Probability Theory & Statistics

Nonlinear Regression

Curve Fitting

Consider a set of m data points, (x_1, y_1), (x_2, y_2) $\ldots$, (x_m, y_m) and a model function

$$\widehat{y} = f(x, \boldsymbol{\beta}), \tag{1}$$

that depends on an independent variable x and a set of n parameters, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_n)$, with $m \geq n$, where $\boldsymbol{\beta}$ is a *vector* of parameters. The hat notation $\widehat{\cdot}$ denotes that $\widehat{y}$ is an *estimate of* y.

Curve fitting corresponds to finding a vector of parameters that best fits the curve given the data, i.e., $\widehat{y}$ is a best estimate of y.

A common measure of best fit is in the *least-squares* sense, where the sum of squared residuals

$$S = \sum_{i=1}^{m} r_i^2 \tag{2}$$

is minimized, given the residuals (the *in-sample prediction errors*)

$$r_i = y_i - f(x_i, \boldsymbol{\beta}) \tag{3}$$

for $i = 1, 2, \ldots, m$.

The minimum value of S occurs when the gradient is zero. (See the example below for a numerical example). Since the model contains n parameters there are n gradient equations:[1]

$$\frac{\partial S}{\partial \beta_j} = 2 \sum_i r_i \frac{\partial r_i}{\partial \beta_j} = 0, \qquad (j = 1, 2, \ldots, n). \tag{4}$$

In a nonlinear system, the derivatives $\partial r_i / \partial \beta_j$ are functions of both the independent variable *and* the parameters, so don't typically have closed-form solutions.[2] Instead, numerical methods must be employed with initial values provided for the parameters.

The parameters are iteratively refined by successive approximations.

$$\beta_j \approx \beta_j^{k+1} = \beta_j^k + \Delta \beta_j \tag{5}$$

where k is the kth successive iteration. The problem can be reformulated as a system of n simultaneous linear equations that form the basis for a Gauss-Newton solver for nonlinear least squares problems.

[1] Can you derive this equation?

[2] Why?

Convergence Criteria

The iterative method isn't guaranteed to find a minimum, nor is it guaranteed to even decrease from one iteration to the next. One possible criterion for convergence is,

$$\left| \frac{S^k - S^{k+1}}{S^k} \right| < a \tag{6}$$

where a can be chosen somewhat arbitrarily, e.g., setting $a = 0.001$ is a common value. Many other criteria exist for stopping the algorithm. Because the method isn't always guaranteed to converge, a limit is usually set on k.

Second-order Polynomial Fitting

Consider the following dataset consisting of four response datapoints, y_1, y_2, y_3, y_4 measured at four concentrations, x_1, x_2, x_3, x_4, in a dose-response experiment.

x_i	y_i	r_i
1	6	$6 - \beta_1(1)^2$
2	5	$5 - \beta_1(2)^2$
3	7	$7 - \beta_1(3)^2$
4	10	$10 - \beta_1(4)^2$

Table 1: Hypothetical measurements from a dose-response experiment.

The response model function is assumed to be a second-order polynomial function of the concentration,[3]

$$y = \beta_3 + \beta_2 x + \beta_1 x^2 \tag{7}$$

with β_3 and β_2 further set to zero in this experiment. Table 1 calculates the residuals for this model function. Substituting the residuals into the equation for the sum of squared residuals gives,

$$S(\beta_1) = (6 - \beta_1)^2 + (5 - 4\beta_1)^2 + (7 - 9\beta_1)^2 + (10 - 16\beta_1)^2 \tag{8}$$

that has a single partial derivative,

$$\frac{\partial S}{\partial \beta_1} = 708\beta_1 - 498. \tag{9}$$

[3]Although the independent variables are nonlinear, f is a linear function of the *parameters*, corresponding to a linear regression as opposed to a nonlinear regression. The example is nonetheless still useful for understanding least-squares methods.

Setting the partial derivative to zero and solving yields

$$\beta_1 = \frac{498}{708} = 0.7033898\ldots \tag{10}$$

giving

$$\widehat{y} = 0.703x^2 \tag{11}$$

the best model fit in a least-squares sense.

