ASTR509 - 10

Data modelling
Parameter estimation

Pierre-Simon, Marquis de Laplace 1749-1827

Under Napoleon, Laplace was a member, then chancellor, of the Senate, and received the Legion of Honour in 1805. However Napoleon, in his memoirs written on St Hélène, says he removed Laplace from the office of Minister of the Interior, which he held in 1799, after only six weeks:- *“... because he brought the spirit of the infinitely small into the government.”*

Last time ....

We fought our way through some non-parametric tests for samples.

=> Despite power and versatility, these are still part of classical testing, a process of 'rejection'; they do not prove our alternative, the research hypothesis.

=> Each one requires the 4-step methodology of classical hypothesis-testing: (1) set up H_0, H_1; (2) specify a priori significance-level α we're prepared to accept and choose the test, set up the sampling distribution with its rejection area(s) totalling α; (3) compute the sampling statistic from our data, rejecting H_0 if it is a value in the rejection region; (4) carry out the terminal action.

=> We looked at the non-parametric tests for comparing single samples and for comparing two (or more) samples: chi-square test (single-sample and two to k samples), Kolmogorov-Smirnov test (single- and two-sample), Runs test for randomness, single sample, Fisher exact probability test for two small samples Wilcoxon-Mann-Whitney U test for two samples.

=> Three-table summary, to help in choice of best test for problem.

Once more - the Chi-Square Test (Pearson 1900)

A point that cannot be emphasized enough -

If we have **observational data which can be binned**, and a model/hypothesis which predicts **the population of each bin**,

Then the chi-square statistic describes the **goodness-of-fit** of the data to the model.

With the **observed** numbers in each of **k** bins as **O_i**, and the expected values from the model as **E_i**, then this statistic is

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}.$$

NB: REAL NUMBERS!!!! This is crucial. You cannot use normalized numbers or try to test one model against another model - the test depends on Poisson statistics holding good, i.e. on the 'expected' scatter being due to Poisson statistics alone.

Data Modelling; Parameter Estimation

What are we doing?

Say we have a model. For example,

if our N data Z_i follow a Gaussian distribution

$$\mathrm{prob}(z) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(z-\mu)^2}{2\sigma^2}\right],$$

then the statistic

$$m = \frac{1}{N}\sum_i Z_i$$

is a good estimator for μ and has a known distribution (a Gaussian again) which can be used for calculating confidence limits.

Or, take the Bayesian way - calculate a probability distribution for *μ*, given the data.

Any data modelling procedure is just a more elaborate version of this.

Data Modelling: what are we doing? - 2

Suppose our data ***Z_i*** were measured at various values of some independent variable ***X_i***, and we believed that they were“really” scattered, with Gaussian errors, around the underlying functional relationship, with ***(α_1,α_2, ….)*** unknown parameters (slopes, intercepts, …) of the relationship.

$$\mu = \mu(x,\alpha_1,\alpha_2, \ldots.)$$

We then have

$$\mathrm{prob}(z \mid \alpha_1, \alpha_2 \ldots) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(z-\mu(x,\alpha_1,\alpha_2\ldots))^2}{2\sigma^2}\right],$$

and, by Bayes' theorem, we have the posterior probability distribution for the parameters

$$\mathrm{prob}(\alpha_1, \alpha_2 \ldots \mid Z_i, \mu) \propto \Pi_i \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(Z_i-\mu(x,\alpha_1,\alpha_2\ldots))^2}{2\sigma^2}\right] \mathrm{prob}(\alpha_1, \alpha_2 \ldots)$$

including as usual our prior information. We have included ***μ*** as one of the “givens" to emphasize that **everything depends on it being the correct model**.

We are done! We have a probability distribution for the parameters of our model, given the data.

Data Modelling: what are we doing? - 3

Good features:

- it can be used to **update models as new data arrive**, as the posterior from one stage of the experiments becomes the prior for the next.

- we can also deal with **unwanted parameters (“nuisance parameters”).**

What are these? Typically we will end up with a probability distribution for various parameters, some of interest (say, cosmological parameters) and some not (say, instrumental calibrations). We can **‘marginalize out'** the unwanted parameters by an integration, leaving us with the distribution of the variable of interest that takes account of all plausible values of the unwanted variables.

Data Modelling: what are we doing? - 4

Bad features:

- **Modelling can be very expensive**, always involving finding a maximum or minimum of some merit function. This means evaluating the function, plus its derivatives, many times. The model itself may be the result of a complex computation; evaluating it over a multi-dimensional grid of parameters is even worse.

