9. Fitting models to data

Sebastian Funk
20 June 2018

Improving health worldwide

www.lshtm.ac.uk

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

OBJECTIVES

By the end of this lecture you should be able to:

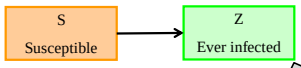
1. Explain the need to fit models to data.
2. Define key terms about model fitting/calibration and sensitivity analysis.
3. Explain the principles behind the least squares, weighted least squares and maximum likelihood methods of fitting.
4. Implement these methods in MS Excel and Berkeley Madonna.

INTRODUCTION

Infectious disease models

- Structure
- Parameters and initial conditions
- Data to fit to

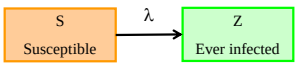
EXAMPLE: CATALYTIC MODEL



Note: This is the seroprevalence model we fitted in Practical 7.

Structure

EXAMPLE: CATALYTIC MODEL



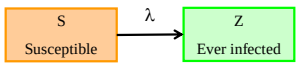
S(0) = 1 Z(0) = 0

Structure

Parameters

Initial conditions

EXAMPLE: CATALYTIC MODEL



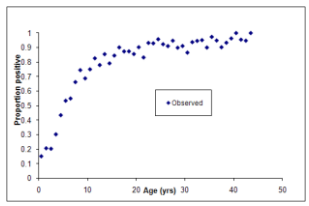
S(0) = 1 Z(0) = 0

Structure

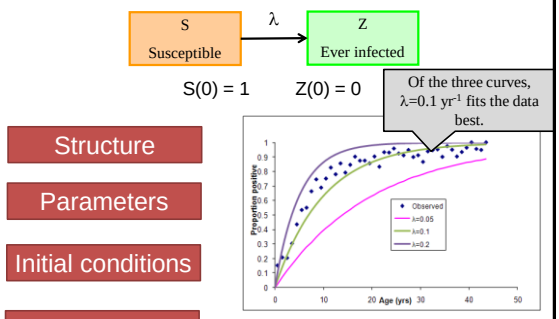
Parameters

Initial conditions

Data to fit to



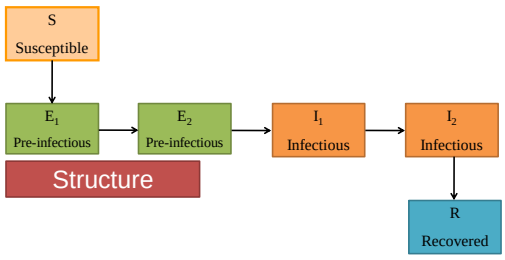
EXAMPLE: CATALYTIC MODEL



EXAMPLE: H1N1 MODEL

A more complex example

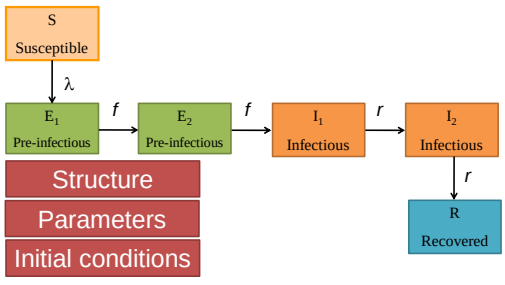
(Baguelin, Van Hoek *et al.* Vaccine 2010; 28:2370)



EXAMPLE: H1N1 MODEL

A more complex example

(Baguelin, Van Hoek *et al.* Vaccine 2010; 28:2370)

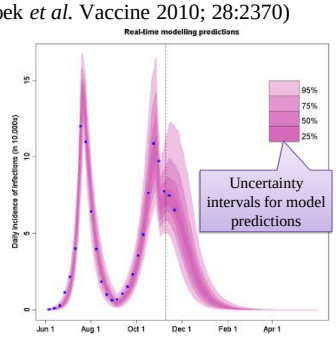


EXAMPLE: H1N1 MODEL

A more complex example

(Baguelin, Van Hoek *et al.* Vaccine 2010; 28:2370)

Number of influenza A/H1N1v infections in England:
Fitted model predictions (pink shades)
Case estimates until 15 Nov (blue dots)