Figure **??** plots the hypothetical data and the best fit in a least-squares sense assuming a quadratic model function. In the supporting R code we compare equation **??** to R's built in (nonlinear) least-squares numerical optimizer with starting value $\beta_1 = 1$.

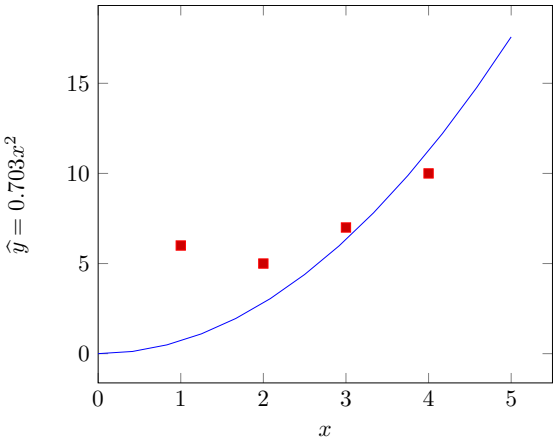

Figure 1: Plot of example data and computed best fit.

Matrix Notation

It's often useful to represent fitting and optimization problems as *matrices*. In matrix notation, the equations without residuals are $\boldsymbol{y} = \boldsymbol{X}\beta$ with

$$\boldsymbol{y} = [6, 5, 7, 10]^T, \qquad \boldsymbol{X} = [1, 4, 9, 16]^T \tag{12}$$

where $[\cdot]^T$ represents a *matrix transpose* and $\beta = [\beta_1]$ is a *scalar*, i.e., just a number.

Solving for β yields the estimate,

$$\widehat{\beta} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y} = [0.703], \tag{13}$$

In the supplementary R code we solve this system using R's base matrix operations, yielding $\beta = 0.703$ in agreement with the other calculations.

Uncertainty

Definition

The problem stated so far has been a *deterministic* algorithm for minimizing the sum of squared residuals. The parameters that optimize the problem produce a best fit curve in a least-squares sense. The *point estimates* calculated in the fitting (each change in the independent variable x results in a single point $y = 0.703x$ in the curve fit) don't provide a measure of uncertainty, i.e., there is a single curve, and we are completely confident that it is the best curve.

But are we really completely certain the "best fit" really is the best fit? Instead of a point estimate, it would be useful if could instead provide an interval of fit, with a probability associated to that interval, conditional on the value of the independent variable x.

Symbolically we write this as,

$$P\left(a \leq Y(X) \leq b \mid X = x\right) \in [0, 1] \tag{14}$$

where the *response*, $Y = y + Z$, is now understood as the true model function, y, plus some uncertainty, Z, related to a random or deterministic *covariate*, X. (The independent variable is now referred to as the covariate because "independent" has a special meaning in statistics.)

The value of the probability function, P, depends on the size of the interval, the value of the covariate, and the distribution of the uncertainty Z, the so-called "noise". Rather than calculate the probability given the interval and the distribution of the noise, the probability is usually first defined, and the intervals then calculated conditional on the value of x, i.e.,

$$P^{-1}(a \leq Y \leq b \mid X = x) = p_\alpha = 1 - \alpha \tag{15}$$

where α is referred to as a *probability threshold* or *degree of confidence*. But how do we calculate the intervals?

Linear Regression

It's useful to review some ideas in *linear regression* before moving on to more complex methods required to calculate probability intervals for models with *nonlinear functions* of their parameters. Linear regression provides estimates and other inferential results for the parameters $\boldsymbol{\beta}$ in models of the form,[4]

$$Y_i = f\left(x, \boldsymbol{\beta}\right) + Z_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_P x_{iP} + Z_i = \left(x_{i1}, \ldots, x_{iP}\right)\boldsymbol{\beta} + Z_i \tag{16}$$

where Y_i is the random variable representing the ith response, $i = 1, 2, \ldots, N$, composed of a *deterministic* and now *stochastic* part. The deterministic part, $\left(x_{i1}, \ldots, x_{iP}\right)\boldsymbol{\beta}$,

[4] In the previous example we were in fact dealing with a linear regression with a single scalar parameter.

depends on the parameters $\boldsymbol{\beta}$ and the *covariates*, x_{ip}, where $p = 1,\ 2, \ldots, P$ are the number of parameters in the linear regression; the stochastic part is represented by the random variable Z_i corresponding to a disturbance that perturbs the response.

