that $\sum x_i = 0$ we can make $\hat{a}_{LS}$ and $\hat{b}_{LS}$ independent.

# Weighted Linear Least Squares

Suppose the $i^{th}$ residual, $\{\epsilon_i\}$, is assumed to be drawn from some underlying pdf with mean zero and variance $\sigma_i^2$, where the variance is allowed to be different for each residual.  ( Common in astronomy )

# Hubble diagram of distant Type Ia supernovae
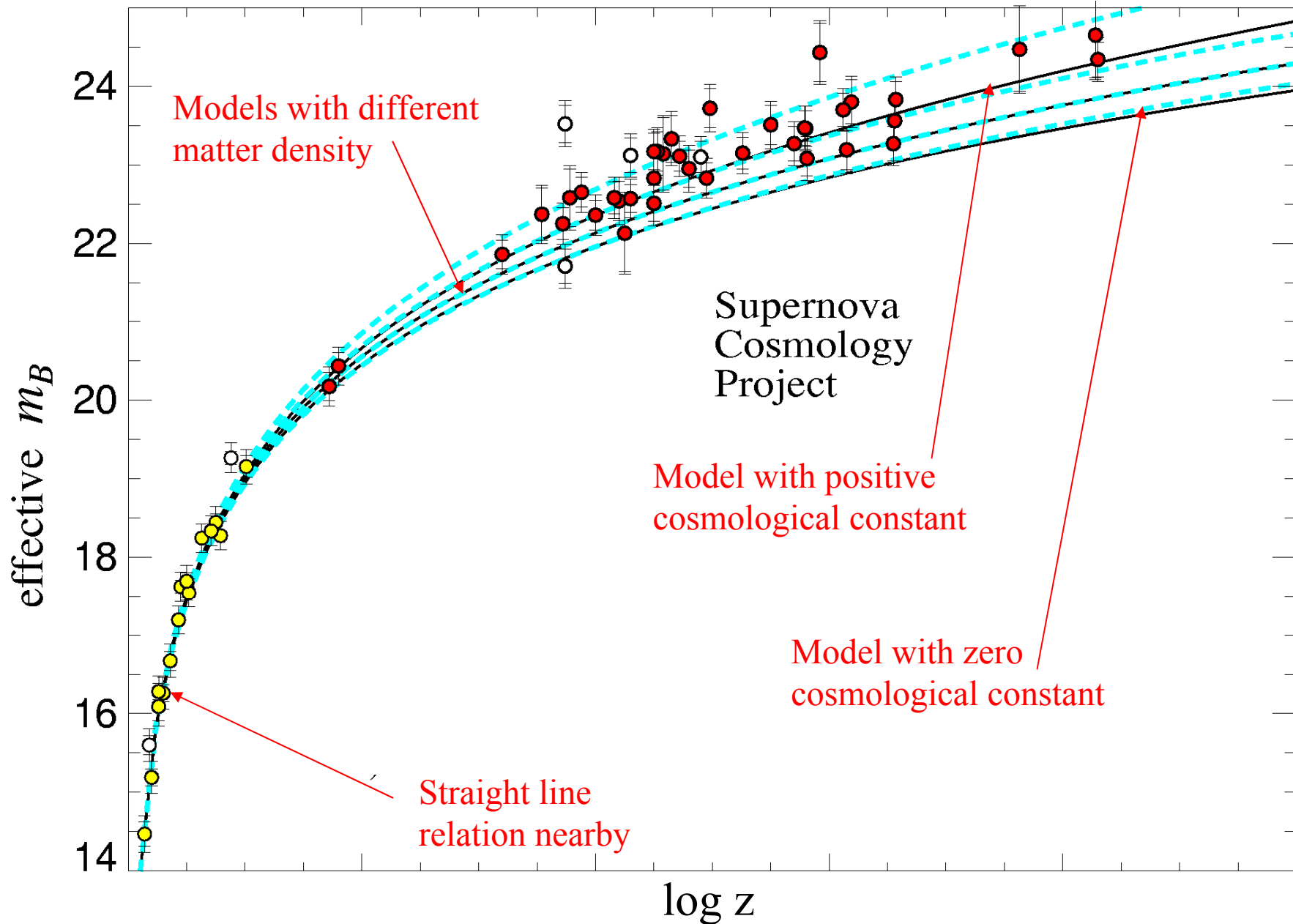


Models with different matter density

Supernova Cosmology Project

Model with positive cosmological constant

Model with zero cosmological constant

Straight line relation nearby

effective $m_B$

log z

# Weighted Linear Least Squares

Suppose the $i^{th}$ residual, $\{\epsilon_i\}$, is assumed to be drawn from some underlying pdf with mean zero and variance $\sigma_i^2$, where the variance is allowed to be different for each residual. **( Common in astronomy )**

Define

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

(3.16)

Again we find Least Squares estimators of $a$ and $b$ satisfying

$$\frac{\partial S}{\partial a} = 0 \qquad \frac{\partial S}{\partial b} = 0$$

Solving, we find

$$\hat{a}_{\mathrm{WLS}} = \frac{\sum \frac{y_i}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \sum \frac{y_i x_i}{\sigma_i^2} \sum \frac{x_i}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \left(\sum \frac{x_i}{\sigma_i^2}\right)^2} \tag{3.17}$$

$$\hat{b}_{\mathrm{WLS}} = \frac{\sum \frac{1}{\sigma_i^2} \sum \frac{y_i x_i}{\sigma_i^2} - \sum \frac{y_i}{\sigma_i^2} \sum \frac{x_i}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \left(\sum \frac{x_i}{\sigma_i^2}\right)^2} \tag{3.18}$$

Also

$$\operatorname{var}(\hat{a}_{\mathrm{WLS}}) = \frac{\sum \frac{x_i^2}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \left(\sum \frac{x_i}{\sigma_i^2}\right)^2} \tag{3.19}$$

$$\operatorname{var}(\hat{b}_{\mathrm{WLS}}) = \frac{\sum \frac{1}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \left(\sum \frac{x_i}{\sigma_i^2}\right)^2} \tag{3.20}$$

$$\operatorname{cov}(\hat{a}_{\mathrm{WLS}}, \hat{b}_{\mathrm{WLS}}) = \frac{-\sum \frac{x_i}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \left(\sum \frac{x_i}{\sigma_i^2}\right)^2} \tag{3.21}$$

