Statistics and Data Science: Lecture 2

Roger Barlow
Huddersfield University

Cockcroft Lecture Series

30^{th} May 2022

Fitting

General problem: you have a dataset $\{x_1, x_2 \ldots x_N\}$ and a probability (density) function $P(x; a)$. (The x_i may be multidimensional. So may a.)
You need to know:

1. What is the best value for a? **Estimation**
2. How accurate is that? **Errors**
3. Does the model truly describe this data? **Goodness of fit**

Estimation

Very general definition: an estimator is a function of the data which returns a value for the parameter you want to know about.

$\hat{a}(x_1 \dots x_N)$

gives a number hopefully close to the true value of a.
(N.b. not a rough guess. Carefully evaluated)
A good estimator is:

1. Consistent. $\hat{a} \to a_{true}$ as $N \to \infty$.
   If you take enough data it will give the right answer
2. Unbiassed. $\langle \hat{a} \rangle = a_{true}$.
   A particular instance may be too high or too low but over many measurements this balances.
3. Efficient. $\left\langle (\hat{a} - a_{true})^2 \right\rangle$ should be small
   It turns out there is a limit to the efficiency: the *Minimum Variance Bound* (MVB)

Introducing Likelihood

The likelihood is just the combined probability (density) for the dataset

$$L(x_1 ... x_N; a) = P(x_1; a)P(x_2; a)...P(x_N; a)$$

Averaging over many repetitions gives *Expectation Values*

$$\langle f \rangle \equiv E(f) = \int \int \cdots \int f(x_1, x_2 \ldots x_N) L(x_1, x_2 \ldots x_N; a) dx_1\, dx_2 \ldots dx_N$$

Note: expectation values are functions of a but not of x - that's all been integrated away

Reminder

$L(x_1 \ldots x_N; a)$ is the likelihood for a particular set of results, given some value for the parameter a. It is *not* the likelihood for a having a particular value.

Maximum Likelihood Estimation

General principle for $\hat{a}(x_1 \dots x_N)$: choose the value of a which maximises $L(x_1 \dots x_N; a)$ (In practice: maximise $\ln L = \sum_i \ln P(x_i; a)$.)

Example

N measurements of something, each Gaussian with standard deviation σ_i

$\ln P(x_i; a) = -\frac{1}{2}\frac{(x_i-a)^2}{\sigma_i^2} - \ln \sigma_i \sqrt{2\pi}$

Find maximum by

$\frac{d \ln L}{da} = 0 = \sum_i \frac{x_i - \hat{a}}{\sigma_i^2}$

$\implies \hat{a} = \sum \frac{x_i}{\sigma_i^2} / \sum \frac{1}{\sigma_i^2}$

Example

Measurements are unknown mixture of signal, $S(x)$, and background, $B(x)$. a is fraction that is signal.

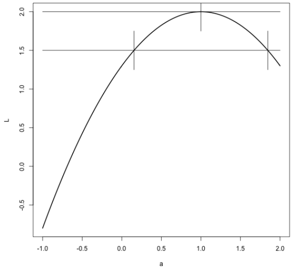
$P(x; a) = aS(x) + (1-a)B(x)$

$\implies \sum \frac{S(x_i)-B(x_i)}{aS(x_i)+(1-a)B(x_i)} = 0.$ Solve numerically

ML estimators

- Are consistent
- Are in general biassed – but the bias falls line $\frac{1}{N}$
- Are as efficient as allowed by the MVB: $V(\hat{a}) = -1/\left\langle \frac{\partial^2 L}{\partial a^2} \right\rangle$

- Differentiate and solve algebraically
- Differentiate and solve numerically
- Maximise numerically



Solving numerically one reads off $\sigma_{\hat{a}}$ from points where $\Delta \ln L = -\frac{1}{2}$, approximating $\left\langle \frac{\partial^2 L}{\partial a^2} \right\rangle = \frac{\partial^2 L}{\partial a^2}$

Fitting for several variables

Same technique

Example

N measurements of something, Gaussian but both mean and σ unknown

$\ln P(x_i; \mu, \sigma) = -\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2} - \ln \sigma\sqrt{2\pi}$

Find maximum by

$\frac{d\ln L}{d\mu} = 0 = \sum_i \frac{x_i-\hat{\mu}}{\hat{\sigma}^2}$

$\frac{d\ln L}{d\sigma} = 0 = \sum_i \frac{(x_i-\hat{\mu})^2}{\hat{\sigma}^3} - \frac{N}{\hat{\sigma}}$

$\implies \hat{\mu} = \frac{1}{N}\sum x_i$ and $\hat{\sigma}^2 = \frac{1}{N}\sum(x_i - \hat{\mu})^2$

(Notice how this doesn't include Bessel's correction.)