For N observations, the model can be written as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z} \tag{17}$$

where $\boldsymbol{Y}$ is a *random vector* representing the data, and $\boldsymbol{X}$ is an $N \times P$ matrix of covariates,

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & \cdots & x_{1P} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{NP} \end{bmatrix} \tag{18}$$

The term $\boldsymbol{X}\boldsymbol{\beta}$ is the *model function* for the responses. Since a non-zero mean can always be incorporated into the model function, we can also assume that,

$$E(\boldsymbol{Z}) = 0 \tag{19}$$

or, equivalently,

$$E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta} \tag{20}$$

where $E(\cdot)$ is the *expected value* (see Appendix). Given the second definition, the term $\boldsymbol{X}\boldsymbol{\beta}$ is sometimes referred to as the *expectation function*. If we further assume that $\boldsymbol{Z}$ is normally distributed with

$$Var(\boldsymbol{Z}) = E\left(\boldsymbol{Z}\boldsymbol{Z}^T\right) = \sigma^2 \boldsymbol{I} \tag{21}$$

where $\boldsymbol{I}$ is the $N \times N$ identity matrix, then the joint probability density function for $\boldsymbol{Y}$ given $\boldsymbol{\beta}$ and the *variance* σ^2 is,

$$p(\boldsymbol{y} \mid \boldsymbol{\beta},\ \sigma^2) = \left(2\pi\sigma^2\right)^{-N/2} \exp\left(\frac{-\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2}{2\sigma^2}\right) \tag{22}$$

where $\|\cdot\|$ denotes the length of a vector. Given a derivative matrix X and a vector of observed data, $\boldsymbol{y}$, we aim to make inferences about σ^2 and the P parameters $\boldsymbol{\beta}$.

Note, the distinction between $\{X,\ Y\}$ and $\{x,\ y\}$ is important. Lower-case letters are used to denote observed or known data, while capital letters are used to denote random variables (think of this as before we observe the data). The covariate is thought of as *fixed* or *observed*, but the error term is a random variable. The latter implies that the outcome is also a random variable.

Least Squares Estimates

The *likelihood function* or the *likelihood*, $L(\boldsymbol{\beta},\ \sigma \mid \boldsymbol{y})$ is regarded as a function of the parameters conditional on the observed data, rather than a function of responses conditional on the values of the parameters. Suppressing the constant $(2\pi)^{-N/2}$, we write

$$L(\boldsymbol{\beta},\ \sigma \mid \boldsymbol{y}) \propto \sigma^{-N} \exp\left(\frac{-\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2}{2\sigma^2}\right) \tag{23}$$

The likelihood is maximized with respect to $\boldsymbol{\beta}$ when the *residual sum of squares,*

$$S(\boldsymbol{\beta}) = \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2 = \sum_{i=1}^{N}\left(y_i - \left(\sum_{i}^{j} x_{ip}\beta_p\right)\right)^2 \tag{24}$$

is a minimum. Thus, the *maximum likelihood estimate* $\widehat{\boldsymbol{\beta}}$ is the value of $\boldsymbol{\beta}$ which minimizes $S(\boldsymbol{\beta})$. This is called the *least squares estimate* and can be written as

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{25}$$

This is the same result we proved in the deterministic sections. A least squares estimate is only appropriate under the following conditions:

1. The expectation function is correct, i.e., of the form $(x_{i1},\ldots,x_{iP})\boldsymbol{\beta}$
2. The response is represented as a function plus a disturbance, i.e., $Y = f(\boldsymbol{x},\ \boldsymbol{\beta}) + Z$
3. The noise is independent of the expectation function.
4. Each noise term is normally distributed.
5. Each noise term has zero mean.
6. The noise terms have equal variances.
7. The noise terms are independently distributed.

If all of these conditions are satisfied, we can derive several results described in the next section.