Also

$$\text{var}(\hat{a}_{\text{WLS}}) = \frac{\sum \frac{x_i^2}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \left(\sum \frac{x_i}{\sigma_i^2}\right)^2} \tag{3.19}$$

$$\text{var}(\hat{b}_{\text{WLS}}) = \frac{\sum \frac{1}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \left(\sum \frac{x_i}{\sigma_i^2}\right)^2} \tag{3.20}$$

$$\text{cov}(\hat{a}_{\text{WLS}}, \hat{b}_{\text{WLS}}) = \frac{-\sum \frac{x_i}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \left(\sum \frac{x_i}{\sigma_i^2}\right)^2} \tag{3.21}$$

In the case where $\sigma_i^2$ is constant, for all $i$, these formulae reduce to those for the unweighted case.

# Extensions and Generalisations

o  Errors on *both* variables?

Need to modify merit function accordingly.

$$\chi^2(a,b) = \sum_{i=1}^{N} \frac{(y_i - a - bx_i)^2}{\sigma_{y\,i}^2 + b^2 \sigma_{x\,i}^2}$$

(3.22)

# Extensions and Generalisations

o Errors on *both* variables?

Need to modify merit function accordingly.

$$\chi^2(a, b) = \sum_{i=1}^{N} \frac{(y_i - a - bx_i)^2}{\sigma_{y\,i}^2 + b^2 \sigma_{x\,i}^2}$$

(3.22)

Renders equations *non-linear*; no simple analytic solution!

Not examinable, but see e.g.
Numerical Recipes 15.3

# Extensions and Generalisations

o   General linear models?

e.g.
$$y(x) = a_1 + a_2 x + a_3 x^2 + \cdots + a_M x^{M-1}$$ (3.23)

We can write
$$\chi^2 = \sum_{i=1}^{N} \left[ \frac{y_i - \sum_{k=1}^{M} a_k X_k(x_i)}{\sigma_i} \right]^2$$ (3.24)

Can formulate as a matrix equation and solve for parameters

# Matrix approach to Least Squares

Define
$$\boldsymbol{a} = \begin{bmatrix} a_1 \\ \vdots \\ a_M \end{bmatrix}$$

Vector of model parameters

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

Vector of observations

$$\boldsymbol{X} = \begin{bmatrix} X_1(x_1) \cdots X_M(x_1) \\ \vdots \qquad \vdots \\ X_1(x_N) \cdots X_M(x_N) \end{bmatrix}$$

Design matrix of model basis functions

**Model:**

Vector of model parameters

$$y = Xa + \varepsilon \qquad (3.25)$$

Vector of errors

Vector of observations

Design matrix of model basis functions

$$\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{bmatrix}$$

where we assume $\varepsilon_i$ is drawn from some pdf with mean zero and variance $\sigma^2$

# Matrix approach to Least Squares: weighting by errors

Define 
$$\mathbf{a} = \begin{bmatrix} a_1 \\ \vdots \\ a_M \end{bmatrix}$$

Vector of model parameters

$$\mathbf{b} = \begin{bmatrix} y_1/\sigma_1 \\ \vdots \\ y_N/\sigma_N \end{bmatrix}$$

Vector of weighted observations

$$\mathbf{A} = \begin{bmatrix} \dfrac{X_1(x_1)}{\sigma_1} & \cdots & \dfrac{X_M(x_1)}{\sigma_1} \\ \vdots & & \vdots \\ \dfrac{X_1(x_N)}{\sigma_N} & \cdots & \dfrac{X_M(x_N)}{\sigma_N} \end{bmatrix}$$

Weighted design matrix of model basis functions

**Weighted Model:**

Vector of model parameters

$$b = Aa + e$$

(3.26)

Vector of weighted errors

Vector of weighted observations

Weighted design matrix of model basis functions

$$e = \begin{bmatrix} \varepsilon_1 / \sigma_1 \\ \vdots \\ \varepsilon_N / \sigma_N \end{bmatrix}$$

where we assume $\varepsilon_i$ is drawn from some pdf with mean zero and variance $\sigma_i^2$

We solve for the parameter vector $\hat{\boldsymbol{a}}_{LS}$ that minimises

$$S = \boldsymbol{e}^T \cdot \boldsymbol{e} = \sum_{i=1}^{n} e_i^2$$

This has solution

$$\hat{\boldsymbol{a}}_{LS} = \left( A^T A \right)^{-1} A^T \cdot \boldsymbol{b} \qquad (3.27)$$

$M \times M$ *matrix*

and

$$\mathrm{cov}\left( \hat{\boldsymbol{a}}_{LS} \right) = \left( A^T A \right)^{-1} \qquad (3.28)$$

# Extensions and Generalisations

o   Non-linear models?     $y_i^{\text{model}} \quad \equiv \quad y^{\text{model}}(x_i; \theta_1, ..., \theta_k)$

Model parameters

Suppose     $y_i^{\text{obs}} \quad = \quad y_i^{\text{model}} + \epsilon_i$

$\varepsilon_i$ drawn from pdf with mean zero, variance $\sigma_i^2$

Then

$$S = \quad \chi^2 \quad = \quad \sum_{i=1}^{n} \left[ \frac{y_i^{\text{obs}} - y_i^{\text{model}}}{\sigma_i} \right]^2$$

(3.29)

# Extensions and Generalisations

o   Non-linear models?

$$y_i^{\text{model}} \equiv y^{\text{model}}(x_i; \theta_1, ..., \theta_k)$$

<span style="color:red">Model parameters</span>

Suppose

$$y_i^{\text{obs}} = y_i^{\text{model}} + \epsilon_i$$

$\varepsilon_i$ drawn from pdf with mean zero, variance $\sigma_i^2$

Then

$$S = \chi^2 = \sum_{i=1}^{n} \left[ \frac{y_i^{\text{obs}} - y_i^{\text{model}}}{\sigma_i} \right]^2 \qquad (3.29)$$

But no simple analytic method to minimise sum of squares
( e.g. no analytic solutions to $\partial S / \partial \theta_i = 0$ )

## 3.3    The principle of maximum likelihood

**Frequentist approach**:

*A parameter is a fixed (but unknown) constant*

From actual data we can compute Likelihood,

$L =$ probability of obtaining the observed data, given the value of the parameter $\theta$

# 3.3   The principle of maximum likelihood

**Frequentist approach**:

*A parameter is a fixed (but unknown) constant*

From actual data we can compute Likelihood,
$L =$ probability of obtaining the observed data, given the value of the parameter $\theta$

Now define **likelihood function**: (infinite) family of curves generated by regarding $L$ as a function of $\theta$ , for data fixed.