Errors are a bit different. '1-sigma' error region given by $\Delta \ln L = -1.14$

From ML to least squares

Gaussians again

Suppose each data point is an (x, y) pair, with a predicted function $f(x; a)$ and y is Gaussian, mean $f(x; a)$ and standard deviation σ.

Log likelihood is then $\ln L = -\frac{1}{2}\sum\left(\frac{y_i - f(x_i;a)}{\sigma_i}\right)^2 = -\frac{1}{2}\chi^2$

Maximising likelihood $\equiv$ Minimising χ^2. Hence the name 'least squares'

And $\Delta \ln L = -\frac{1}{2} \equiv \Delta\chi^2 = 1$

Differentiate and set to zero $\implies \sum_i \frac{\partial f(x_i;a)}{\partial a}\frac{f(x_i;a)}{\sigma_i^2} = \sum_i \frac{\partial f(x_i;a)}{\partial a}\frac{y_i}{\sigma_i^2}$

If $f(x; a)$ is linear in a (e.g. $f(x) = a_0 + a_1 x + a_2 x^3 + a_3 sin(x)$) can write $f(x_i) = \sum_j c_j(x_i)a_j = \sum_j C_{ij}a_j$, $\frac{\partial f(x_i)}{\partial a_j} = C_{ij}$, and equation becomes

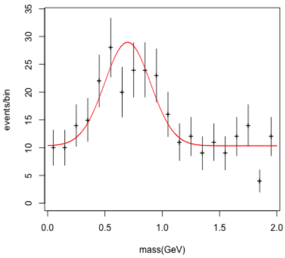
$$\tilde{\mathbf{C}}\mathbf{V_y}^{-1}\mathbf{C}\hat{\mathbf{a}} = \tilde{\mathbf{C}}\mathbf{V_y}^{-1}\mathbf{y}$$

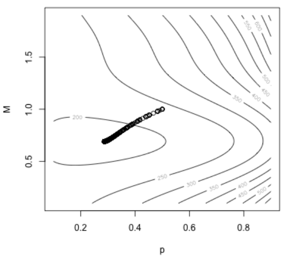
(If the y_i are independent then $\mathbf{V_y}^{-1}$ is diagonal with elements $1/\sigma_i^2$)

$$\hat{\mathbf{a}} = (\tilde{\mathbf{C}}\mathbf{V_y}^{-1}\mathbf{C})^{-1}\tilde{\mathbf{C}}\mathbf{V_y}^{-1}\mathbf{y}$$

Do not invert the matrix! Use `solve(M,v)` or equivalent `linsolve`, `np.linalg.solve` .

An example





300 values drawn from Gaussian of unknown mean, $\sigma = 0.2$, on a flat background

2 parameters: M the Gaussian mean and p the fraction

Method 1: histogram in 20 bins, do χ^2 fit Fit converges to $(M = 0.63, p = 0.37)$

Method 2: Maximum likelihood. Contours shown, Optimizer starts at $(M = 1.0, p = 0.5$ and converges to $(M = 0.69, p = 0.29)$

N.b. true values $M = 0.6, p = \frac{1}{3}$

Example - the program (R version)

```
stepsize=0.1
h=hist(data,breaks=seq(0,2,stepsize),,plot=FALSE) #   data contains the numbers for the fit
y=h$counts
x=h$mids
errors=sqrt(y)
ntot=sum(y)
plot(x,y,pch="+",ylim=c(0,max(y)+1.1*sqrt(max(y))),xlab="mass(GeV)",ylab="events/bin")
for(i in 1:length(x)) lines(x[i]*c(1,1),y[i]+errors[i]*c(1,-1))  # draw error bars

f<-function(p){ return (ntot*stepsize*(p[1]*dnorm(x,p[2],.2)+(1-p[1])*.5)) }

chisq<-function(p){ return(sum(((y-f(p))/errors)^2)) }

o=optim(c(.5,1),chisq,method="CG",control=list(maxit=200,parscale=0.01* c(1,1)))
print(o)

x=seq(0,2,.001)
lines(x,f(o$par),col='red')

NLL<-function(p){ return(-sum(log(p[1]*dnorm(data,p[2],.2)+(1-p[1])*.5))) }

xx=seq(.1,.9,.01)
yy=seq(.1,1.9,.01)
zz=xx %o% yy
for (i in 1:length(xx)){
    for (j in 1:length(yy))
      zz[i,j]=NLL(c(xx[i],yy[j]))
}
contour(xx,yy,zz,xlab='p',ylab='M')

o=optim(c(.5,1),NLL,method="CG",control=list(maxit=200,parscale=0.01* c(1,1)))
print(o)
```

Numerical methods for optimisation

R: `optim`
Python: `scipy.optimize.minimize`
MATLAB. `fminsearch` and `fminunc`
Methods

1. Simplex. Slow but 'safe', shrinking mesh method
2. Gradient-based. Generalisations of Newton's method are faster provided one is close to the true minimum. Gradient may be supplied by the user or evaluated numerically.
3. Annealing. Find minimum, then jump to random point and re-start and check solution is the same.

