Wiki: https://en.wikipedia.org/wiki/Least_squares

- When the observations come from an exponential family and mild conditions are satisfied, least-squares estimates and maximum-likelihood estimates are identical.[1] The method of least squares can also be derived as a method of moments estimator.

1 Least Square

$$S = \sum_{i=0}^{N-1} r_i^2 = \sum_{i=0}^{N-1} [y_i - f(x_i, \beta)]^2 \tag{1-1}$$

The minimum of the sum of squares is found by setting the gradient to zero. Since the model contains m parameters, there are m gradient equations:

$$\frac{\partial S}{\partial \beta_j} = -2 \sum_{i=0}^{N-1} \left([y_i - f(x_i, \beta)] \cdot \frac{\partial f(x_i, \beta)}{\partial \beta_j} \right) = 0 \tag{1-2}$$

1.1 Linear Least Square

A regression model is a linear one when the model comprises a linear combination of the parameters, i.e.

$$f(x, \beta) = \sum_{j=0}^{m-1} \beta_j \cdot \emptyset_j(x) \tag{1-3}$$

Letting

$$X_{i,j} = \frac{\partial f(x_i, \beta)}{\partial \beta_j} = \emptyset_j(x_i) \tag{1-4}$$

Then equation (1-2) becomes to

$$\begin{aligned}
&\sum_{i=0}^{N-1} \left([y_i - f(x_i, \beta)] \cdot X_{i,j} \right) = 0 \\
<=> &\sum_{i=0}^{N-1} y_i \cdot X_{i,j} = \sum_{i=0}^{N-1} f(x_i, \beta) \cdot X_{i,j} \\
<=> &\sum_{i=0}^{N-1} y_i \cdot X_{i,j} = \sum_{i=0}^{N-1} \left[\left(\sum_{j=0}^{m-1} \beta_j \cdot X_{i,j} \right) \cdot X_{i,j} \right] \\
<=> &\boldsymbol{X}_{:,\boldsymbol{j}}{}^T \cdot \vec{\boldsymbol{y}} = \sum_{i=0}^{N-1} [X_{i,j} \cdot (\boldsymbol{X}_{\boldsymbol{i},:} \cdot \widehat{\boldsymbol{\beta}})] \\
<=> &\boldsymbol{X}_{:,\boldsymbol{j}}{}^T \cdot \vec{\boldsymbol{y}} = \boldsymbol{X}_{:,\boldsymbol{j}}{}^T \cdot \boldsymbol{X}_{\boldsymbol{i},:} \cdot \widehat{\boldsymbol{\beta}} \\
<=> &\boldsymbol{X}^{\boldsymbol{T}} \cdot \vec{\boldsymbol{y}} = \boldsymbol{X}^{\boldsymbol{T}} \cdot \boldsymbol{X} \cdot \widehat{\boldsymbol{\beta}}
\end{aligned} \tag{1-5}$$

Then the parameters $\widehat{\boldsymbol{\beta}}$ can be estimated by

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\boldsymbol{T}} \cdot \boldsymbol{X})^{-\mathbf{1}} \cdot \boldsymbol{X}^{\boldsymbol{T}} \cdot \vec{\boldsymbol{y}} \tag{1-6}$$

1.2 Linear Least Square Estimation/Equalization

1.2.1 Signal model

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{s}) \tag{1-7}$$

Where $\boldsymbol{y}$ is the received signal, regression model to estiamte $\boldsymbol{s}$ using LLS

$$\hat{\boldsymbol{s}} = \boldsymbol{\beta}^{\boldsymbol{H}}\boldsymbol{y} \tag{1-8}$$

LLS to minimize

$$\boldsymbol{J} = (\boldsymbol{s} - \hat{\boldsymbol{s}})(\boldsymbol{s} - \hat{\boldsymbol{s}}) \tag{1-9}$$

Base on (1-6) and (1-9)

1.3 Non-linear Least Square

The formula (1-6) is also valid for non-linear case but with the $X_{i,j} = \frac{\partial f(x_i,\beta)}{\partial \beta_j}$.