- **Numerical integration** may be another difficulty. Interesting problems have many parameters. Marginalizing these out, or calculating evidences for discriminating between models, involves multi-dimensional integrals - time-consuming, and hard to check.

Any analytical help we can get is especially welcome in doing integrations.

Powerful theorems may allow great simplifications.

Data Modelling: what are we doing? - 5

The most important thing to remember about models is - **they may be wrong.**

Mistaken assumptions about the **distribution of residuals** about the model represent the biggest danger.

- The **wrong parameters** will be deduced from the model.

- **Wrong errors on the parameters** will be obtained.
  It is important to have a range of models, and always to check optimized models against the data.

  Runs Test!

Analytic approximations were developed in past centuries for very good reasons:

- in the limiting case of diffuse priors, the Bayesian approach is very closely related to **maximum likelihood**;
- if the distribution of the residuals from the model is indeed Gaussian, it is closely related to **least squares**.

Maximum Likelihood - 1

The likelihood function is **the posterior probability** from Bayes' Theorem.

Suppose our data are described by the pdf ***f(x;α)***, where ***x*** is a variable, ***α*** is a parameter (maybe many parameters) characterizing the known form of ***f***.

We want *α*

If $X_1, X_2, .. X_N$ are data, presumed independent and all drawn from ***f***, then the likelihood function is

$$\begin{aligned} \mathcal{L}(x_1, x_2, ..x_N) &= f(x_1, x_2, ..x_N; \alpha) \\ &= f(x_1;\alpha) f(x_2;\alpha) ... f(x_N;\alpha) \\ &= \prod^{N} f(x_i;\alpha) \end{aligned}$$

From the **classical** point of view this is the probability, given ***α***, of obtaining the data.

From the **Bayesian** point of view it is propto ***prob(α)***, given the data and assuming that the priors are “diffuse'‘, i.e. they change little over the peaked region of the likelihood.

Maximum Likelihood - 2

The peak value of **L** seems likely to be a useful choice of the “best" estimate of ***α***.

Formally, the Maximum-Likelihood Estimator (MLE) of ***α*** is

$\underline{\alpha}$ = (that value of ***α*** which maximizes **L *(α)*** for all variations of ***α***).

