Chapter 7

Least Squares Estimation

7.1. Introduction

Least squares is a time-honored estimation procedure, that was developed independently by Gauss (1795), Legendre (1805) and Adrain (1808) and published in the first decade of the nineteenth century. It is perhaps the most widely used technique in geophysical data analysis. Unlike maximum likelihood, which can be applied to any problem for which we know the general form of the joint pdf, in least squares the parameters to be estimated must arise in expressions for the means of the observations. When the parameters appear linearly in these expressions then the least squares estimation problem can be solved in closed form, and it is relatively straightforward to derive the statistical properties for the resulting parameter estimates.

One very simple example which we will treat in some detail in order to illustrate the more general problem is that of fitting a straight line to a collection of pairs of observations (x_i, y_i) where $i = 1, 2, \ldots, n$. We suppose that a reasonable model is of the form

$$y = \beta_0 + \beta_1 x, \tag{1}$$

and we need a mechanism for determining β_0 and β_1. This is of course just a special case of many more general problems including fitting a polynomial of order p, for which one would need to find $p + 1$ coefficients. The most commonly used method for finding a model is that of least squares estimation. It is supposed that x is an **independent** (or predictor) variable which is known exactly, while y is a **dependent** (or response) variable. The least squares (LS) estimates for β_0 and β_1 are those for which the predicted values of the curve minimize the sum of the squared deviations from the observations. That is the problem is to find the values of β_0, β_1 that minimize the residual sum of squares

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{2}$$

Note that this involves the minimization of vertical deviations from the line (not the perpendicular distance) and is thus not symmetric in y and x. In other words if x is treated as the dependent variable instead of y one might well expect a different result.

To find the minimizing values of β_i in (2) we just solve the equations resulting from setting

$$\frac{\partial S}{\partial \beta_0} = 0, \qquad \frac{\partial S}{\partial \beta_1} = 0, \tag{3}$$

namely

$$\begin{aligned} \sum_i y_i &= n\hat{\beta}_0 + \hat{\beta}_1 \sum_i x_i \\ \sum_i x_i y_i &= \hat{\beta}_0 \sum_i x_i + \hat{\beta}_1 \sum_i x_i^2 \end{aligned} \tag{4}$$

Solving for the $\hat{\beta}_i$ yields the least squares parameter estimates:

$$\begin{aligned}\hat{\beta}_0 &= \frac{\sum x_i^2 \sum_i y_i - \sum x_i \sum x_i y_i}{n \sum x_i^2 - (\sum x_i)^2} \\ \hat{\beta}_1 &= \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}\end{aligned} \tag{5}$$

where the $\sum$'s are implicitly taken to be from $i = 1$ to n in each case. Having generated these estimates, it is natural to wonder how much faith we should have in $\hat{\beta}_0$ and $\hat{\beta}_1$, and whether the fit to the data is reasonable. Perhaps a different functional form would provide a more appropriate fit to the observations, for example, involving a series of independent variables, so that

$$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \tag{6}$$

or decay curves

$$f(t) = Ae^{-\alpha t} + Be^{-\beta t}, \tag{7}$$

or periodic functions

$$f(t) = A\cos\omega_1 t + B\sin\omega_1 t + C\cos\omega_2 t + D\sin\omega_2 t. \tag{8}$$

In equations (7) and (8) the functions $f(t)$ are linear in A, B, C and D, but **nonlinear** in the other parameters α, β, ω_1, and ω_2. When the function to be fit is linear in the parameters, then the partial derivatives of S with respect to them yield equations that can be solved in closed form. Typically non-linear least squares problems do not provide a solution in closed form and one must resort to an iterative procedure. However, it is sometimes possible to transform the nonlinear function to be fitted into a linear form. For example, the Arrhenius equation models the rate of a chemical reaction as a function of temperature via a 2-parameter model with an unknown constant frequency factor C and activation energy E_A, so that

$$\alpha(T) = Ce^{-E_A/kT} \tag{9}$$

Boltzmann's constant, k is known *a priori*. If one measures α at various values of T, then C and E_A can be found by a linear least squares fit to the transformed variables, $\log \alpha$ and $\frac{1}{T}$:

$$\log \alpha(T) = \log C \; - \; \frac{E_A}{kT} \tag{10}$$

7.2. Fitting a Straight Line

We return to the simplest of LS fitting problems, namely fitting a straight line to paired observations (x_i, y_i), so that we can consider the statistical properties of LS estimates, assess the goodness of fit in the resulting model, and understand how regression is related to correlation.