Sampling Theory Inference Results

The least squares estimator has a number of useful properties [**?**]:

1. The least squares estimator $\widehat{\boldsymbol{\beta}}$ is normally distributed. This follows because the estimator is a linear function of $\boldsymbol{Y}$ which in turn is a linear function of $\boldsymbol{Z}$. Since $\boldsymbol{Z}$ is assumed to be normally distributed, $\widehat{\boldsymbol{\beta}}$ is also normally distributed.
2. $E\left(\widehat{\boldsymbol{\beta}}\right) = \boldsymbol{\beta}$ so, the least squares estimator is unbiased.

3. $Var\left(\widehat{\boldsymbol{\beta}}\right) = \sigma^2\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$: the covariance matrix of the least squares estimator depends on the variance of the noise, σ^2 and the derivative matrix $\boldsymbol{X}$.

4. A $1-\alpha$ *joint confidence region* for the parameter vector $\boldsymbol{\beta}$ is the ellipsoid

$$\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)^T\boldsymbol{X}^T\boldsymbol{X}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right) \leq Ps^2F(P,\ N-P;\alpha) \tag{26}$$

where

$$s^2 = \frac{S(\widehat{\boldsymbol{\beta}})}{N-P} \tag{27}$$

is the residual mean square or variance estimate based on $N-P$ degrees of freedom, and $F(P, N-P;\alpha)$ is the upper α quantile for Fisher's F-distribution with P and $N-P$ degrees of freedom.

5. A $1-\alpha$ *confidence band* for the response function at any $\boldsymbol{x}$ is given by,

$$\boldsymbol{x}^T\widehat{\boldsymbol{\beta}} \pm s\sqrt{\boldsymbol{x}^T\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{x}}\cdot\sqrt{P\cdot F(P,N-P;\alpha)} \tag{28}$$

In the next Section we use some measured data to compute parameter estimates and joint and marginal inference regions.

Worked Example

Using the following transformed PCB data

The regions are summarized with $\widehat{\boldsymbol{\beta}}$, s^2, $\boldsymbol{X}^T\boldsymbol{X}$, and $\nu = N-P$. For the PCB data we have $\widehat{\boldsymbol{\beta}} = (-2.391,\ 2.300)^T$, $s^2 = 0.246$, and $\nu = 26$ degrees of freedom.

Hence,

$$X^TX = \begin{bmatrix} 28.000 & 46.941 \\ 46.941 & 83.367 \end{bmatrix} \tag{29}$$

and

$$\left(X^TX\right)^{-1} = \begin{bmatrix} 0.6374 & -0.3589 \\ -0.3589 & 0.2141 \end{bmatrix} \tag{30}$$

giving the *joint 95% inference region* as

$$28.00\,(\beta_1+2.391)^2 + 93.88\,(\beta_1+2.391)\,(\beta_2-2.300) + 83.37\,(\beta_2-2.300)^2 = 1.66 \tag{31}$$

The marginal 95% inference interval for the parameter β_1 is,

$$-2.391 \pm (0.496)\sqrt{0.6374}\ (2.056) \tag{32}$$

or

$$-3.21 \leq \beta_1 \leq -1.58 \tag{33}$$

and the marginal 95% inference interval for the parameter β_2 is,

$$2.300 \pm (0.496)\sqrt{0.2141}\ (2.056) \tag{34}$$

or

$$1.83 \leq \beta_2 \leq 2.77 \tag{35}$$

The 95% *inference band* for any value x is given by,

$$-2.391 + 2.300x \pm (0.496)\sqrt{0.637 - 0.718x + 0.214x^2} \cdot \sqrt{2(3.37)} \tag{36}$$

The results are plot in the following Figures. The R code is added as supplementary material.

Sigmoid Curve Fitting

Consider the following *sigmoid* model function,

$$y = f(x) = \frac{L}{1 + \theta_1 e^{\theta_2 x}} + b \tag{37}$$

Often used to fit dose-response curves with $L = 100$, and $b = 0$. The following Figure plots a sigmoidal function with $\theta_1 =$ and $\theta_2 =$.