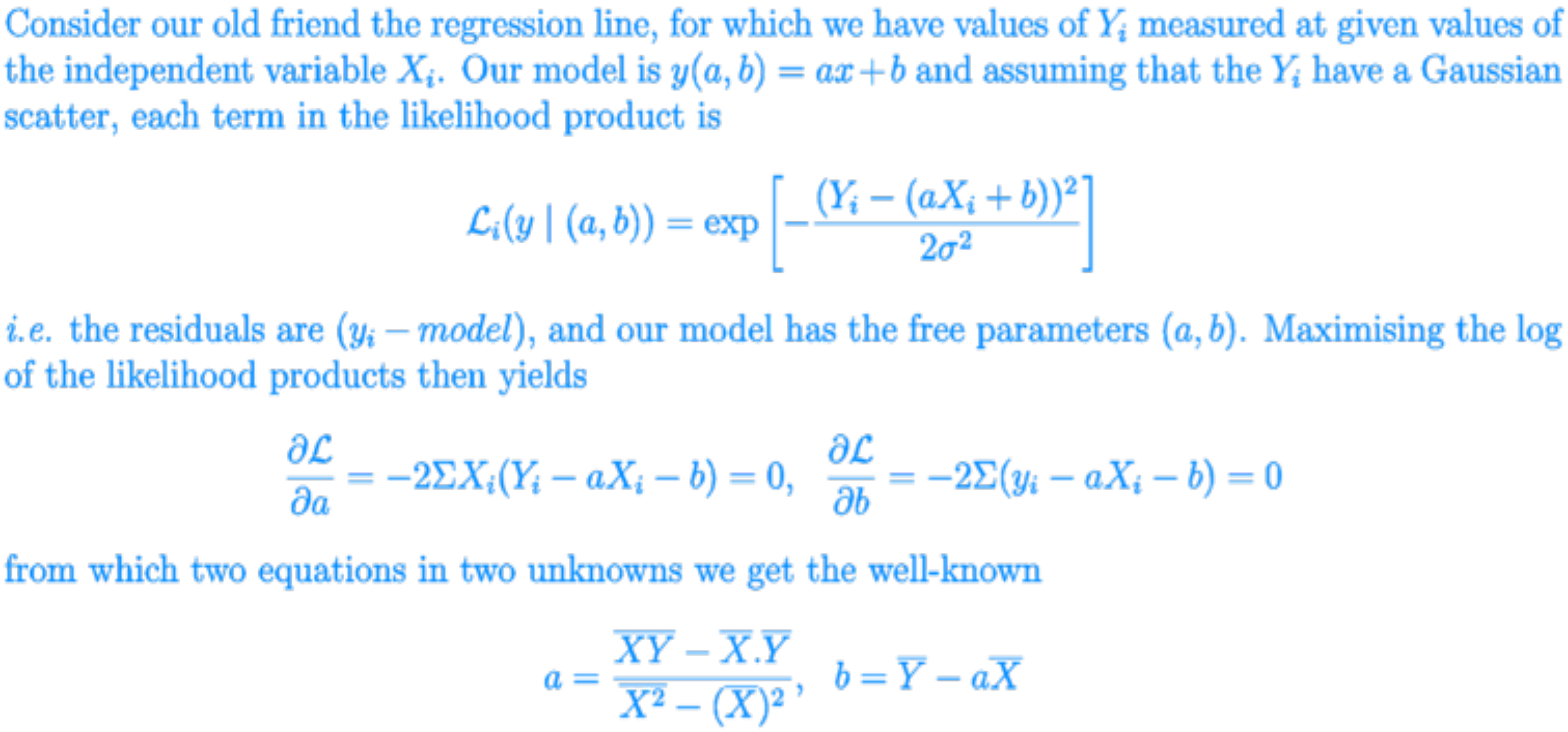
Often we can find this from

$$\frac{\partial}{\partial\alpha}\ln\mathcal{L}(\alpha)\,|_{\alpha=\underline{\alpha}}=0$$

Maximizing the **logarithm** is often convenient.

The MLE is a **statistic** - it depends only on the data, not on any parameters.

Maximum Likelihood - Example 1



Consider our old friend the regression line, for which we have values of Y_i measured at given values of the independent variable X_i. Our model is $y(a,b) = ax+b$ and assuming that the Y_i have a Gaussian scatter, each term in the likelihood product is

$$\mathcal{L}_i(y \mid (a,b)) = \exp\left[-\frac{(Y_i-(aX_i+b))^2}{2\sigma^2}\right]$$

$i.e.$ the residuals are $(y_i - model)$, and our model has the free parameters (a,b). Maximising the log of the likelihood products then yields

$$\frac{\partial \mathcal{L}}{\partial a} = -2\Sigma X_i(Y_i - aX_i - b) = 0, \quad \frac{\partial \mathcal{L}}{\partial b} = -2\Sigma(y_i - aX_i - b) = 0$$

from which two equations in two unknowns we get the well-known

$$a = \frac{\overline{XY} - \overline{X}.\overline{Y}}{\overline{X^2} - (\overline{X})^2}, \quad b = \overline{Y} - a\overline{X}$$

Maximum Likelihood - Example 1 continued

We have (accidentally?) derived the standard OLS, the Ordinary Least Squares estimate of y on the independent variable x. But note that this is:

- given that the Y_i were **Normally distributed** with their scatter described by a single deviation σ;

- given that a **straight-line model was correct**.

We could have started knowing that each Y_i had its own σ_i, or even that the distribution in y about the line was not Gaussian, perhaps say uniform, or dependent on $|Y_i - model|$.

The formulation is identical, but the the ensuing algebra is messier.

Maximum Likelihood - 2

After the MLE estimate has been obtained, it is essential to perform a final check – **does the MLE model fit the data reasonably**? If it does not

- the **data are erroneous** when the model is known to be right;
- the assumed **model is wrong**; or
- there's been a **blunder** of some kind.

There are many ways of carrying out checks; e.g. chi-square test, K-S test, etc.

The strongest reason for picking the MLE of a parameter is that it has desirable properties - e.g. **minimum variance** compared to any other estimate, and **asymptotically distributed** around the true value.

But the MLE is **not always unbiased**.

Key feature of MLE : powerful theorems allow simplification. Instances are given p132-133 of *Practical Statistics for Astronomers*, 2nd ed.: proof of the asymptotic property, and that spread is described by the **covariance matrix *C***, which is calculated from ***C = 1/(E[H]),*** with ***H*** the famous **Hessian matrix**, formed via 2nd derivatives of the likelihood function.

Maximum Likelihood - Example 2

Example of this in action:

A Gaussian of true mean ***μ***, variance ***σ*²**, ***N*** data ***X**ᵢ*.

The log(likelihood) is

$$\log \mathcal{L} = \frac{-1}{2\sigma^2} \sum_i (X_i - \mu)^2 - N \log \sigma, \quad \text{and} \quad \frac{-\partial^2 \log \mathcal{L}}{\partial \mu^2} = \frac{N}{\sigma^2}.$$

The latter expression gives the “**Hessian matrix"**.