To make progress on these fronts we need to adopt some kind of statistical model for the noise associated with the measurements. In the **standard statistical model** (SSM) we suppose that y is a linear function of x plus some random noise,

$$y_i = \beta_0 + \beta_1 x_i + e_i \qquad\qquad i = 1, \ldots, n. \tag{11}$$

In (11) the values of x_i are taken to be fixed, while the e_i are independent random variables with $E(e_i) = 0$ and $Var(e_i) = \sigma^2$, but for the time being we make no further assumption about the exact distribution underlying the e_i.

Under the SSM it is straightforward to show that the LS estimate for a straight line is unbiased: that is $E[\beta_j] = \beta_j$. To do this for β_0 we make use of the fact that $E[y_i] = \beta_0 + \beta_1 x_i$, and take the expected value of β_0 in equation (5). This yields:

$$\begin{aligned} E[\hat{\beta}_0] &= \frac{\sum x_i^2 \sum_i E[y_i] - \sum x_i \sum x_i E[y_i]}{n \sum x_i^2 - (\sum x_i)^2} \\ &= \frac{\sum x_i^2 (n\beta_0 + \beta_1 \sum_i x_i) - \sum x_i (\beta_0 \sum x_i + \beta_1 \sum_i x_i^2)}{n \sum x_i^2 - (\sum x_i)^2} \\ &= \beta_0 \end{aligned} \tag{12}$$

A similar proof establishes that $E[\hat{\beta}_1] = \beta_1$. Note that this proof only uses $E[e_i] = 0$ and the fact that the errors are additive: we did not need them to be iid.

Under the SSM, $Var[y_i] = \sigma^2$ and $Cov[y_i, y_j] = 0$ for $i \neq j$. Making use of this it is possible (see Rice p. 513) to calculate the variances for $\hat{\beta}_i$ as

$$\begin{aligned} Var[\hat{\beta}_0] &= \frac{\sigma^2 \sum_i x_i^2}{n \sum x_i^2 - (\sum x_i)^2} \\ Var[\hat{\beta}_1] &= \frac{n\sigma^2}{n \sum x_i^2 - (\sum x_i)^2} \\ Cov[\hat{\beta}_0, \hat{\beta}_1] &= \frac{-\sigma^2 \sum_i x_i}{n \sum x_i^2 - (\sum x_i)^2} \end{aligned} \tag{13}$$

To show this we make use of the fact that equation (5) can be rewritten in the form:

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{\sum_i (x_i - \bar{x})(y_i)}{\sum_i (x_i - \bar{x})^2}$$

Then

$$Var[\hat{\beta}_1] = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}$$

Similarly for the other expressions in (13).

We see from (13) that the variances of the slope and intercept depend on x_i and σ^2. The x_i are known, so we just need a means of finding σ^2. In the SSM, $\sigma^2 = E[y_i - \beta_0 - \beta_1 x_i]$. So we can estimate σ^2 from the average squared deviations of data about the fitted line:

$$RSS = \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \tag{14}$$

We will see later that

$$s^2 = \frac{RSS}{n-2} \tag{15}$$

is an unbiased estimate of σ^2. The number of degrees of freedom is $n-2$ because 2 parameters have been estimated from the data. So our recipe for estimating $Var[\hat{\beta}_0]$ and $Var[\hat{\beta}_1]$ simply involves substituting s^2 for σ^2 in (13). We call these estimates $s^2_{\hat{\beta}_0}$ and $s^2_{\hat{\beta}_1}$, respectively.

When the e_i are independent normally distributed random variables then $\hat{\beta}_0, \hat{\beta}_1$ will be too, since they are just linear combinations of independent normal RV's. More generally if the e_i are independent and satisfy some not too demanding assumptions, then a version of the Central Limit Theorem will apply, and for large n, $\hat{\beta}_0$ and $\hat{\beta}_1$ are approximately normal RV's.

An immediate and important consequence of this is that we can invoke either exact or approximate confidence intervals and hypothesis tests based on the $\hat{\beta}_i$ being normally distributed. It can be shown that

$$\frac{\hat{\beta}_i - \beta_i}{s_{\hat{\beta}_i}} \sim t_{n-2} \tag{16}$$

and we can use the t-distribution to establish confidence intervals and for hypothesis testing. Perhaps the commonest application of hypothesis testing is in determining whether the β_i are significantly different from zero. If not there may be a case for excluding them from the model.