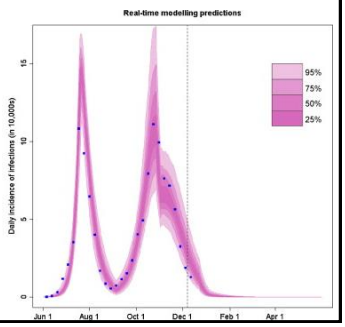
Structure
Parameters
Initial conditions
Data to fit to

EXAMPLE: H1N1 MODEL

A more complex example

(Baguelin, Van Hoek *et al.* Vaccine 2010; 28:2370)

Number of influenza A/H1N1v infections in England:
Fitted model predictions (pink shades)
Case estimates until 3 Dec (blue dots)



Structure
Parameters
Initial conditions
Data to fit to

TERMINOLOGY

Some terminology

Parameterisation	Giving values to the parameters in a model (by whatever means eg. guessing, directly from the literature, fitting to data).
Fitting (or calibration)	Parameterisation of a model by finding a parameter set that produces model results with a "good fit" to outcome data.
Validation	Comparing results of a parameterised model to data to see if the results are "valid" (either have face validity or satisfy some statistical test).
(Parametric) sensitivity analysis	Altering the parameters of a model to see what effect this has on results.

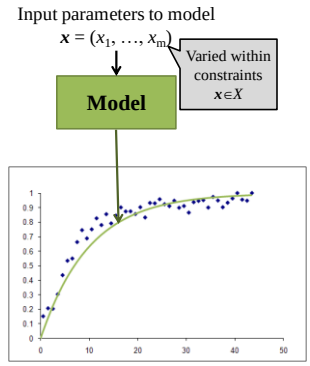
TERMINOLOGY

Two key issues:

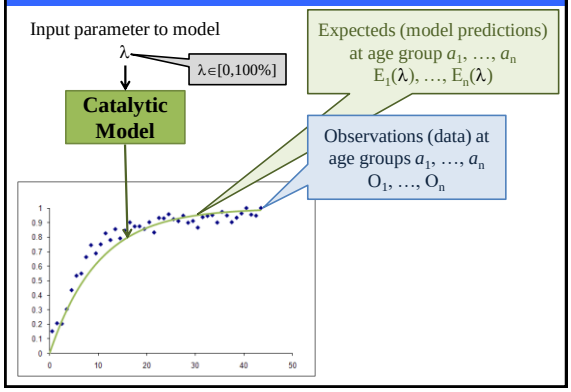
1) How do we decide whether a model provides a "good" fit to data?
→Need a **goodness of fit** metric (statistical issue).

2) How do we find the "best fitting" parameters?
→Need a **fitting algorithm** (computational issue).

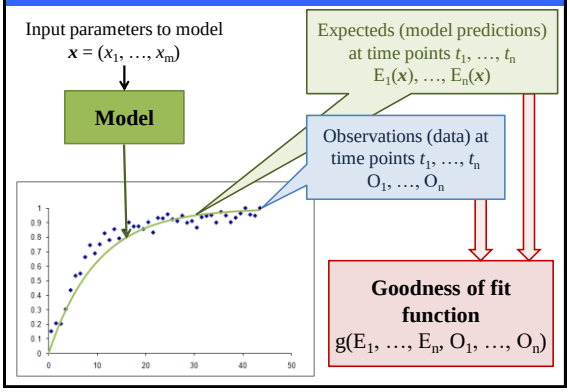
TERMINOLOGY



TERMINOLOGY



TERMINOLOGY



TERMINOLOGY

The **goodness of fit function**, $g(E_1, \ldots, E_n, O_1, \ldots, O_n)$, describes how well the model predictions $E_1(\boldsymbol{x}), \ldots, E_n(\boldsymbol{x})$ fit data $O_1, \ldots, O_n$ at time points $t_1, \ldots, t_n$ for a given value of $\boldsymbol{x}$.

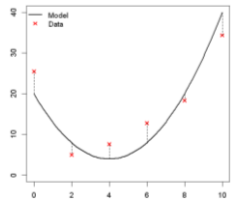
The larger (or smaller, depending on the function) it is, the better the model fit to data. Hence it is called the **objective function** or **maximand** (minimand).