**Principle of Maximum Likelihood**

A good estimator of $\theta$  maximises $L$  -

i.e.   $\dfrac{\partial L}{\partial \theta} = 0$   and   $\dfrac{\partial^2 L}{\partial \theta^2} < 0$

## 3.3    The principle of maximum likelihood

**Frequentist approach**:

*A parameter is a fixed (but unknown) constant*

From actual data we can compute Likelihood,

$L =$ probability of obtaining the observed data, given the value of the parameter $\theta$

Now define **likelihood function**:  (infinite) family of curves generated by regarding $L$ as a function of $\theta$ , for data fixed.

**Principle of Maximum Likelihood**

A good estimator of $\theta$ maximises $L$ -

i.e.    $\dfrac{\partial L}{\partial \theta} = 0$   and   $\dfrac{\partial^2 L}{\partial \theta^2} < 0$

We set the parameter equal to the value that makes the actual data sample we *did* observe – out of all the possible random samples we *could have* observed – the most likely.

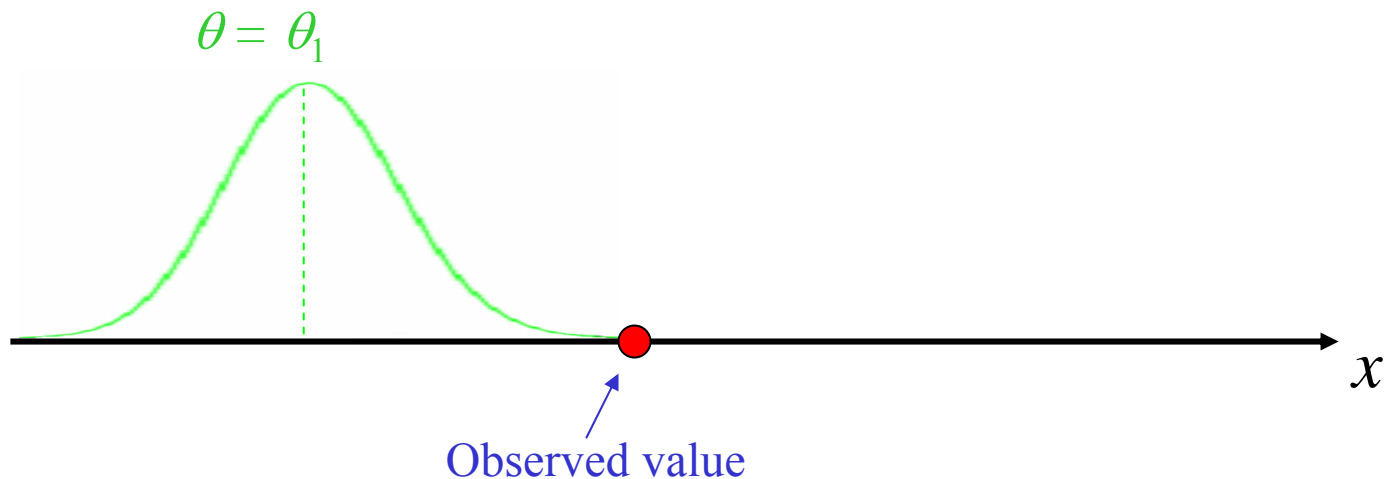Likelihood function has same definition in Bayesian probability theory, but subtle difference in meaning and interpretation – no need to invoke idea of (infinite) ensemble of different samples.

## Principle of Maximum Likelihood

A good estimator of $\theta$ maximises $L$ -

i.e. $\dfrac{\partial L}{\partial \theta} = 0$ and $\dfrac{\partial^2 L}{\partial \theta^2} < 0$

$\theta = \theta_1$



Observed value

## Principle of Maximum Likelihood

A good estimator of $\theta$ maximises $L$ -

i.e. $\dfrac{\partial L}{\partial \theta} = 0$ and $\dfrac{\partial^2 L}{\partial \theta^2} < 0$

$\theta = \theta_2$

$x$

Observed value

## Principle of Maximum Likelihood

A good estimator of $\theta$ maximises $L$ -

i.e. $\dfrac{\partial L}{\partial \theta} = 0$   and   $\dfrac{\partial^2 L}{\partial \theta^2} < 0$

$\theta = \theta_3$

$x$

Observed value

## Principle of Maximum Likelihood

A good estimator of $\theta$ maximises $L$ -

i.e. $\dfrac{\partial L}{\partial \theta} = 0$ and $\dfrac{\partial^2 L}{\partial \theta^2} < 0$

$\theta = \hat{\theta}_{\mathrm{ML}}$

$x$

Observed value

# 3.4   Least squares as maximum likelihood estimators

To see the maximum likelihood method in action,  let's consider again
weighted least squares  for the simple model   $y_i \quad = \quad a + bx_i + \epsilon_i$

*(From Section 3.3)*

Suppose the $i^{th}$ residual, $\{\epsilon_i\}$, is assumed to be drawn from some underlying pdf with mean zero and variance $\sigma_i^2$, where the variance is allowed to be different for each residual.

Let's assume the pdf is a Gaussian

## 3.4   Least squares as maximum likelihood estimators

To see the maximum likelihood method in action,  let's consider again
weighted least squares  for the simple model   $y_i \quad = \quad a + bx_i + \epsilon_i$

*(From Section 3.3)*

Suppose the $i^{th}$ residual, $\{\epsilon_i\}$, is assumed to be drawn from some underlying pdf with mean zero and variance $\sigma_i^2$, where the variance is allowed to be different for each residual.