Taking its expectation, then the inverse gives the variance on the estimate of the mean as ***σ*²/*N***, the anticipated result.

A marker : The Fisher Matrix

The negative average of the Hessian matrix is even more famous, and is important enough to have a name, to which there are millions of references in the literature:

This is the Fisher Information Matrix (Fisher 1935).

It describes the width of the likelihood function and hence the scatter in the Maximum-Likelihood estimators.

The Fisher matrix can be calculated for various experimental designs as a measure of how well the experiment will perform.

We'll be back…..

Maximum Likelihood - Example 3

Jauncey showed in 1967 that **ML** was an excellent way of estimating **the slope of the number - flux-density relation**, the dependence of source surface density on intensity, for extragalactic radio sources.

The source count is assumed to be of the form

$$N(>S) = k\, S^{-\gamma}$$

where ***N*** is the number of sources on a patch of sky with flux densities greater than ***S***, ***k*** is a constant, and ***γ*** is the exponent (slope in the ***log N - log S*** plane), to estimate.

The probability distribution for S (the chance of getting a source with a flux density near S) is then

$$\mathrm{prob}(S) = \gamma k S^{-(\gamma+1)}$$

and k is determined by the normalization to unity

$$\int_{S_0}^{\infty} \mathrm{prob}(S)\, dS = 1, \quad \rightarrow \quad k = \frac{\gamma}{S_0^{\gamma}}.$$

(We take the maximum possible flux density to be ∞, small error for steep counts.)

Maximum Likelihood - Example 3 concluded

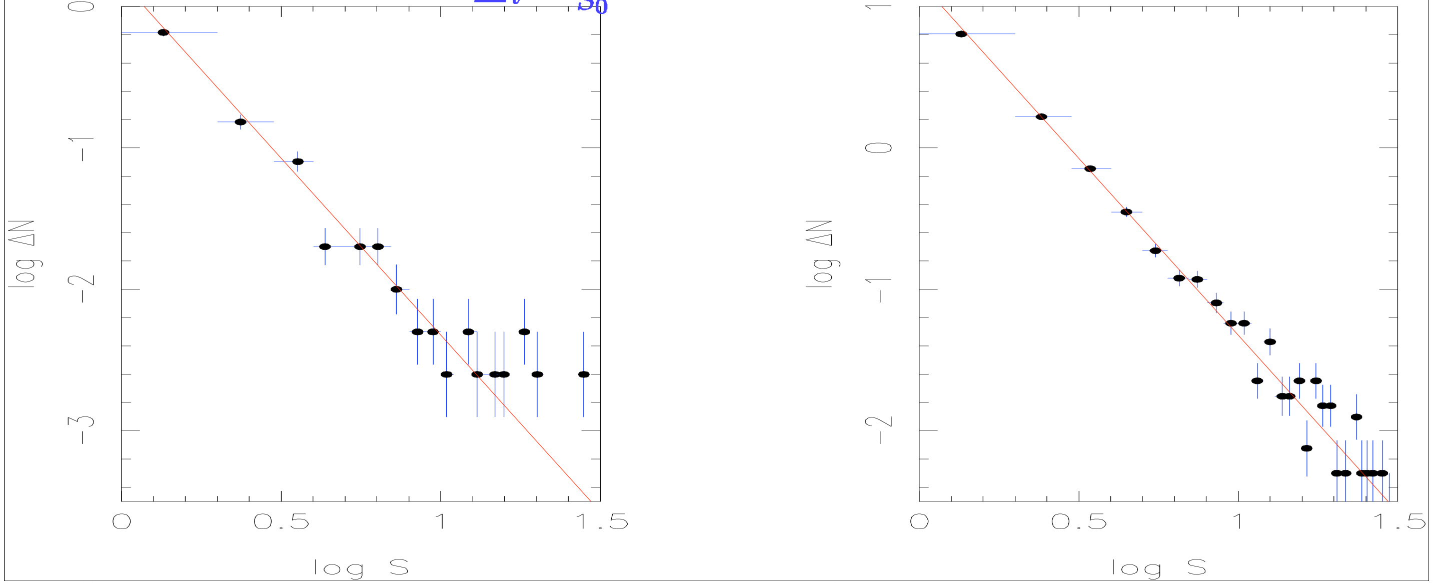
The log-likelihood is, dropping constants,

$$\ln \mathcal{L}(\gamma) = M \ln \gamma - \gamma \sum_i \ln \frac{S_i}{S_0}$$

where we have observed ***M*** sources with flux densities ***S*** brighter than ***S***$_0$.
Differentiating this with respect to ***γ*** to find the maximum then gives the equation for $\underline{\boldsymbol{\gamma}}$, the MLE of ***γ*** :