Other arguments involve limits on parameters (best to avoid), step sizes, tolerances for claiming solution, etc.

Very obvious point

When maximising a function using a minimiser, don't forget the minus sign

Errors: another way to evaluate them

Using χ^2 we have $\hat{\mathbf{a}} = (\tilde{\mathbf{C}}\mathbf{V_y}^{-1}\mathbf{C})^{-1}\mathbf{C}\mathbf{V_y}^{-1}\mathbf{y}$

Errors on a are due to errors on y, the σ_i, and the usual combination of errors formula can be used

After some algebra,

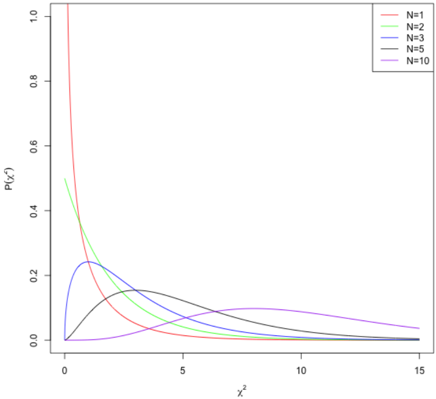
$$\mathbf{V_a} = (\tilde{\mathbf{C}}\mathbf{V_y}^{-1}\mathbf{C})^{-1}$$

Showing that this is compatible with the errors from $\Delta\chi^2 = 1$ is left as an exercise for the reader.

χ^2 and goodness-of-fit

Remember, writing $f_i \equiv f(x_i; a)$, $\qquad \chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - f_i}{\sigma_i}\right)^2$

Obviously, χ^2 should be about N
Bit of algebra:
$P(\chi^2; N) = \frac{\chi^{N-2} e^{-\chi^2/2}}{2^{N/2}\Gamma(N/2)}$
Function available as `dchisq` in R, `chi2pdf` in MATLAB and `chi2.pdf` from scipy.stats
If the function has been fitted then $N \to N_{DF} = N_{points} - N_{pars}$ number of "degrees of freedom"
Turns out
$\langle \chi^2 \rangle = \int_0^\infty \chi^2 P(\chi^2; N)\, d\chi^2 = N$

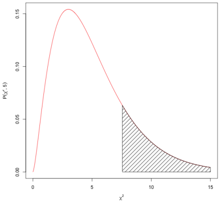


If $\chi^2 >> N$ then (i) your model is wrong or (ii) your data is wrong or (iii) your errors are underestimated or (iv) you are unlucky
If $\chi^2 << N$ then (i) your errors are overestimated or (ii) you are lucky

Hypotheses and p-values

p-value

Probability under the null hypothesis of getting a result this extreme (or worse)



Suppose $N = 5$ and you get $\chi^2 = 7.56$. Is that bad?
p-value is $\int_{7.56}^{\infty} P(\chi^2; 5)\, d\chi^2 = 0.18$
If there is a distribution which really is $P(\chi^2, 5)$, the probability of getting a χ^2 value of 7.56 or more is 18%

This is an instance of *hypothesis testing*. To make the case for an effect, hypothesis H_1, you have to show that the null hypothesis H_0 is implausible
E.g. to show a medicine works, you have to show that this many cases would not have occurred by chance
Here: if you want to show that y does vary with x, for 6 values, you have to show that the hypothesis $y = constant$ is implausible (and in this case you havn't)

Goodness of fit from Likelihood

Very short slide

Can't be done. The actual value of the likelihood tells you *nothing* about the fit quality. Even if you include all the constant factors

WIlks' theorem says that (for large N) $\Delta \ln L$ behaves like χ^2, but this only applies to the difference in log likelihood for two models, where one is the limiting case of the other, and does not have any parameters which are meaningless in the first model. So you can use it, e.g. to see whether a cubic gives a meaningfully better fit than a parabola. But not whether either fit is valid.

Bayesian Methods

Really needs a whole lecture, not just one slide...