Formally, we write the following:

Maximise $f(\boldsymbol{x})$ subject to $\boldsymbol{x} \in X$

where $f(\boldsymbol{x}) = g(E_1(\boldsymbol{x}), \ldots, E_n(\boldsymbol{x}), O_1, \ldots, O_n)$

GOODNESS OF FIT METRICS



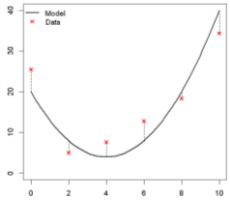
Some obvious choices for the goodness of fit metric

1) Linear distances between the data point and model prediction (the **residual**)

$$g(x) = \sum_i \left(E_i(x) - O_i\right)$$

Problem:
Positive and negative distances cancel out.

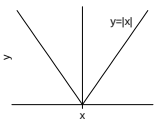
GOODNESS OF FIT METRICS



Some obvious choices for the goodness of fit metric

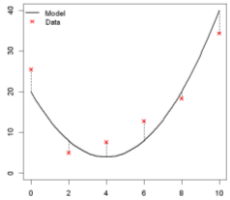
2) Absolute distances between the data point and model prediction

$$g(x) = \sum_i |E_i(x) - O_i|$$



Problem:
Strange behaviour at zero makes the function difficult to deal with both analytically and computationally.

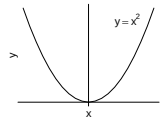
GOODNESS OF FIT METRICS



Some obvious choices for the goodness of fit metric

3) Sum of squared distances between the data point and model prediction

$$g(x) = \sum_i (E_i(x) - O_i)^2$$

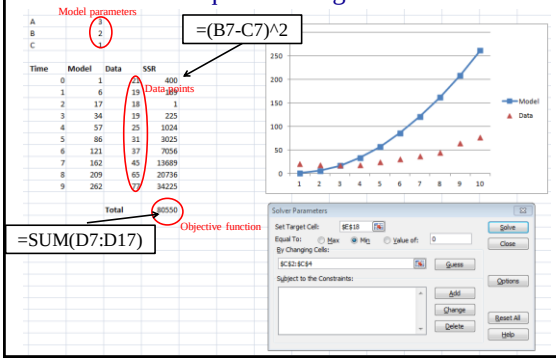


This is called the **residual sum of squares (SSR or SSQ)**, and is a very popular goodness of fit metric.

The method is called the **method of least squares**.

METHOD OF LEAST SQUARES

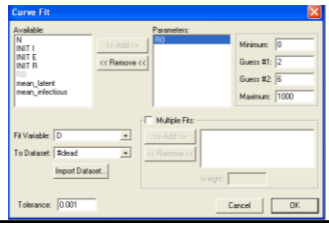
Least-squares fitting in Excel



METHOD OF LEAST SQUARES

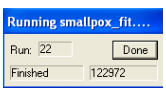
Least-squares fitting in BerkeleyMadonna

- Recall that previously you used Berkeley Madonna to fit a model of pandemic influenza to disease incidence data by changing two parameters (β_1 and β_2).
- Berkeley Madonna can fit one or more parameters, although it is less reliable for multiple parameters.



METHOD OF LEAST SQUARES

Least-squares fitting in BerkeleyMadonna



- The "fit" in Berkeley Madonna is the root mean squared difference between the data and model prediction.

$$Fit = \sqrt{\frac{1}{n}\sum_i (E_i - O_i)^2} = \sqrt{\frac{1}{n} SSR}$$

- Choose parameter value(s) which give the smallest value of "fit" (if doing several fits).

- Note that Berkeley Madonna fits to the cumulative numbers of cases. This isn't ideal, because the fitting is carried out using the equivalent of least squares, so that the fitting routine will aim to fit the "later" data points best (because in doing so, it can make the SSR as small as possible). There are ways around this, but it's a bit fiddly.

METHOD OF LEAST SQUARES

The advantage of the least squares method is that it is simple, intuitive and easy to set up in MS Excel or a programming language.

Disadvantages:

1. The SSR doesn't actually tell us how "good" a fit the model is to data (aside from the fact that we want it to be as small as possible!).
2. We aren't taking into account how uncertain we are about our data (observations). Every data point is given equal weight.

It can be shown (but won't be here) that (2) is "not a problem" if the errors around the observations are uncorrelated, have roughly equal variances and are (preferably) normally distributed.

If they obviously don't have equal variances, one way around is to weigh the observations.

