Data Modeling

Patrice Koehl
Department of Biological Sciences
National University of Singapore

http://www.cs.ucdavis.edu/~koehl/Teaching/BL5229
koehl@cs.ucdavis.edu

Data Modeling

- Data Modeling: least squares
- Data Modeling: robust estimation

Data Modeling

- Data Modeling: least squares
- Data Modeling: robust estimation

Least squares

Suppose that we are fitting N data points (x_i, y_i) (with errors σ_i on each data point) to a model Y defined with M parameters a_j:

$$Y(x; a_1, a_2, \ldots, a_M)$$

The standard procedure is least squares: the fitted values for the parameters a_j are those that minimize:

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - Y(x; a_1, \ldots, a_M)}{\sigma_i} \right)^2$$

Where does this come from?

Model Fitting

Let us work out a simple example. Let us consider we have N students, $S_1, \ldots, S_N$ and let us "evaluate" a variable x_i for each student such that:

$x_i = 1$ if student S_i owns a Ferrari, and $x_i = 0$ otherwise.

We want an estimator of the probability p that a student owns a Ferrari.

The probability of observing x_i for student S_i is given by:

$$f(x_i, p) = p^{x_i} (1-p)^{1-x_i}$$

The likelihood of observing the values x_i for all N students is:

$$L(p) = f(x_1, \ldots x_N; p) \approx f(x_1; p) \ldots f(x_N; p)$$

Model Fitting

$$L(p) = p^{\sum x_i} (1-p)^{n - \sum x_i}$$

The maximum likelihood estimator of p is the value pm that maximizes L(p):

$$p_m = \underset{p}{\operatorname{argmax}}\, L(p)$$

This is equivalent to maximizing the logarithm of L(p) (log-likelihood):

$$\log(L(p)) = \log(p) \sum_{i=1}^{N} x_i + \log(1-p) \left(n - \sum_{i=1}^{N} x_i \right)$$

Model Fitting

$$\frac{\partial \log(L(p))}{\partial p} = \left(\frac{1}{p}\right)\sum_{i=1}^{N} x_i - \left(\frac{1}{1-p}\right)\left(n - \sum_{i=1}^{N} x_i\right) = 0$$

Multiplying by *p(1-p)*:

$$(1-p_m)\sum_{i=1}^{N} x_i - p_m\left(n - \sum_{i=1}^{N} x_i\right) = 0$$

$$\sum_{i=1}^{N} x_i - p_m \sum_{i=1}^{N} x_i - p_m n + p_m \sum_{i=1}^{N} x_i = 0$$

$$p_m = \frac{\sum_{i=1}^{N} x_i}{n}$$

This is the most intuitive value… and it matches with the maximum likelihood estimator.

Maximum Likelihood Estimators

Let us suppose that:

- The data points are independent of each other
- Each data point has a measurement error that is random, distributed as a Gaussian distribution around the "true" value $Y(x_i)$:

$$f(y_i;Y) = \exp\left[-\frac{1}{2}\left(\frac{y_i - Y(x_i)}{\sigma_i}\right)^2\right]$$

The likelihood function is:

$$L(Y) = f(y_1,\ldots,y_N;Y) \approx f(y_1;Y)\ldots f(y_N;Y)$$

$$L(Y) = \prod_{i=1}^{N}\left\{\exp\left[-\frac{1}{2}\left(\frac{y_i - Y(x_i)}{\sigma_i}\right)^2\right]\right\}$$

A Bayesian approach

Let us suppose that:

- The data points are independent of each other
- Each data point has a measurement error that is random, distributed as a Gaussian distribution around the "true" value $Y(x_i)$

The probability of the data points, *given the model Y* is then:

$$P(data/Model) \propto \prod_{i=1}^{N}\left\{\exp\left[-\frac{1}{2}\left(\frac{y_i - Y(x_i)}{\sigma_i}\right)^2\right]\right\}$$

A Bayesian approach

Application of Bayes 's theorem:

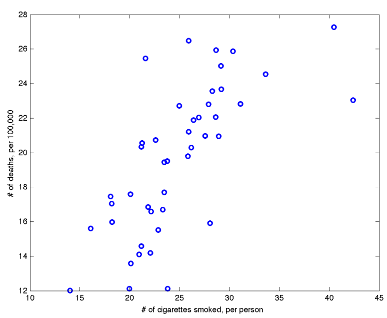
$$P(Model/Data) \propto P(Data/Model)P(Model)$$

With no information on the models, we can assume that the prior probability P(Model) is constant.

Finding the coefficients a1,…aM that maximizes P(Model/Data) is then equivalent to finding the coefficients that maximizes P(Data/Model).
This is equivalent to maximizing its logarithm, or minimizing the negative of its logarithm, namely:

$$\sum_{i=1}^{N} \frac{1}{2}\left(\frac{y_i - Y(x)}{\sigma_i}\right)^2$$

Fitting data to a straight line



Fitting data to a straight line

This is the simplest case:

$$Y(x) = ax + b$$

Then:

$$\chi^2 = \sum_{i=1}^{N}\left(\frac{y_i - ax_i - b}{\sigma_i}\right)^2$$

The parameters *a* and *b* are obtained from the two equations:

$$\frac{\partial \chi^2}{\partial a} = 0 = -2\sum_{i=1}^{N} \frac{x_i(y_i - ax_i - b)}{\sigma_i^2}$$

$$\frac{\partial \chi^2}{\partial b} = 0 = -2\sum_{i=1}^{N} \frac{y_i - ax_i - b}{\sigma_i^2}$$