7.3. Assessing Fit

The most basic thing to do in assessing the fit is to use the residuals from the model, in this case:

$$\hat{e}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \tag{17}$$

They should be plotted as a function of x, which allows one to see systematic misfit or departures from the SSM. These may indicate the need for a more complex model or transformation of variables.

When the variance of errors is a constant independent of x then the errors are said to be **homoscedastic**, when the opposite is true they are **heteroscedastic**. Rice provides some good examples for this in Chapter 14 - see Figs 14.11 and 14.12. When the variance varies with x it is sometimes possible to find a transformation to correct the problem. For example, instead of $y = \beta x$ one could try $\sqrt{y} = \gamma\sqrt{x}$. Then $\hat{\beta} = \hat{\gamma}^2, \ldots$

A common scenario one might wish to test is whether the intercept is zero. This can be done by calculating both slope and intercept, and finding $s_{\hat{\gamma}_0}$. Then one could use the $t-$test on the hypothesis $H_0 : \gamma_0 = 0$ with

$$t = \frac{\hat{\gamma}_0}{s_{\hat{\gamma}_0}}$$

Another strategy in assessing fit is to look at the sample distribution of residuals, compared to a normal probability plot. Q-Q plots of the residuals can provide a visual means of assessing things like gross departures from normality or identifying outliers. LS estimates are not robust against outliers, which can have a large effect on the estimated coefficients, their standard errors and s. This is especially true if outliers correspond to extreme values of x.

7.4. Correlation and Regression

As must by now be obvious there is a close relationship between correlation and fitting straight lines. We define

$$\begin{aligned}
S_{xx} &= \frac{1}{n}\sum_i (x_i - \bar{x})^2 \\
S_{yy} &= \frac{1}{n}\sum_i (y_i - \bar{y})^2 \\
S_{xy} &= \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y})
\end{aligned} \tag{18}$$

The correlation coefficient r expressing the degree of linear correlation between x and y is

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}, \tag{19}$$

while the slope of the best fitting straight line is

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \tag{20}$$

implying that

$$r = \hat{\beta}_1 \sqrt{\frac{S_{xx}}{S_{yy}}}$$

Clearly zero correlation occurs only when the slope is zero. We can also see that if we calculate **standardized residuals**

$$u_i = \frac{x_i - \bar{x}}{\sqrt{S_{xx}}} \qquad v_i = \frac{y_i - \bar{y}}{\sqrt{S_{yy}}}, \tag{21}$$

then $S_{uu} = S_{vv} = 1$ and $S_{uv} = r$. In this standardized system $\beta_0 = \bar{v} - r\bar{u} = 0$, since u and v are centered on the mean values of the original variables, and the predicted values are just

$$\hat{v}_i = r u_i \tag{22}$$

(22) clearly describes the concept of regression toward mediocrity, originally espoused by Galton. When offsprings heights are paired with those of their parents, there is a tendency for larger (or smaller) than average parents to have offspring whose sizes are closer to the mean of the distribution.

7.5. The Matrix Approach to Least Squares

In more complex least squares problems there are substantial advantages to adopting an approach that exploits linear algebra. The notation is more compact and can provide theoretical insights as well as computational advantages of a purely practical nature. Programming packages like Matlab are an excellent example of this.

In least squares estimation the parameters to be estimated must arise in expressions for the means of the observations. That is for an observation y we can write:

$$\mathcal{E}(y) \;=\; \theta_1 x_1 \;+\; \theta_2 x_2 \;+\; \ldots + \theta_p x_p \tag{23}$$

where $x_1,\; x_2,\; \ldots, x_p$ are known and $\theta_1,\; \theta_2,\; \ldots, \theta_p$ are unknown parameters. Equivalently, we can write

$$y \;=\; \theta_1 x_1 \;+\; \theta_2 x_2 \;+\; \ldots \;+\; \theta_p x_p \;+\; \epsilon \tag{24}$$

where $\mathcal{E}(\epsilon) \;=\; 0$; ϵ is a random variable, usually regarded as a measurement error for y. We then have a **linear model** for the dependence of y on p other variables $x_1,\; x_2,\; \ldots, x_p$ whose values are assumed exactly known.