METHOD OF LEAST SQUARES

Weighted sum of squares:

$$g(\mathbf{x}) = \sum_i w_i \left(E_i(\mathbf{x}) - O_i\right)^2$$

Possible choices of weights w_i:

1. Reciprocal of variance of O_i, $1/\sigma_i$ (if known).
2. Reciprocal of model values $1/E_i(\boldsymbol{x})$ – this is called the **Pearson chi-squared statistic**.

MAXIMUM LIKELIHOOD

Probability

If each serum from a panel has a probability of 0.5 of being seropositive, what is the **probability** of obtaining 5 seropositive sera in a random sample of 10?

Let X = number of seropositive sera obtained
p = probability of each serum being positive
n = total number of sera in the sample

X has a **binomial distribution.** The probability it takes a particular value given some values of p and n, P(X|p,n), is given by:

$$P(X \mid p, n) = {}^nC_r\, p^r(1-p)^{n-r}$$

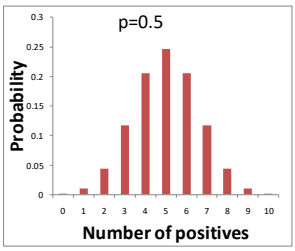
We write X ~ Bin(n,p).

This is the Binomial coefficient: ${}^nC_r = \frac{n!}{r!(n-r)!}$
It represents the number of ways of choosing r positive sera from a sample of n sera.

MAXIMUM LIKELIHOOD

Probability

If each serum from a panel has a probability of 0.5 of being seropositive, what is the **probability** of obtaining 5 seropositive sera in a random sample of 10?



If X = number of positive sera obtained, then

$P(X \mid p, n) = {}^nC_r\, p^r(1-p)^{n-r}$

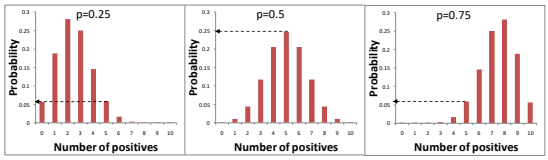
Here n=10 and p=0.5, so
$P(X=5 \mid p=0.5, n=10)$
$= {}^{10}C_5\,(0.5)^5(0.5)^5$
$=0.25$

MAXIMUM LIKELIHOOD

Likelihood

If we find 5 seropositive sera in a sample of 10, what is the **likelihood** that each serum has a probability of 0.5 of being positive?

To answer this question, think about the probability of obtaining 5 seropositive sera in a sample of 10 for different values of p...



$P(X=5|p=0.25) = 0.06$ $P(X=5|p=0.5) = 0.25$ $P(X=5|p=0.75) = 0.06$

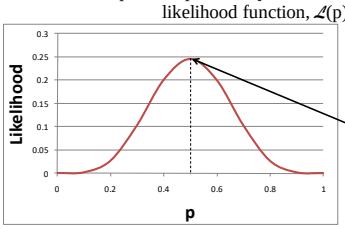
We can then plot this probability as a function of p – this is the likelihood function, $\mathcal{L}(p)$.

MAXIMUM LIKELIHOOD

Likelihood

If we find 5 seropositive sera in a sample of 10, what is the **likelihood** that each serum has a probability of 0.5 of being positive?

We can then plot this probability as a function of p – this is the likelihood function, $\mathcal{L}(p)$.



Notice that the likelihood function is maximised at p=0.5. We say that p=0.5 is the **maximum likelihood estimate** for X=5 (having 5 positive sera).

MAXIMUM LIKELIHOOD

Likelihood

If we find 5 seropositive sera in a sample of 10, what is the **likelihood** that each serum has a probability of 0.5 of being positive?

How do we calculate the likelihood function?

We know that X ~ Bin(10,p)

$$\mathcal{L}(p) = P(X=5|n=10,p) = {}^nC_r\, p^r(1-p)^{n-r} = {}^{10}C_5\, p^5(1-p)^5$$

Reminder: Binomial coefficient – number of ways of choosing 5 positive sera from a sample of 10 sera.

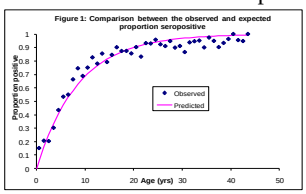
MAXIMUM LIKELIHOOD

Maximum likelihood estimation is a method of finding best fitting parameter values by maximising the likelihood function of the model given the observed data.