Let's assume the pdf is a Gaussian

Likelihood     $$L \quad = \quad \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[ -\frac{1}{2}\frac{\varepsilon_i^2}{\sigma_i^2} \right] \qquad (3.30)$$

# 3.4 Least squares as maximum likelihood estimators

To see the maximum likelihood method in action, let's consider again
weighted least squares for the simple model $y_i = a + bx_i + \epsilon_i$

*(From Section 3.3)*

Suppose the $i^{th}$ residual, $\{\epsilon_i\}$, is assumed to be drawn from some underlying pdf with mean zero and variance $\sigma_i^2$, where the variance is allowed to be different for each residual.

Let's assume the pdf is a Gaussian

Likelihood $\qquad L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[ -\frac{1}{2}\frac{\varepsilon_i^2}{\sigma_i^2} \right]$ $\qquad$ (3.30)

*(note: $L$ is a product of 1-D Gaussians because we are assuming the $\varepsilon_i$ are independent)*

Substitute $\varepsilon_i = y_i - a - bx_i$

$$\Rightarrow \quad L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2}\frac{(y_i - a - bx_i)^2}{\sigma_i^2}\right] \qquad (3.31)$$

and the ML estimators of $a$ and $b$ satisfy $\partial L/\partial a = 0$ and $\partial L/\partial b = 0$

Substitute $\varepsilon_i = y_i - a - bx_i$

$$\Rightarrow \quad L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[ -\frac{1}{2} \frac{(y_i - a - bx_i)^2}{\sigma_i^2} \right] \qquad (3.31)$$

and the ML estimators of $a$ and $b$ satisfy $\partial L / \partial a = 0$ and $\partial L / \partial b = 0$

But maximising $L$ is equivalent to maximising $\ell = \ln L$

Here
$$\ell = -\frac{n}{2}\ln(2\pi) - \ln \sum_{i=1}^{n} \sigma_i - \frac{1}{2} \sum_{i=1}^{n} \left( \frac{y_i - a - bx_i}{\sigma_i} \right)^2 \qquad (3.32)$$

$$= \text{constant} - \frac{1}{2} S$$

This is exactly the same sum of squares we defined in Section 3.3

Substitute $\quad \varepsilon_i = y_i - a - bx_i$

$$\Rightarrow \quad L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[ -\frac{1}{2} \frac{(y_i - a - bx_i)^2}{\sigma_i^2} \right] \qquad (3.31)$$

and the ML estimators of $a$ and $b$ satisfy $\partial L/\partial a = 0$ and $\partial L/\partial b = 0$

But maximising $L$ is equivalent to maximising $\ell = \ln L$

Here $\quad \ell = -\frac{n}{2}\ln(2\pi) - \ln\sum_{i=1}^{n}\sigma_i - \frac{1}{2}\sum_{i=1}^{n}\left( \frac{y_i - a - bx_i}{\sigma_i} \right)^2 \qquad (3.32)$

$\quad = \text{constant} - \frac{1}{2}S$ ← This is exactly the same sum of squares we defined in Section 3.3

So in this case maximising $L$ is **exactly equivalent** to minimising the sum of squares.

i.e. for Gaussian, independent errors, ML and weighted LS estimators are identical.

## 3.5  Chi-squared goodness of fit test

In the previous 3 sections we have discussed how to estimate parameters of an underlying pdf model from sample data.

We now consider the related question:

### *How good is our pdf model in the first place?*

We now illustrate the frequentist approach to this question using the chi-squared goodness of fit test, for the (very common) case where the model pdf is a Gaussian.



We take an example from Gregory (Chapter 7)

*(book focusses mainly on Bayesian probability, but is very good on frequentist approach too)*
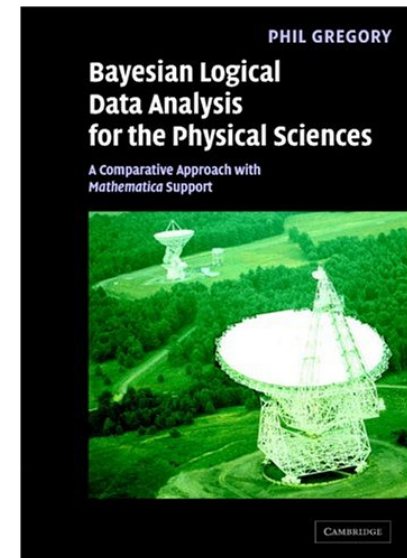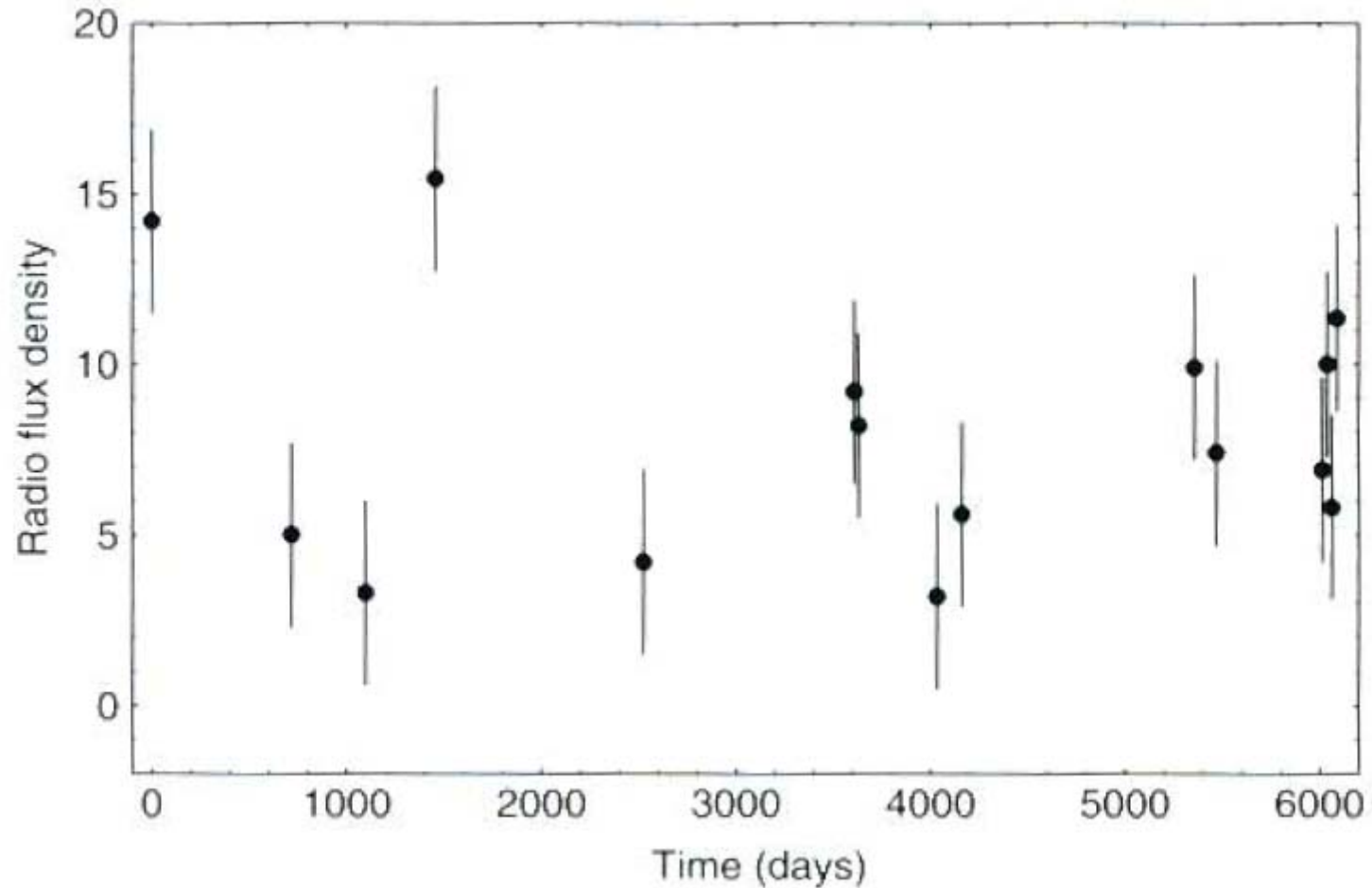
Model: radio emission from a galaxy is constant in time.