Now suppose that we measure y a total of n times yielding $y_i,\; i \;=\; 1,\; \ldots,\; n$ each time using a different set of values for $x_1,\; \ldots,\; x_p$, denoted by $x_{i1},\; x_{i2},\; \ldots,\; x_{i,p}$ for the ith experiment; we assume all these values for x are known. Then our expression becomes

$$y_i = \theta_1 x_{i1} + \theta_2 x_{i2} + \ldots + \theta_p x_{ip} + \epsilon_i \qquad i = 1, \ldots, n \tag{25}$$

with $\mathcal{E}(\epsilon_i) \;=\; 0$ for all i and also $n \;\geq\; p$.

How do we estimate the unknown parameters $\theta_1, \theta_2,\; \ldots, \theta_p$ using the observed values $\{y_i\}$ and the known values $\{x_{i1}\}, \ldots, \{x_{ip}\}, i = 1, \ldots, n$? The **principle of least squares** chooses as estimates of $\theta_1,\; \theta_2,\; \ldots, \theta_p$ those values $\hat{\theta}_1,\; \hat{\theta}_2, \ldots, \hat{\theta}_p$ which minimize

$$S(\theta_1, \theta_2, \ldots, \theta_p) = \sum_{i=1}^{n} \{y_i - \theta_1 x_{i1} - \theta_2 x_{i2} - \ldots - \theta_p x_{ip}\}^2 \tag{26}$$

We can find the solution to this problem as we did before for the straight line fitting problem, by simply setting $\frac{\partial S}{\partial \theta_i} = 0$ yielding the following:

$$\hat{\theta}_1 \sum_{i=1}^{n} x_{i1} x_{ik} + \hat{\theta}_2 \sum_{i=1}^{n} x_{i2} x_{ik} + \ldots \hat{\theta}_p \sum_{i=1}^{n} x_{ip} x_{ik} = \sum_{i=1}^{n} y_i x_{ik} \qquad k = 1, \ldots, p \tag{27}$$

But the index notation is a bit messy and we now want to push forward with developing the matrix approach, rewriting (27) in terms of vectors and matrices. We start again from (25),

$$\vec{Y} = X\vec{\theta} + \vec{\epsilon} \tag{28}$$

$\vec{Y} \in I\!R^n$, $\vec{\theta} \in I\!R^p$. X is an $n \times p$ matrix (not a random variable) and is known as the **design matrix**; its rows are the $x_1, x_2, \ldots, x_p$'s for each measurement of y:

$$\vec{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \vec{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{pmatrix} \qquad X \;=\; \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1,p} \\ x_{21} & x_{22} & \ldots & x_{2,p} \\ \vdots & \ddots & \vdots & \\ x_{n1} & x_{n2} & \ldots & x_{n,p} \end{pmatrix} \tag{29}$$

We can also define a **residual vector** $\vec{r} \in \mathbb{R}^n$

$$\vec{r} = \vec{Y} - X\vec{\theta} \tag{30}$$

We see that $S = \vec{r}.\vec{r}$ or the Euclidean length of $\vec{r}$. In other words the least-squares solution is the one with the smallest misfit to the measurements, as given by

$$\vec{r}^T\vec{r} = r_i r_i = \sum_{i=1}^{n} r_i^2 = \vec{r}.\vec{r} = \|\vec{r}\|_2^2 \tag{31}$$

We want $\vec{\theta} \;=\; \hat{\theta}$ such that

$$\nabla_{\vec{\theta}}[\vec{r}(\vec{\theta}).\vec{r}(\vec{\theta})] = \vec{0} \tag{32}$$

Equivalently,

$$\begin{aligned} \nabla_{\vec{\theta}}\left[(\vec{Y} - X\vec{\theta})^T(\vec{Y} - X\vec{\theta})\right] &= \vec{0} \\ \nabla_{\vec{\theta}}\left[\vec{Y}^T\vec{Y} - \vec{Y}^T X\vec{\theta} - \vec{\theta}^T X^T\vec{Y} + \vec{\theta}^T X^T X\vec{\theta}\right] &= \vec{0} \end{aligned} \tag{33}$$

Since $\vec{Y}^T X\vec{\theta} = \vec{\theta}^T X^T\vec{Y} = (X^T\vec{Y})^T\vec{\theta}$ this becomes

$$\nabla_{\vec{\theta}}\left[\vec{Y}^T\vec{Y} - 2(X^T\vec{Y})^T\vec{\theta} + \vec{\theta}^T X^T X\vec{\theta}\right] = 0 \tag{34}$$

whence

$$-2X^T\vec{Y} + 2X^T X\vec{\theta} = 0 \tag{35}$$

which is to say

$$\hat{\theta} \;=\; (X^T X)^{-1} X^T\vec{Y} \tag{36}$$

provided $(X^T X)^{-1}$ exists. These are the **normal equations**, the formal solution to the least-squares problem. As we will see later, constructing $(X^T X)$ can sometimes introduce considerable roundoff error so in practice it may not be desirable to solve the equations in this particular form. We discuss some alternative formulations later. Here we note that in order for a unique solution to exist for the normal equations, we require $(X^T X)$ to be non-singular: this will be the case if and only if the rank of X equals p.