$$\underline{\gamma} = \frac{M}{\sum_i \ln \frac{S_i}{S_0}}$$

- a nicely intuitive result. **No binning!**



Differential source counts generated via Monte Carlo sampling obeying the source-count law N(>S) = kS$^{-1.5}$. The straight line in each shows the anticipated count with slope -2.5. Left : k = 1.0, 400 trials, Right : k = 10.0, 4000 trials. The ML results for the slopes are -2.52 ± 0.09 and -2.49 ± 0.03, the range given by the points at which the log likelihood function has dropped from its maximum by a factor of 2.
The anticipated errors in the two exponents,
given by $|\text{slope}|^{\sqrt{M}}$, are 0.075 and 0.024.

ML - Example 2 concluded even more

The Figure legend repeated…

Differential source counts generated via Monte Carlo sampling obeying the source-count law N(>S) = kS$^{-1.5}$. The straight line in each shows the anticipated count with slope -2.5. Left : k = 1.0, 400 trials, Right : k = 10.0, 4000 trials. The ML results for the slopes are -2.52 ± 0.09 and -2.49 ± 0.03, the range given by the points at which the log likelihood function has dropped from its maximum by a factor of 2. The anticipated errors in the two exponents, given by **|slope|/√M, are 0.075 and 0.024**.

Uh, wait, where did you say that |slope|/√M was the anticipated error in *γ*?

We have just one parameter. The variance on ***γ*** is (recall ***C=(E[H])$^{-1}$***) :

$$\frac{-1}{E[\frac{\partial^2 \mathcal{L}(\gamma)}{\partial \gamma^2}]} \rightarrow \frac{\gamma^2}{M}$$

(The expectation calculation is easy in this case.) However, we see that the error is given in terms of the thing we want to know, namely ***γ***.

As long as the errors are small we can approximate them by *γ*/√M.

Least Squares: Regression Analysis

Laplace! Justification follows immediately from the method of ML. If the distribution of the residuals is Gaussian, then the ***log(likelihood)*** is a sum of squares of the form

$$\log \mathcal{L} = \mathrm{constant} - \sum_{i=1}^{N} \xi_i (X_i - \mu(\alpha_1, \alpha_2 \ldots))^2$$

where the **ξ** are the weights, obviously inversely proportional to the variance on the measurements.

Usually the weights are assumed equal for all the data, and least-squares is just that. We seek the model parameters which minimize

$$\log \mathcal{L} = \mathrm{constant} - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (X_i - \mu(\alpha_1, \alpha_2 \ldots))^2.$$

These will just be the maximum-likelihood estimators, and everything carries over - asymptotically distributed like a multivariate Gaussian.

If we do not know the error level (the **σ**) we do not need to use it, but **we will not be able to infer errors on the MLE.** We will get get a model fit, but we will never know how good or bad the model is.

Least Squares: Regression Analysis 2

The matrix of 2nd derivatives defining the covariance matrix of the estimates, the **Hessian matrix**, is often used by **numerical algorithms which find the minimum.**

There are many powerful variations on these algorithms (e.g. AMOEBA: see Numerical Recipes).

Typically the value of the Hessian matrix, at the minimum, pops out as a by-product of the minimization.

We can use this **directly** to work out the covariance matrix, as long as our model is **linear** in the parameters.

In this case, the expectation operation is straightforward and the matrix does not depend on any of the parameters.

Least Squares : what's a LINEAR Model then?

Suppose our data ***X_i*** are measured as a function of some independent variable ***Z_i***.

Then a linear model - linear in the **parameters** - might be

$\alpha z^2 + \beta \exp(-z)$,

whereas

$\alpha \exp(-\beta z)$

is **not** a linear model.

Of course a model may be approximately linear near the MLE.

How close must it be? This illustrates again the general feature of the asymptotic Normality of the MLE - we can use the approximation, but

we can't tell how good it is.

Usually things will start to go wrong first in the **wings of the inferred distributions**, and so high degrees of significance usually cannot be trusted unless they have been calculated exactly, or simulated by Monte Carlo methods.

Try working out the MLE using **different assumptions on the residuals** – e.g. a simple exponential, or a t-distribution –

Are your outliers driving the answer?