Once fit to data, e.g., with a least-squares algorithm, the fitting curve is used to infer lethal concentrations,

$$y^{-1}(x) = 50 \tag{38}$$

This can be calculated directly as

$$x = \frac{1}{\theta_2} \cdot \log\left(\frac{1}{\theta_1}\left(\frac{L}{y} - 1\right)\right) \tag{39}$$

The sigmoid function can be transformed into a linear model,

$$z = \log\left(\frac{L}{y} - 1\right) = \theta_2 x + \log(\theta_1) = \beta_2 x + \beta_1 \tag{40}$$

where $\beta_1 = \log(\theta_1)$, it is *transformably linear*. This means we can easily calculate an initial guess for the parameters using the linear regression algorithm described in the first section. Note, although initial estimates of the parameters are possible, we cannot calculate prediction bands because the transformation breaks the property condition 6

requiring the noise terms have equal variance. The following figures show a sigmoid function $f(x,\ \theta) + \varepsilon$, and its transform. The transformed function clearly has larger variances at the ends of the function.

In the next section the *delta method* provides an approximation of the prediction band for models with nonlinear functions of parameters. Nonlinear model functions are defined as models with partial derivatives that contain at least ...

$$\frac{\partial}{\partial \theta_1} f(x,\ \theta) = \tag{41}$$

and, ..

Simulated dose-response,

Transformed into a linear coordinate system,

Delta Method

Bates & Watts (1988) provide a method to calculate confidence and prediction intervals for model functions that are nonlinear in their parameters. In this section we provide some background results for understanding the delta method and then show how it is used to calculate prediction bands.

The Geometry of Least Squares

Appendix

Answers

1) The constraint on the number of data points, m, given the number of parameters, n, is due to the requirement that the system is *consistent*, i.e., there are at least as many or more equations than unknowns (here the unknowns are the unknown parameter values).

2) Show that,

$$\frac{\partial S}{\partial \beta_j} = 2 \sum_i r_i \frac{\partial r_i}{\partial \beta_j} = 0 \tag{42}$$

Starting with the definition of S

$$S = \sum_{i=1}^{m} r_i^2 \tag{43}$$

The partial derivative is,

$$\frac{\partial S}{\partial \beta_j} = \frac{\partial}{\partial \beta_j} \sum_{i=1}^{m} r_i^2 = \sum_{i=1}^{m} \frac{\partial}{\partial \beta_j} r_i^2 . \tag{44}$$

Noting that, by linearity of the summation operator, each term can be considered independently as a function of the parameter vector,

$$z = r^2(\boldsymbol{\beta}) \tag{45}$$

so that,

$$z = f(u) = u^2 \tag{46}$$

and

$$u = g(\boldsymbol{\beta}) \tag{47}$$

giving

$$\frac{\partial z}{\partial u} = f'(u) = 2u \tag{48}$$

and,

$$\frac{\partial u}{\partial \boldsymbol{\beta}} = g'(\boldsymbol{\beta}) . \tag{49}$$

By the product rule

$$\frac{\partial z}{\partial \boldsymbol{\beta}} = \frac{\partial z}{\partial u} \cdot \frac{\partial u}{\partial \boldsymbol{\beta}} = 2u \cdot \frac{\partial u}{\partial \boldsymbol{\beta}} \tag{50}$$

the partial derivatives can be written as the gradient equations,

$$\frac{\partial S}{\partial \beta_j} = 2 \sum_i r_i \frac{\partial r_i}{\partial \beta_j} \tag{51}$$

3) Very few partial differential equations (PDEs) have a known closed-form solution, so unless f is chosen to satisfy one of those known solutions, numerical methods are the only possible way to find a solution.

4) TBD

Probability Theory Recap

Expectations

Define and describe the properties of mathematical expectations.

Variances

Same for variances.

Proof of Sampling Theory Inference Results

Definition

Start with the definition of the simple linear regression method.

Method of Least Squares

Given the definition, an inference can be obtained with a least squares approach, essentially unconditional distributions with sample mean and variances feeding into the parameter estimations.

Maximum Likelihood Estimation

A special case of the simple linear regression is when we assume residuals are Normal. In this case, a few results can be achieved, including an ML estimate of the parameters.

Linear Regression in Matrix Format

Discuss how to represent linear regressions as matrices and how derive some equivalent results.