The matrix notation is readily understood if we use as an example the straight line fitting from an earlier section. In this context (29) produces

$$\vec{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \vec{\theta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \qquad X \;=\; \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \tag{37}$$

We can form the normal equations as in (36) by

$$X^T X = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \tag{38}$$

yielding

$$X^T X = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix} \tag{39}$$

Inverting this we find

$$(X^T X)^{-1} = \frac{1}{n\sum_i x_i^2 - (\sum_i x_i)^2} \begin{pmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{pmatrix} \tag{40}$$

along with

$$X^T Y = \begin{pmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{pmatrix} \tag{41}$$

Now we get $\hat{\beta}$ using (36)

$$\begin{aligned} \hat{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} &= (X^T X)^{-1} X^T \vec{Y} \\ &= \frac{1}{n\sum_i x_i^2 - (\sum_i x_i)^2} \begin{bmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{bmatrix} \begin{bmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{bmatrix} \\ &= \frac{1}{n\sum_i x_i^2 - (\sum_i x_i)^2} \begin{bmatrix} (\sum_i y_i)(\sum_i x_i^2) & -(\sum_i x_i)(\sum_i x_i y_i) \\ n(\sum_i x_i y_i) & -(\sum_i x_i)(\sum_i y_i) \end{bmatrix} \end{aligned} \tag{42}$$

as we had before in (5).

7.6. Statistical Properties of LS Estimates

If the errors in the original measurements are uncorrelated, *i.e.*, $Cov(\epsilon_i, \epsilon_j) = 0 \quad \forall\, i \neq j$ and they all have the same variance, σ^2, then we write the data covariance matrix as $C_{\epsilon\epsilon} = \sigma^2 I$. I is an n by n identity matrix. When this property holds for the data errors, each $\hat{\theta}_k$ is an unbiased estimate of θ_k

$$\mathcal{E}(\hat{\theta}_k) = \theta_k \quad \text{for all } \ k = 1, \ldots, p \tag{43}$$

Also the variance-covariance matrix for $\hat{\theta}$ is $C_{\hat{\theta}\hat{\theta}} = \sigma^2 (X^T X)^{-1}$, so that

$$\begin{aligned} Cov(\hat{\theta}_k, \hat{\theta}_l) &= k,\ l\text{th element of } \sigma^2 (X^T X)^{-1} \\ Var(\hat{\theta}_k) &= k, k\text{th diagonal element of } \sigma^2 (X^T X)^{-1} \end{aligned} \tag{44}$$

$C_{\hat{\theta}\hat{\theta}}$ is a p by p matrix, and (44) is just a generalization of (13). Observe that even though the uncertainties in the original measurements are uncorrelated the parameter estimates derived from them are in general correlated. Under this model an unbiased estimate for σ^2 is

$$\hat{\sigma}^2 = s^2 = \frac{||\vec{Y} - \hat{Y}||^2}{n - p} = \frac{RSS}{n - p} \tag{45}$$

Also note that the normal equations provide estimators $\hat{\theta}$ that are linear combinations of the observations. The **Gauss-Markov** theorem states that the least squares estimators are the **best**

linear unbiased estimates (BLUE) for the parameters θ_i. "Best" in this context means most efficient or minimum variance. Suppose the LS estimate for θ_i is given by

$$\hat{\theta}_i = a_1 y_1 + a_2 y_2 + \ldots + a_n y_n \tag{46}$$

a linear combination of the observations. Then any other linear unbiased estimate for θ_i, for example,

$$\theta_i^* = b_1 y_1 + b_2 y_2 + \ldots + b_n y_n \tag{47}$$

will always have

$$Var(\theta_i^*) \geq Var(\hat{\theta}_i) \tag{48}$$

with equality if $a_i = b_i , i = 1, \ldots, n$

Unbiased estimators do not always have the smallest possible variance. It can be shown that minimum mean square error (MSE) estimators look like

$$\hat{\theta}_{MMSE} = (\alpha I + X^T X)^{-1} X^T \vec{Y} \tag{49}$$

where α is a suitably chosen scalar.