Least Squares - Example: the ‘Regression Line’

Least-squares fit through **N pairs of (X_i,Y_i)** by minimizing squares of residuals:

$$y = ax + b; \quad a = \frac{\overline{XY} - \overline{X}.\overline{Y}}{\overline{X^2} - (\overline{X})^2}, \quad b = \overline{Y} - a\overline{X}.$$

You can fit **any two-parameter curve** this way with simple coord transformations:

1. an exponential, $y = b \exp a$ requires y_i to be changed to $\ln y_i$ in the above expressions,
2. a power-law, $y = bx^a$; change y_i to $\ln y_i$ and x_i to $\ln x_i$;
3. a parabola, $y = b + ax^2$; change x_i to $\sqrt{x_i}$.

Note : - the residuals **cannot** be Gaussian for all of these transformations (and may not be Gaussian for any);

- it is **always** possible to minimize the squares of the residuals, but the formal justification?
- the **one-sample hypothesis tests** can be revealing as to which (if any) model fits, particularly the runs test.

This simple formulation of the **least-squares fit for *y* on *x*** represents the tip of an iceberg ……………………

The Regression Line - 2

A glimpse of the iceberg - there is an enormous variety of least-squares linear regression procedures. Amongst the issues involved in choosing a procedure:

1. Are the data to be treated **weighted or unweighted**?

2. Do all the data have the same properties, e.g. in the simple case of y on x, is one σ_y^2 applicable to all y? Or does it depend on y? In the uniform σ case, the data are described as **homoskedastic**, and in the opposite case, **heteroskedastic**.

3. Is the right fit the **standard ordinary least squares solution** y on x **(OLS(Y/X))** or x on y **(OLS(X/Y)?** Or something different?

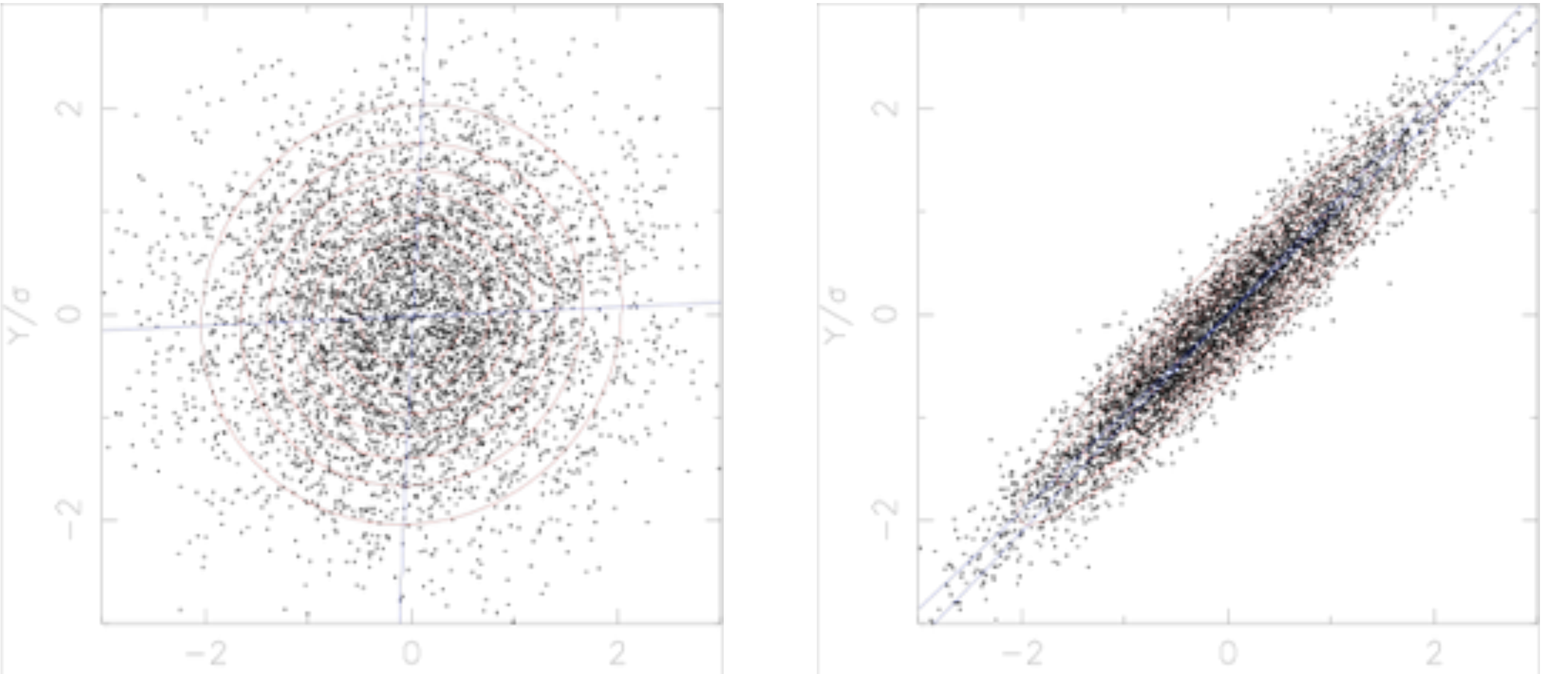
4. If we **know** we have heteroskedasticity, but with **known uncertainties** in each Y_i and each X_i, how do we use this information to estimate the uncertainty in the fit?