Assume residuals are iid, drawn from N(0,σ)

# Goodness-of-fit Test:   the basic ideas

1. Choose as our null hypothesis that the galaxy has an unknown but constant flux density. If we can demonstrate that this hypothesis is absurd at say the 95% confidence level, then this provides indirect evidence that the radio emission is variable. Previous experience with the measurement apparatus indicates that the measurement errors are independently normal with a $\sigma = 2.7$.

2. Select a suitable statistic that (a) can be computed from the measurements, and (b) has a predictable distribution. More precisely, (b) means that we can predict the distribution of values of the statistic that we would expect to obtain from an infinite number of repeats of the above set of radio measurements under identical conditions. We will refer to these as our hypothetical reference set. More specifically, we are predicting a probability distribution for this reference set.

   To refute the null hypothesis, we will need to show that scatter of the individual measurements about the mean is larger than would be expected from measurement errors alone.

3. Evaluate the $\chi^2$ statistic from the measured data. Let's start with the expression for the $\chi^2$ statistic for our data set:

*From Gregory, pg. 164*

$$\chi^2 = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{\sigma^2},$$   (3.33)

**The $\chi^2$ pdf**

$$p_\nu\left(\chi^2\right) = p_0 \times \left(\chi^2\right)^{\frac{\nu}{2}-1} e^{-\chi^2/2}$$

(3.34)

Here $\nu$ is known as the number of **degrees of freedom** of the pdf.

The mean value of the pdf is $\nu$ and the variance is $2\nu$.
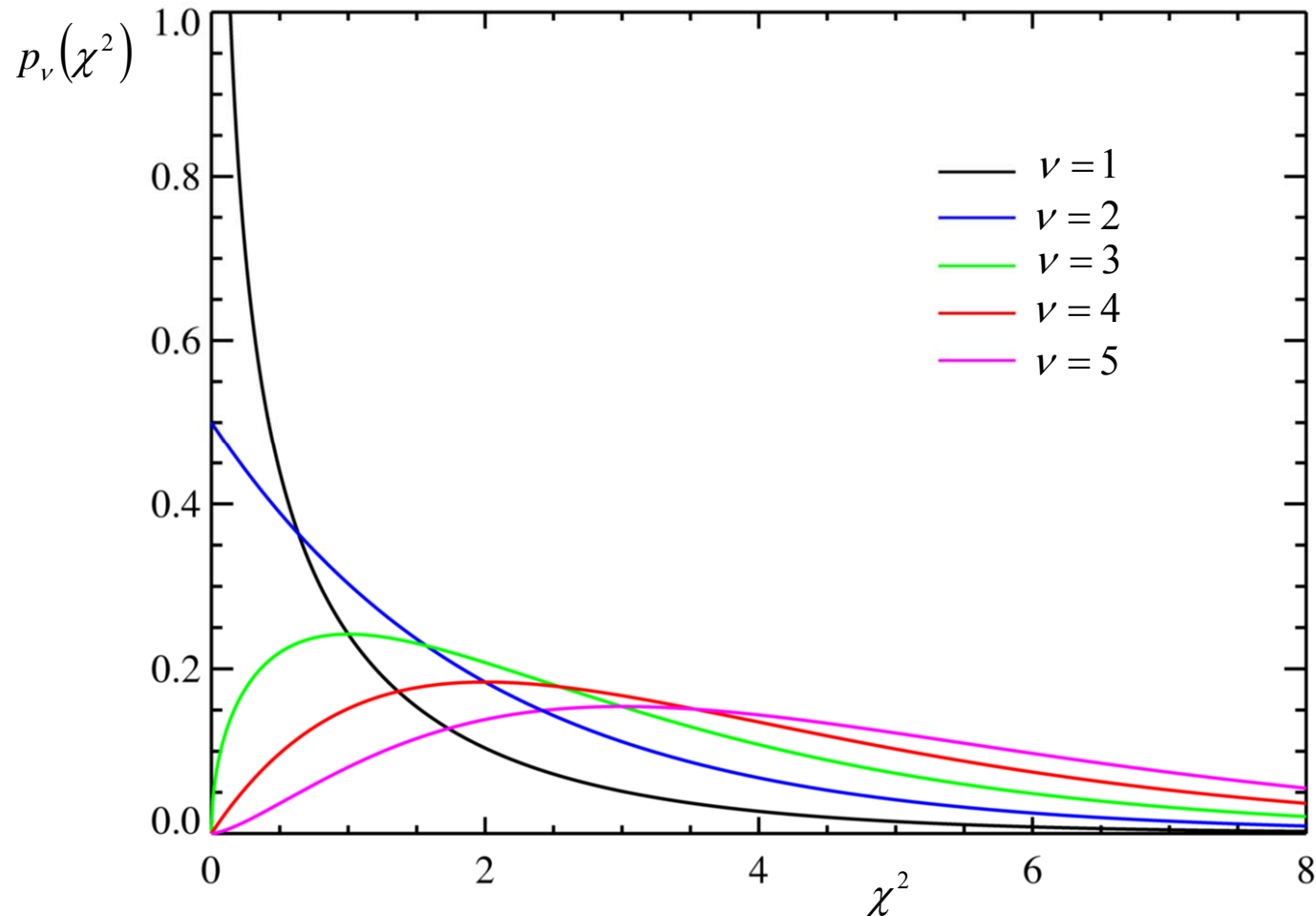
$$\chi^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(x_i - 7.98)^2}{2.7^2} = 26.76.$$