7.7. Inferences about $\vec{\theta}$

Suppose that we now further restrict the e_i to be independent and normally distributed. Since the $\hat{\theta}_k$ are linear combinations of individual normally distributed random variables they are also normal with mean θ_k and variance $\sigma^2 \Sigma_{kk}$ where we now have $\Sigma = (X^T X)^{-1}$. The standard error in $\hat{\theta}_k$ can be estimated as

$$s_{\hat{\theta}_k} = s\sqrt{\Sigma_{kk}} \tag{50}$$

Using this result we can once again construct confidence intervals and hypothesis tests. These will be exact under the case of normal data errors and approximate otherwise (courtesy a version of the Central Limit Theorem) since the $\hat{\theta}_k$ are a linear combination of the RV's ϵ_i.

With the normality assumption

$$\frac{\hat{\theta}_k - \theta_k}{s_{\hat{\theta}_k}} \sim t_{n-p}. \tag{51}$$

From this we get that a $100(1 - \alpha)\%$ confidence interval for θ_k is

$$\hat{\theta}_k \pm t_{n-p}(\frac{\alpha}{2}) s_{\hat{\theta}_k}. \tag{52}$$

Similarly if we have the null hypothesis $H_0 : \theta_j = \theta_{j_0}$ with θ_{j_0} some fixed number we can test with the statistic

$$t = \frac{\hat{\theta}_j - \theta_{j_0}}{s_{\hat{\theta}_j}} \tag{53}$$

This statistic will follow a t distribution with $n - p$ degrees of freedom under H_0. The most common test in this context is $H_0 : \theta_j = 0$ implying that x_j has no predictive value for the data y_i. Rice Fig 14.5.5 has some examples.

One criterion often used as a rule of thumb in evaluating models is the **squared multiple correlation coefficient** or coefficient of determination

$$R^2 = \frac{S_y^2 - S_{\hat{\epsilon}}^2}{S_y^2} \tag{54}$$

This provides a measure of the variance reduction achieved by the model: in other words it indicates how much of the variance in the original data is explained by the model.

If we have an independent estimate of the uncertainties in our measurements, that is we know σ^2 ahead of time then we would expect the variance of the standardized residuals r_i/σ_i to follow a χ^2 distribution with n degrees of freedom and the $E[\chi_n^2] = n$ (see Section 3.9).

7.8. Equivalence of ML and LS Estimates for Multivariate Normal

Suppose that we now assume that the ϵ_i are drawn from a Gaussian population. Then recalling the method of maximum likelihood described in Chapter 5 we can write the log-likelihood function for $\theta_1, \ \ldots, \theta_p$ as

$$\begin{aligned} l(\theta_1, \ldots, \theta_p) &= ln\big[\big(\frac{1}{2\pi\sigma^2}\big)^{n/2}\big] - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\theta_j X_{ij})^2 \\ &= -\frac{n}{2}ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|\vec{Y} - X\vec{\theta}\|^2 \\ &= -\frac{n}{2}ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|\vec{r}\|^2 \end{aligned} \tag{55}$$

Thus maximizing the likelihood function in this case will lead to identical results to minimizing the sum of the squares of the residuals.

Note that this is a special property of Gaussian errors, and recall the results of Section 5.5.5. If we had instead exponentially distributed uncertainties, i.e., $\epsilon_j \sim exp\{-|x_j|/x_0\}$, then l is maximized when $\sum_i |y_i - \sum_j \theta_j X_{ij}|$ is minimum. This is the one-norm of the vector of residuals, and θ is no longer a fixed linear combination of the y_i at its minimum.

7.9. Weighted Least Squares

What do we do when the ϵ_i are covariant, with general covariance matrix $C \neq \sigma^2 I$? Then the Gauss-Markov Theorem is no longer valid, and equivalence between LS and ML will not hold even under the assumption of normality. However, going back to our earlier description of covariance matrices (section 4.6), we remember that we can transform to a system where the observations are uncorrelated. We form new pseudo-data

$$\vec{Y}' = L\vec{Y} \tag{56}$$

so that $Var[\vec{Y}'] = \sigma^2 I$. We proceed as before with the new variables.