5. Are the data **truncated or censored**; do we wish to include upper limits in our fit? This is perfectly possible.

See the thorough papers of Eric Feigleson et al. (1990-1992) – bootstrap and jackknife resampling to get the errors, and much more.

And what is the scientific question?

An Example of the Regression Line Example



Linear contours of the bivariate Gaussian probability distribution. Left: ***ρ*** = 0.05, a bivariate distribution with weak connection between ***x*** and ***y***; right: ***ρ*** = 0.95, indicative of a strong connection. In each case 5000 (***x,y***) pairs are plotted, selected at random from the appropriate distribution. Two lines are shown as fits for each distribution, the **OLS(X/Y)** and the **OLS(Y/X)**.

But we know the answer! A line of slope 45^{o} should result? **No. What's the question?**

If we need **'the relation'** and we have no prior – then use the **bisector line** (average OLS), the **orthogonal regression** line (minimizes perpendiculars), or **PCA** – does not assume which variable is 'in control'.

Minimum Chi-Square Method

♣ a dominant classical modelling process, a simple **extension of the chi-square goodness-of-fit test** and closely related to (weighted) least squares methods.

♣ minimum chi-square statistic has **asymptotic properties similar to ML.**

♣ for observational data which can be (or are already) binned, with a model predicting population of each bin. Chi-square statistic describes the goodness-of-fit of the data to the model. If the **observed** numbers in each of ***k*** bins are ***O_i***, and the **expected** values from the model are ***E_i***, then

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

♣ it's the squares of the residuals **weighted by effectively the variance** if the procedure is governed by Poisson statistics.

♣ now **minimize the χ^2 statistic by varying the parameters** of the model.

♣ parameter search is OK as long as there are less than 4; otherwise we need a proper **search procedure** - see Numerical Recipes for a great range of these.

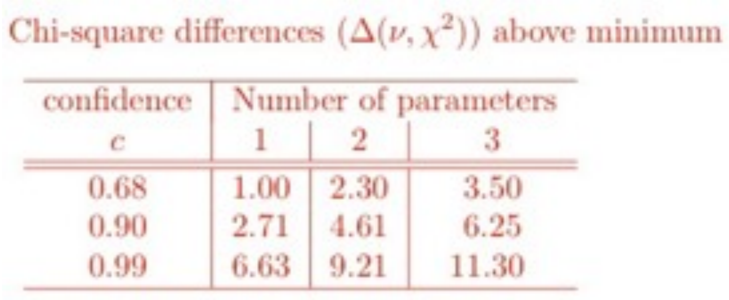
♣ Of these, **AMOEBA is good** – 'downhill simplex' method of steepest descent.

Minimum Chi-Square Method - 2

♣ confidence limits? (A mysterious sum; nobody but me will tell you its origin……)

$$\chi^2_\alpha = \chi^2_{min} + \Delta(\nu, \alpha)$$

where ***Δ*** is from:



Chi-square differences ($\Delta(\nu, \chi^2)$) above minimum

confidence	Number of parameters		
c	1	2	3
0.68	1.00	2.30	3.50
0.90	2.71	4.61	6.25
0.99	6.63	9.21	11.30

♣ appropriate dof ***v*** to associate = ***(k -1 - N), k bins, N parameters.***

♣ Debit and loss: (+) **additive**, so results of different data sets that may fall in different bins, bin sizes, or compare different aspects of same model, may be tested all at once.

(+) **contribution of each bin** may be examined - regions of good or bad fit delineated.

(+) **model-testing for free**. Min model should have value of order 0.5 – remember peak of **χ^2 distribution** is ~ ***v*** when ***v*** > 4.

(-) low bin-populations in the chi-square sums will cause **severe instability**. **80%** of the bins must have ***E_i > 5***.

(-) **data-binning is bad**. It loses information and efficiency. It can unskew skewed distributions.

And here is something known only to us ....