(3.35)

$n = 15$ data points, but $\nu = 14$ degrees of freedom, because $\chi^2$ statistic involves the *sample mean* and not the true mean.

We subtract one d.o.f. to account for this.

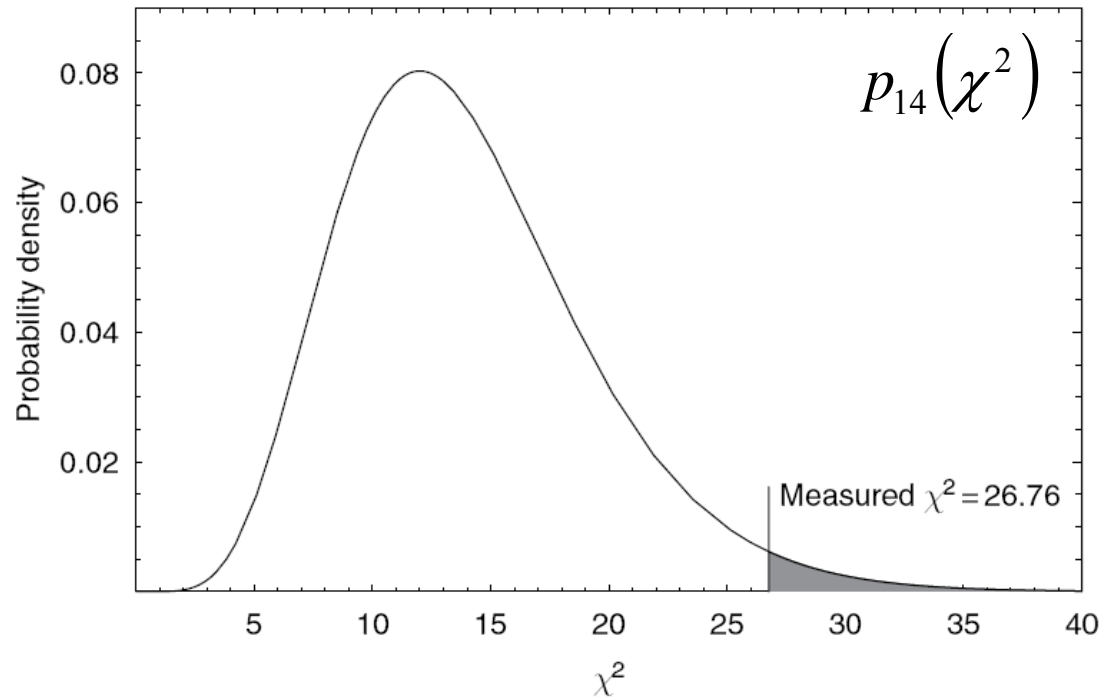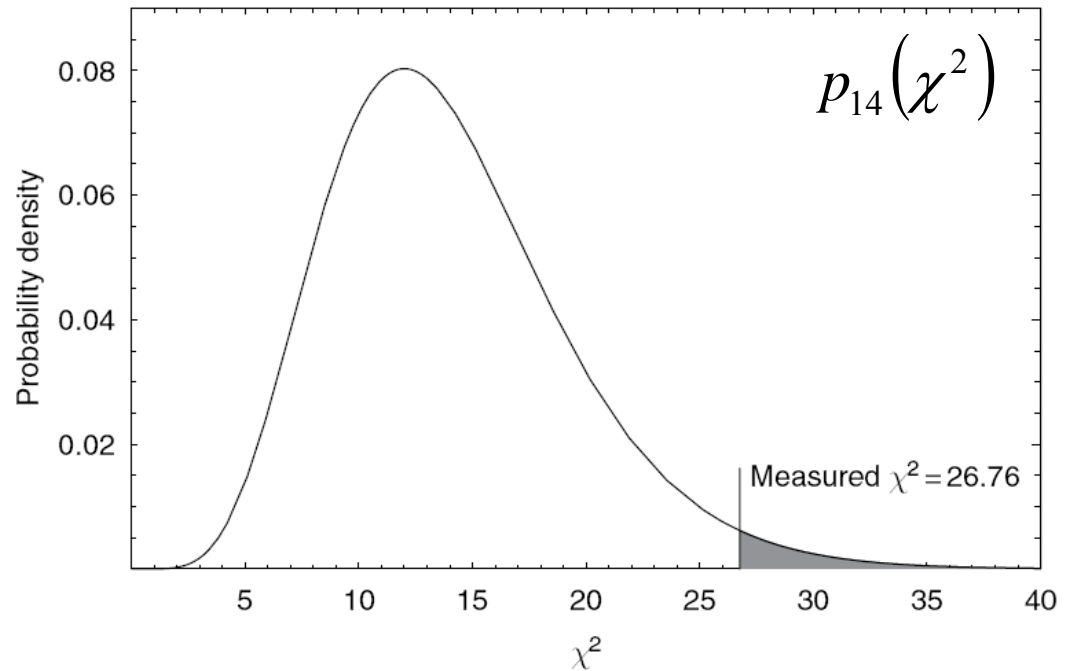| Day Number | Flux Density (mJy) |
|---|---|
| 0.0 | 14.2 |
| 718.0 | 5.0 |
| 1097.0 | 3.3 |
| 1457.1 | 15.5 |
| 2524.1 | 4.2 |
| 3607.7 | 9.2 |
| 3630.1 | 8.2 |
| 4033.1 | 3.2 |
| 4161.3 | 5.6 |
| 5355.9 | 9.9 |
| 5469.1 | 7.4 |
| 6012.4 | 6.9 |
| 6038.3 | 10.0 |
| 6063.2 | 5.8 |
| 6089.3 | 11.4 |