Note that if the covariance matrix for the ϵ_i is C, the likelihood function may be written as

$$l(\theta_1, \ldots, \theta_p) = -\frac{1}{2}log\big[(2\pi)^n det(C)\big] - \frac{1}{2}\big[(\vec{Y} - X\vec{\theta})^T C^{-1}(\vec{Y} - X\vec{\theta})\big] \tag{57}$$

Maximizing l with respect to $\vec{\theta}$ is equivalent to minimizing the quadratic form

$$\|\vec{r}'\|^2 = (\vec{Y} - X\vec{\theta})^T C^{-1} (\vec{Y} - X\theta) \tag{58}$$

A commonly occurring case is that of **weighted least squares**, when the observations are assumed uncorrelated, but with different variances. This leads to the minimization of

$$\|\vec{r}'\|^2 = \sum_{i=1}^{n} \frac{1}{\sigma_i^2} (y_i - \sum_j \theta_j X_{ij})^2 \tag{59}$$

or the **generalized normal equations**

$$X C^{-1} \vec{Y} = (X C^{-1} X^T) \hat{\theta} \tag{60}$$

7.10. Numerically Stable Solutions to LS Problems

Numerical solution of the normal equations needs to be performed with some care as the matrix $(XC^{-1}X^T)^{-1}$ may be very poorly conditioned. In fact it is sometimes not a good idea even to form the normal equations since the additions made in forming $(XC^{-1}X^T)$ can produce undesirable round-off error. X and $X^T X$ have the same null space and therefore the same rank. Thus the normal equations have a unique solution if and only if the columns of X are linearly independent.

The QR algorithm is a very numerically stable method for solution of a least squares problem * If X is an $n \times p$ matrix and is of rank p then we can write

$$X_{n \times p} = Q_{n \times p} R_{p \times p}$$

where $Q^T Q = I$, that is the columns of Q are orthogonal, and R is upper triangular, that is $r_{ij} = 0$ for all $i > j$. If we substitute $X = QR$ in the normal equations it is straightforward to show that the least squares estimate can be expressed as $\hat{\beta} = R^{-1} Q^T Y$ or equivalently

$$R\hat{\beta} = Q^T Y$$

so a very stable estimate for $\hat{\beta}$ can be obtained by back substitution.

Another alternative is to use the Cholesky decomposition

$$X^T X = R^T R,$$

with R upper triangular again. Then $R^T v = X^T Y$, again solved by back substitution for v and $\hat{\beta}$ is recovered from $R\beta = v$.

7.11. Non-Linear Least Squares (NLLS)

Earlier in this chapter we discussed how in some circumstances non-linear models can be exactly transformed into linear ones which can be solved directly. One example of this is given below:

* See C. L. Lawson and R. J. Hanson (1974) *Solving Least Squares Problems* (Prentice-Hall).

The Arrhenius relationship for thermally activated semiconduction in minerals is $\sigma(t) = \sigma_0 e^{-A/kt}$ where $\sigma(t)$ is the electrical conductivity at temperature t, k is Boltzmann's constant and A is the activation energy. This has been used to model the electrical conductivity data for the mineral olivine as a function of temperature. Olivine is a major constituent of Earth's mantle, and it is of some interest to understand the relationship between conductivity and the temperature and other physical properties of Earth's interior. For the conductivity example we can solve for the activation energy A and σ_0 simply by working in the log domain, and the transformed model is shown in Figure 7-1 for conductivity data derived from a sample of Jackson County dunite.