Fitting data to a straight line

Let us define:

$$S=\sum_{i=1}^{N}\frac{1}{\sigma_i^2}\quad S_x=\sum_{i=1}^{N}\frac{x_i}{\sigma_i^2}\quad S_y=\sum_{i=1}^{N}\frac{y_i}{\sigma_i^2}\quad S_{xx}=\sum_{i=1}^{N}\frac{x_i^{\,2}}{\sigma_i^2}\quad S_{xy}=\sum_{i=1}^{N}\frac{x_i y_i}{\sigma_i^2}$$

then

$$\begin{aligned} aS_{xx}+bS_x &= S_{xy} \\ aS_x+bS &= S_y \end{aligned}$$

a and b are given by:

$$\boxed{\begin{aligned} a &= \frac{S_{xy}S-S_xS_y}{S_{xx}S-S_x^2} \\ b &= \frac{S_{xx}S_y-S_xS_{xy}}{S_{xx}S-S_x^2} \end{aligned}}$$

Fitting data to a straight line

We are not done!

Uncertainty on the values of a and b:

$$\boxed{\begin{aligned} \sigma_a^{\,2} &= \frac{S}{SS_{xx}-S_x^2} \\ \sigma_b^{\,2} &= \frac{S_x}{SS_{xx}-S_x^2} \end{aligned}}$$

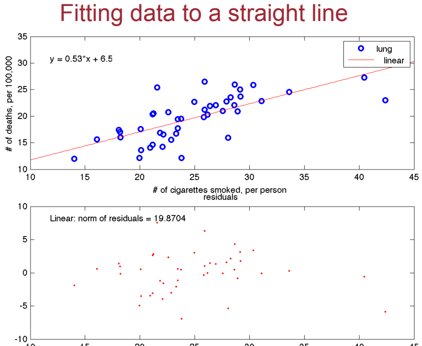
Evaluate goodness of fit:

-Compute $\chi 2$ and compare to N-M (here N-2)

-Compute residual error on each data point: $Y(x_i)-y_i$

-Compute correlation coefficient R^2

Fitting data to a straight line



General Least Squares

$$Y(x) = a_1 X_1(x) + a_2 X_2(x) + \ldots + a_M X_M(x)$$

Then:

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - a_1 X_1(x_i) - \ldots - a_M X_M(x_i)}{\sigma_i} \right)^2$$

The minimization of $\chi 2$ occurs when the derivatives of $\chi 2$ with respect to the parameters $a_1, \ldots a_M$ are 0. This leads to M equations:

$$\frac{\partial \chi^2}{\partial a_k} = \sum_{i=1}^{N} \frac{1}{\sigma_i} \left(y_i - a_1 X_1(x_i) - \ldots - a_M X_M(x_i) \right) X_k(x_i) = 0$$

General Least Squares

Define design matrix A such that $A_{ij} = \dfrac{X_j(x_i)}{\sigma_i}$

$$\begin{array}{cc} & \xleftarrow{\quad} \text{basis functions} \xrightarrow{\quad} \\ & \begin{array}{cccc} X_1(\ ) & X_2(\ ) & \cdots & X_M(\ ) \end{array} \\ \text{data points} \begin{array}{c} x_1 \\ x_2 \\ \vdots \\ x_N \end{array} & \begin{pmatrix} \frac{X_1(x_1)}{\sigma_1} & \frac{X_2(x_1)}{\sigma_1} & \cdots & \frac{X_M(x_1)}{\sigma_1} \\ \frac{X_1(x_2)}{\sigma_2} & \frac{X_2(x_2)}{\sigma_2} & \cdots & \frac{X_M(x_2)}{\sigma_2} \\ \vdots & & & \vdots \\ \frac{X_1(x_N)}{\sigma_N} & \frac{X_2(x_N)}{\sigma_N} & \cdots & \frac{X_M(x_N)}{\sigma_N} \end{pmatrix} \end{array}$$

General Least Squares

Define two vectors ***b*** and ***a*** such that $b_i = \dfrac{y_i}{\sigma_i}$

and ***a*** contains the parameters

Note that $\chi 2$ can be rewritten as:

$$\chi^2 = \left| A\boldsymbol{a} - \boldsymbol{b} \right|^2$$

The parameters ***a*** that minimize $\chi 2$ satisfy:

$$\left(A^T A \right) \boldsymbol{a} = A^T \boldsymbol{b}$$

These are the **normal equations** for the linear least square problem.

General Least Squares

How to solve a general least square problem:

1) Build the design matrix A and the vector b

2) Find parameters $a_1, \ldots a_M$ that minimize

$$\chi^2 = |A\boldsymbol{a} - \boldsymbol{b}|^2$$

(usually solve the normal equations)

3) Compute uncertainty on each parameter a_j:

if $C = A^T A$, then

$$\sigma(a_j)^2 = C^{-1}(j,j)$$

Data Modeling

- Data Modeling: least squares
- Data Modeling: robust estimation

Robust estimation of parameters

Least squares modeling assume a Gaussian statistics for the experimental data points; this may not always be true however. There are other possible distributions that may lead to better models in some cases.

One of the most popular alternatives is to use a distribution of the form:

$$\rho(x) = e^{-|x|}$$

Let us look again at the simple case of fitting a straight line in a set of data points (t_i, Y_i), which is now written as finding a and b that minimize:

$$Z(a,b) = \sum_{i=1}^{N} |Y_i - at_i - b|$$

b = median(Y-at) and a is found by non linear minimization

Robust estimation of parameters