Bayes' Theorem

$P(A|B) = \frac{P(B|A)}{P(B)} P(A)$

Bayes Theorem for parameters

$P(a|x) \propto L(x; a) P(a)$

Posterior $\propto$ likelihood $\times$ Prior

From Posterior you can get best value, errors, etc.
But to use this you need the Prior - which is (usually) not just unknown but meaningless (unless you switch to using subjective probability)
Different priors give different posteriors. A uniform prior is not the answer

Bayesian methods can often be illuminating and sometimes essential. But they come with a whole slew of problems the salesmen don't tell you about

Setting Limits

Really needs a whole lecture, not just one slide...Particle physicists spend a lot of time on this. Accelerator physicists less so. Be grateful!

Searching for a signal which may or may not be there

Discovery: see a signal and use p-value to establish that it is very unlikely ($p < 3 \times 10^{-7}$, or 'five sigma') that a model $H_0(S = 0)$ with zero signal would give a result this extreme (i.e. this large or larger)

Non-discovery: signal is small/zero. Find a signal strength S_+ such that it is quite unlikely ($p < 5\%$, or maybe 10%) that under a model $H_0(S = S_+)$ would give a result this extreme (i.e. this small or smaller). Then quote S_+ as 95% (or maybe 90%) confidence level upper limit.

" We observe only 6 events, with an expected background of 2.1 events. As there is a 2% probability of getting 6 or more events from a Poisson with mean $\mu = 2.1$ we claim evidence for a signal, but not a discovery. For $\mu = 3.3$ or more, the probability of getting 6 events or less is only 5%. We therefore say with 95% confidence that if there is any signal it is below the equivalent of 1.2 events"

Systematic errors (1)

What they are

Errors that are random but shared between measurements

Examples

- Poisson counts from a detector with efficiency $\eta \pm \sigma_\eta$
- BPM measurements where the calibration is $c \pm \sigma_c$
- Disease instances by date where the collection efficiency is $C \pm \sigma_C$

What they are not

Mistakes, faulty equipment, wrong assumptions, misconnected cables

Why they are scary

If you get them wrong, they do not show up as bad χ^2 etc

Systematic errors (2): How to evaluate them?

The ancillary experiment

A separate experiment, often a calibration. Or a Monte Carlo simulation

Guesswork

Expert opinion (based on knowledge, experience, etc) of the uncertainty
Be careful to use 68% sigma-type errors, not tolerances. Do not be tempted to be ‘conservative’.

Systematic errors (3): How to apply them

Standard combination-of-errors-formula including correlations. For $f(x,y)$

$$\sigma_f^2 = \left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_y^2 + 2\left(\frac{\partial f}{\partial x}\right)\left(\frac{\partial f}{\partial y}\right) Cov(x,y)$$

If x and y have individual errors s_x, s_y and shared error S then variance matrix is

$$\mathbf{V} = \begin{pmatrix} s_x^2 + S^2 & S^2 \\ S^2 & s_y^2 + S^2 \end{pmatrix}$$

Matrices for more variables, and experiments with different shared systematics, can be built up in the same way

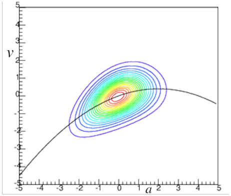
Systematic Errors (4): Nuisance parameters

It can be helpful to think of systematic uncertainties as 'nuisance parameters', ν

Write down the likelihood in terms of the raw measurements with the factors applied explicitly, including prior knowledge from the ancillary experiment etc.

Fit by maximising the likelihood in all parameters

Total error from $\Delta \ln L = -\frac{1}{2}$ in the profile likelihood $L(a, \hat{\hat{\nu}}, x)$ where for each value of a considered, the value of ν is adjusted to give the maximum likelihood for that a



Checks

Although a systematic error is not a mistake, mistakes do happen
Checks are an important safeguard against mistakes

- Analysing subsets of data (this year's data and last year's)
- Making changes which should in principle give the same result (changing bin size for fitting histogram)
- Making measurements for which the answer is known (if measuring mass of the Higgs, check you get the right mass for the Z)

If the differences are small, say "OK" and move on. Do NOT add them to the systematic error
If the differences are large, worry and sort them out. Do NOT just add them to the systematic error.

What happens next

1. Split into groups as before
2. Download `https://barlow.web.cern.ch/barlow/lecture2.dat`
3. The data is a set of measurements containing two Gaussian peaks on a flat background. One has width 0.1 and the other 0,2.
4. Determine the masses, using a histogram and χ^2 and by maximising the unbinned likelihood. Show the results are similar but different
5. Look at how the binning affects your results from the histogram
6. Explore the options of your minimiser package
7. Prepare a short presentation of your results. Time about 10 minutes
8. After lunch (2:00) we re-convene. Groups make their presentations in turn, and the rest of us listen and learn and criticise.