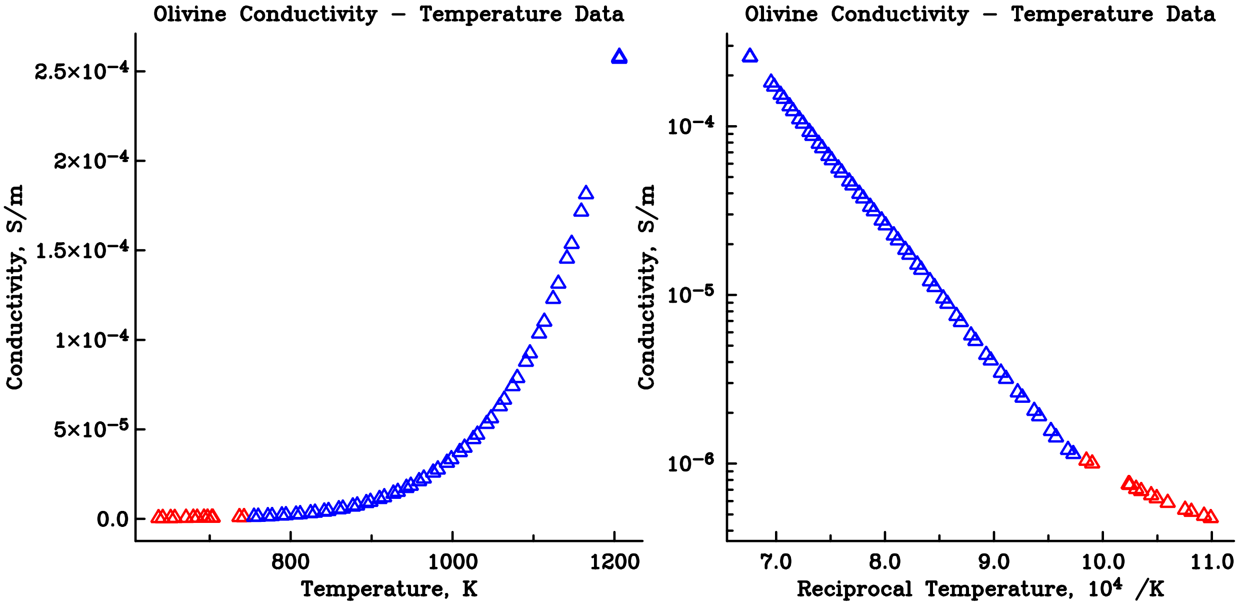

Figure 7 -1: Temperature - conductivity data for Jackson County dunite. The blue symbols appear to follow the Arrhenius relationship reasonably well, but the red parts will require additional non-linear terms.

However, in some circumstances such a transformation may be either not possible or undesirable - see the example in Figure 7-1. If there is no closed form solution then one must resort to solution by non-linear least squares, with a model that is non-linear in p unknown parameters with n observations. The basic strategy is to make some initial guess for a solution, approximate the model by a linear one and refine the parameter estimates by successive iterations.

The linear form of equation (28) is replaced by a nonlinear model function, so that

$$\vec{Y} = f(\vec{\theta}) + \vec{r} \tag{61}$$

with $\vec{\theta}$ again a vector of parameters, and the nonlinear prediction function f replacing the $n \times p$ matrix X. Again we want to miminize

$$S = \vec{r}.\vec{r}$$

which occurs when the gradient of S with respect to $\vec{\theta} = 0$, but in the nonlinear system the derivatives $\nabla_\theta \vec{r}$ will be functions of both the independent variables and the parameters. We choose an initial value for the parameters $\hat{\theta}_0$ and successively update the approximate estimates:

$$\hat{\theta}^{k+1} = \hat{\theta}^k + \Delta\hat{\theta}, \qquad k = 0, 1, 2, \ldots \tag{62}$$

Here the index k represents the iteration number, and at each iteration the vector of increments to the parameters is $\Delta\hat{\theta}$. At each iteration the model f is linearized using a 1st order Taylor expansion

$$\begin{aligned} f(X,\hat{\theta}) &\approx f^k(X,\hat{\theta}) + \nabla_\theta f(X,\hat{\theta}^k)(\hat{\theta}^k - \hat{\theta}) \\ &= f^k(X,\hat{\theta}) + J\Delta\hat{\theta} \end{aligned} \tag{63}$$

the matrix J is known as the Jacobian and is a function of constants, the independent variables, and the parameters $\vec{\theta}$ and varies from one iteration to the next. In terms of this linearized model, we can write

$$\Delta\vec{Y} = \vec{Y} - f^k(\hat{\theta}) = \vec{r}^k \tag{64}$$

and $\nabla_\theta \vec{r} = -J$.

Substituting the iterative update into the residual sum of squares S we recover the normal equations with the Jacobian in place of X:

$$(J^T J)\Delta\theta = J^T \Delta\vec{Y}, \tag{65}$$

forming the basis for the Gauss-Newton algorithm for solution of NLLS. As in the linear case these can be solved using QR, Cholesky or singular value decomposition for stability, but it's important to note that the algorithm will only converge when the objective function is approximately quadratic in the parameters. Various techniques have been devised for dealing with divergence during the iterative process. See Bob Parker's notes on Optimization from the SIO239 math class (or Numerical Recipes, or Practical Optimization, by Gill, Murray and Wright) for details of shift cutting, steepest descent methods, and the Levenberg-Marquadt algorithm.