$p_{14}(\chi^2)$

Measured $\chi^2 = 26.76$

$$\chi^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(x_i - 7.98)^2}{2.7^2} = 26.76.$$

(3.35)

| Day Number | Flux Density (mJy) |
|---|---|
| 0.0 | 14.2 |
| 718.0 | 5.0 |
| 1097.0 | 3.3 |
| 1457.1 | 15.5 |
| 2524.1 | 4.2 |
| 3607.7 | 9.2 |
| 3630.1 | 8.2 |
| 4033.1 | 3.2 |
| 4161.3 | 5.6 |
| 5355.9 | 9.9 |
| 5469.1 | 7.4 |
| 6012.4 | 6.9 |
| 6038.3 | 10.0 |
| 6063.2 | 5.8 |
| 6089.3 | 11.4 |

$n = $ 15 data points, but $\nu = $ 14 degrees of freedom, because $\chi^2$ statistic involves the *sample mean* and not the true mean.

We subtract one d.o.f. to account for this.



$p_{14}(\chi^2)$

Measured $\chi^2 = 26.76$

If the null hypothesis is true, how probable is it that we would measure as large, or larger, a value of $\chi^2$ ?

*If* the null hypothesis were true, how probable is it that we would measure as large, or larger, a value of $\chi^2$ ?

This is an important quantity, referred to as the **P-value**

$$\text{P-value} \;=\; 1 - P\!\left(\chi^2_{\text{obs}}\right) \;=\; 1 - \int_0^{\chi^2_{\text{obs}}} p_0 \; x^{\frac{\nu}{2}-1} \exp\!\left(-\frac{x}{2}\right) dx \;=\; 0.02$$

(3.36)

*If* the null hypothesis were true, how probable is it that we would measure as large, or larger, a value of $\chi^2$?

This is an important quantity, referred to as the **P-value**

$$\text{P-value} \ = \ 1 - P\left(\chi^2_{\text{obs}}\right) \ = \ 1 - \int_0^{\chi^2_{\text{obs}}} p_0 \ x^{\frac{\nu}{2}-1} \exp\left(-\frac{x}{2}\right) dx \ = \ 0.02 \qquad (3.36)$$

What precisely does the P-value mean?

"If the galaxy flux density really *is* constant, and we repeatedly obtained sets of 15 measurements under the same conditions, then only 2% of the $\chi^2$ values derived from these sets would be expected to be greater than our one actual measured value of 26.76"

*From Gregory, pg. 165*

If we obtain a very small P-value (e.g. a few percent?) we can interpret this as providing little support for the null hypothesis, which we may then choose to reject.

*(Ultimately this choice is subjective, but $\chi^2$ may provide objective ammunition for doing so)*

Nevertheless, P-value based frequentist hypothesis testing remains very common in the literature:

| Type of problem | test | References |
|---|---|---|
| Line and curve goodness-of-fit | $\chi^2$ test | NR: 15.1-15.6 |
| Difference of means | Student's $t$ | NR: 14.2 |
| Ratio of variances | $F$ test | NR: 14.2 |
| Sample CDF | K-S test Rank sum tests | NR: 14.3, 14.6 |
| Correlated variables? | Sample correlation coefficient | NR: 14.5, 14.6 |
| Discrete RVs | $\chi^2$ test / contingency table | NR: 14.4 |

*See also supplementary handout*

# 3.7  Minimising and Maximising Functions

Least squares and maximum likelihood involve, in practice, a lot of minimising and maximising of functions – i.e. solving equations of the form:

$$\partial L/\partial \theta_i = 0 \qquad \text{(3.37)}$$

In general these equations may not have an analytic solution, especially if our pdf is a function of two or more parameters.

Some computational strategies for minimising/maximising functions:

1.  Solve $\partial \ell/\partial \theta_i = 0$ where $\ell = \ln L$ (may be easier to solve)

2.  Perform **grid search** over $\boldsymbol{\theta}$, evaluating $L(\boldsymbol{\theta})$ at each point

## 3.7  Minimising and Maximising Functions

Least squares and maximum likelihood involve, in practice, a lot of minimising and maximising of functions – i.e. solving equations of the form:

$$\partial L / \partial \theta_i = 0 \qquad \text{(3.37)}$$

In general these equations may not have an analytic solution, especially if our pdf is a function of two or more parameters.
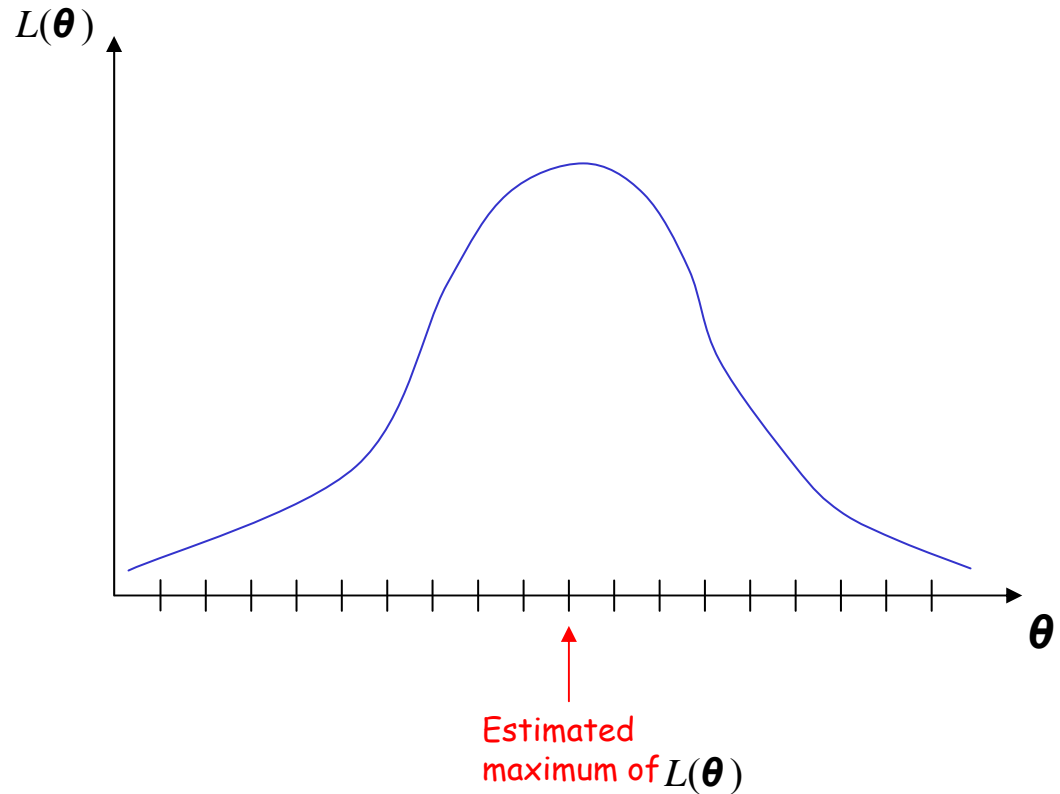
Some computational strategies for minimising/maximising functions:

1. Solve $\partial \ell / \partial \theta_i = 0$ where $\ell = \ln L$ (may be easier to solve)

2. Perform **grid search** over $\boldsymbol{\theta}$, evaluating $L(\boldsymbol{\theta})$ at each point

3. Use gradient ascent / descent for increased efficiency

2. Perform **grid search** over $\theta$, evaluating $L(\theta)$ at each point

## 1-D example

- Regularly spaced grid points.

- No need to compute derivatives of likelihood

- But we need very fine grid spacing to obtain accurate estimate of the maximum

- This is computationally **very costly**, particularly if we need to search a multi-dimensional parameter space.

$L(\theta)$

$\theta$

Estimated maximum of $L(\theta)$

2. Perform **grid search** over $\boldsymbol{\theta}$, evaluating $L(\boldsymbol{\theta})$ at each point

### 1-D example

$L(\boldsymbol{\theta})$

- Regularly spaced grid points.

- No need to compute derivatives of likelihood

- But we need very fine grid spacing to obtain accurate estimate of the maximum

- This is computationally **very costly**, particularly if we need to search a multi-dimensional parameter space.
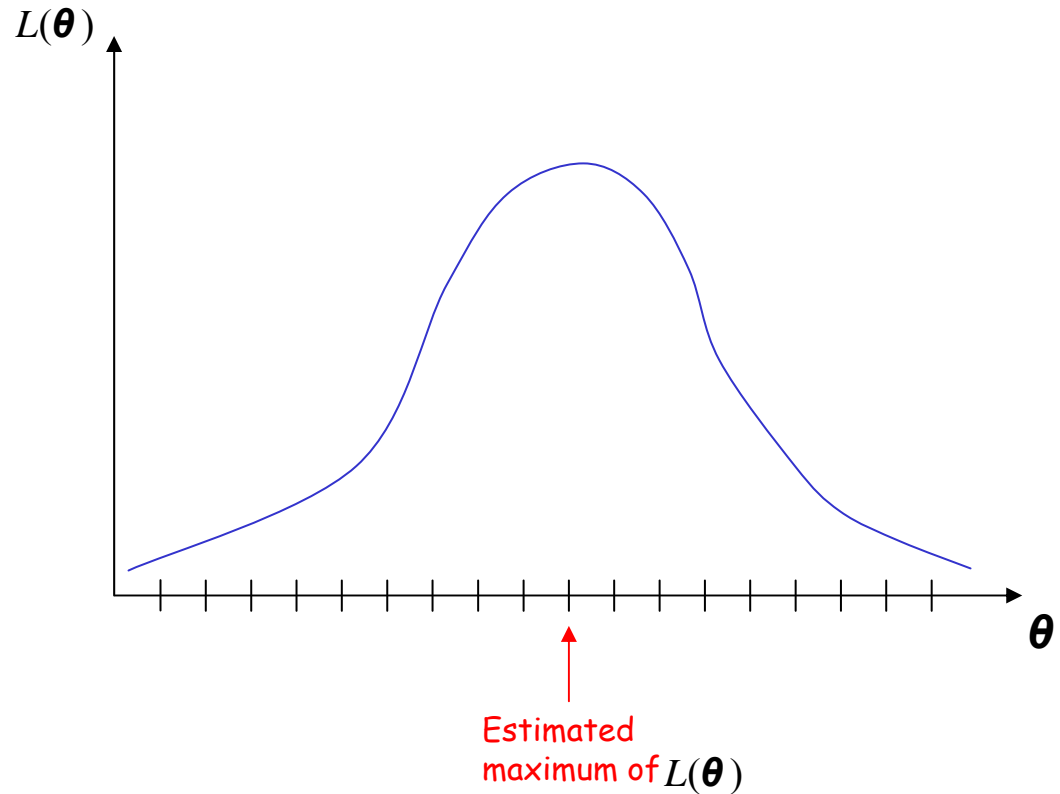
$\boldsymbol{\theta}$

Estimated maximum of $L(\boldsymbol{\theta})$

3. Method of Steepest Ascent / Descent

- Make jumps in direction where gradient of $L(\boldsymbol{\theta})$ is changing most rapidly.

- Need to estimate derivatives of likelihood, i.e. $\nabla L(\theta)$     *(See Num. Rec. Chap. 10)*