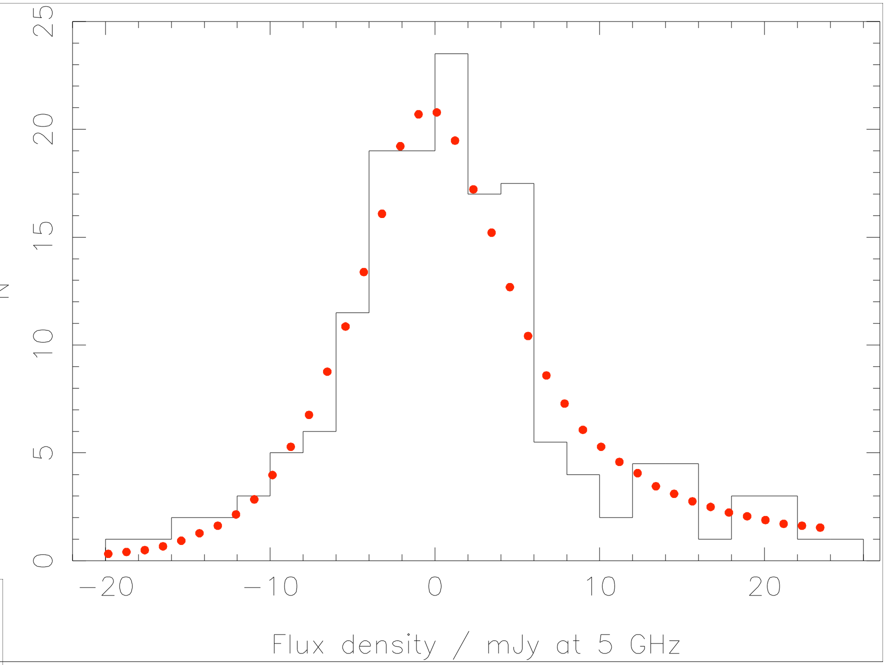
The table indicates that there is a probabilility **c** that this region will contain the true values of the parameters. It is calculated from

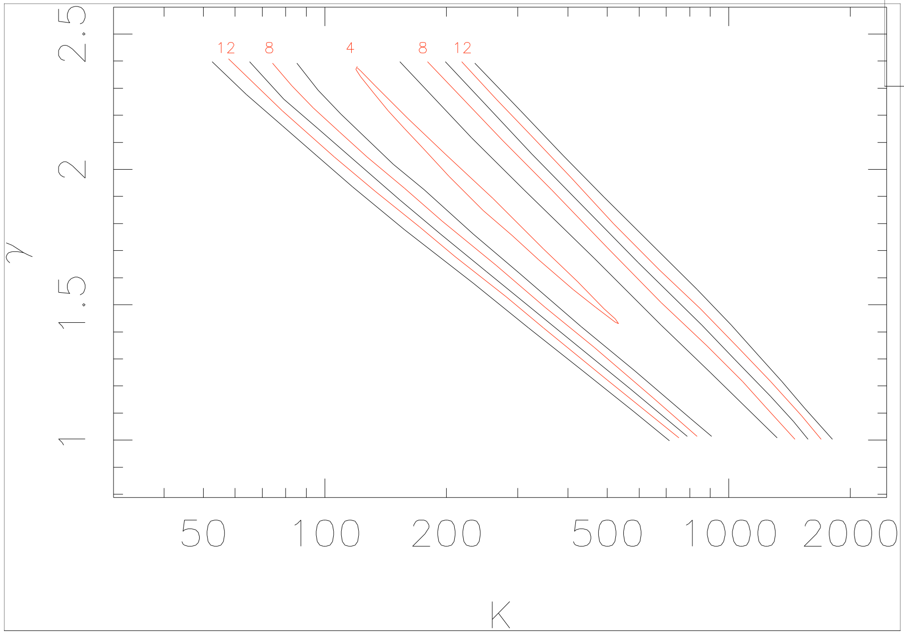
$$c(\nu, \Delta\chi^2) = P(\nu/2, \Delta\chi^2/2)$$

with **P** the incomplete Gamma-function (NumRec – Press et al. 2007)

Minimum Chi-Square Method - Example

Chi-square testing/modelling: the object of the experiment was to estimate the surface-density count (the ***N(S)*** relation) of faint radio sources at 5 GHz, assuming a power-law ***N(>S) = KS***$^{-(\gamma-1)}$**,** ***γ*** and ***K*** to be determined from the distribution of background deflections, the **P(D) method**. The histogram of measured deflections is shown right.





The dotted red curve above represents the optimum model from minimizing χ^2. Contours of χ^2 in the ***γ - K* plane** are shown left.

With the best-fit model, χ^2 = 4 for 7 bins, 2 parameters; thus dof = 4. **Right on.**