Note that if errors around observations are uncorrelated, have equal variances and are normally distributed, then the residual sum of squares is the maximum likelihood estimate for the model (not shown here).

MAXIMUM LIKELIHOOD

Back to the seroprevalence model.



This time we are trying to fit the force of infection, λ, so that model predictions of the number of positive samples at each age, $E_a(\lambda)$, is close to observed number of positive samples, O_a.

The number of positive samples in age group a is binomially distributed, i.e. $E_a(\lambda) \sim \text{Bin}(N_a, p_a)$.

Fixed number of trials (samples) N_a, each with fixed probability of being a success (positive) p_a.

$$\text{So } L(\lambda) = \prod_a L(\lambda, a)$$

$$= \prod_a P(\lambda \mid N_a, O_a)$$

$$= \prod_a {}^{N_a}C_{O_a}\, p_a^{O_a} (1-p_a)^{N_a - O_a} \text{ where } p_a = \frac{E_a(\lambda)}{N_a}$$

MAXIMUM LIKELIHOOD

The likelihood function tells us that we have found the "best fitting" parameters when we have maximised the likelihood.

However, a different model may allow us to get an even better fit to data.

One extreme is a model with a separate variable for every data point – this would reproduce the data perfectly (but at the cost of not giving us any useful insight at all!). Such a model is called a **saturated model**.

The **model deviance** compares the goodness of fit of a model with that of the saturated model, and is defined by:

$$D = -2\log\frac{L(\lambda)}{L(\hat{\lambda})} \quad \text{where} \quad \begin{array}{l} \text{D} = \text{deviance} \\ L(\lambda) = \text{model likelihood} \\ L(\hat{\lambda}) = \text{saturated likelihood} \end{array}$$

MAXIMUM LIKELIHOOD

Example: For the seroprevalence model we fitted in Practical 7:

Model log likelihood

$$\log L(\lambda) = \log \prod_a {}^{N_a}C_{O_a}\, p_a^{O_a} (1-p_a)^{N_a - O_a}$$

$$= \sum_a O_a \log p_a + (N_a - O_a)\log(1-p_a) + \log {}^{N_a}C_{O_a}$$

where $p_a = E_a(\lambda)/N_a$

Constants will cancel out

Saturated log likelihood

$$\log L(\hat{\lambda}) = \log \prod_a {}^{N_a}C_{O_a}\, \hat{p}_a^{O_a} (1-\hat{p}_a)^{N_a - O_a}$$

$$= \sum_a O_a \log \hat{p}_a + (N_a - O_a)\log(1-\hat{p}_a) + \log {}^{N_a}C_{O_a}$$

where $\hat{p}_a = O_a / N_a$

Deviance

$$D = 2(\log L(\hat{\lambda}) - \log L(\lambda))$$

MAXIMUM LIKELIHOOD

Useful properties of the deviance

If we have two models M1 and M2 then we can find the difference

(deviance of M1) – (deviance of M2)

This is χ^2-distributed with degrees of freedom equal to

(number of parameters in M1) – (number of parameters in M2)

This can be used to:

1. Calculate confidence intervals around best fitting parameters.
2. Test hypotheses (are additional parameters in a more complex model significant?).
3. Choose the most parsimonious model using criteria such as the Akaike Information Criterion (AIC).

RECAP

By now you should be able to:

1. Explain the need to fit models to data.
2. Define key terms about model fitting/calibration and sensitivity analysis.
3. Explain the principles behind the least squares, weighted least squares and maximum likelihood methods of fitting.
4. Implement these methods in MS Excel and Berkeley Madonna.

You will get opportunities to fit models to data later in the course. Don't worry – you won't need to understand the technicalities.

FURTHER READING (OPTIONAL)

These books may be useful if you want to explore the concepts that were introduced here in more depth.

- Hilborn and Mangel. The Ecological Detective. Chapter 5 deals with the least squares method, and chapter 7 with maximum likelihood estimation.
- Vanni et al. Pharmacoeconomics 2011; 29(1):35-49. Discusses various goodness of fit metrics and methods of numerical optimisation (goes further than we cover in this course).
- Kincaid and Cheney. Numerical analysis: mathematics of scientific computing. Chapter 11 discusses numerical optimisation.