There is no closed-form solution to a non-linear least squares problem. Instead, numerical algorithms are used to find the value of the parameters $\widehat{\boldsymbol{\beta}}$ that minimizes the objective. Most algorithms involve choosing initial values for the parameters. Then, the parameters are refined iteratively, that is, the values are obtained by successive approximation:

$$\beta_j^{k+1} = \beta_j^k + \Delta\beta_j \tag{1-10}$$

Or

$$\boldsymbol{\beta}^{\boldsymbol{k+1}} = \boldsymbol{\beta}^{\boldsymbol{k}} + \Delta\boldsymbol{\beta} \tag{1-11}$$

where k is an iteration number, and the vector of increments $\Delta\beta_j$ is called the shift vector. In some commonly used algorithms, at each iteration the model may be linearized by approximation to a first-order Taylor series expansion about $\boldsymbol{\beta}^{\boldsymbol{k}}$.

$$\begin{aligned} f(x_i,\beta) &= f^k(x_i,\beta) + \sum_j \frac{\partial f(x_i,\beta)}{\partial \beta_j}\left(\beta_j - \beta_j^k\right) \\ &= f^{k(x_i,\beta)} + \sum_j J_{ij}\Delta\beta_j \end{aligned} \tag{1-12}$$

The Jacobian J is a function of constans, the independent variable and the parameters, so it changes from one iteration to the next. The residuals are given by

$$\begin{aligned} r_i &= y_i - f^k(x_i,\beta) - \sum_{k=0}^{m-1} J_{ik}\Delta\beta_k \\ &= \Delta y_i - \sum_{j=0}^{m-1} J_{ij}\Delta\beta_j \end{aligned} \tag{1-13}$$

To minimize the sum of square of r_i, the gradient equation is set to zero and solved for $\Delta\beta_j$:

$$-2\sum_{i=0}^{N-1} J_{ij}\left(\Delta y_i - \sum_{k=0}^{m-1} J_{ik}\Delta\beta_k\right) = 0 \tag{1-14}$$

Which, on rearrangement, becom m simultaneous linear equations, the normal equations:

$$\sum_{i=0}^{N-1}\sum_{k=0}^{m-1} J_{ij}J_{ik}\Delta\beta_k = \sum_{i=0}^{N-1} J_{ij}\Delta y_i\,, j = 0, \dots, m-1 \qquad (1\text{-}15)$$

The normal equations are written in matrix notation as

$$(\boldsymbol{J}^T\boldsymbol{J})\Delta\boldsymbol{\beta} = \boldsymbol{J}^T\Delta\boldsymbol{y} \qquad (1\text{-}16)$$

These are the defining equations of the Gauss–Newton algorithm.

By rearranging (1-16)

$$\Delta\boldsymbol{\beta} = (\boldsymbol{J}^T\boldsymbol{J})^{-\mathbf{1}}\boldsymbol{J}^T\Delta\boldsymbol{y} \qquad (1\text{-}17)$$

One example of channel estimation moduel, it uses a simplified LS method:

$$\Delta\boldsymbol{\beta} \cong \boldsymbol{u}\boldsymbol{J}^T\Delta\boldsymbol{y} \qquad (1\text{-}18)$$

```
/* update channel estimate */

/* h[i] = h[i] + mu sum_{k=L}^{tsclen-1} tsc[k-i]*e[k+L], i=0:L */
for (i = 0; i <= L; i++)
{    temp.r = temp.i = 0;
     for (k = L; k < tsclen; k++)
     {    temp.r += tsc[k-i]*e[k].r;
          temp.i += tsc[k-i]*e[k].i;
     }
     h[i].r += temp.r*chanest_mu16/32768;
     h[i].i += temp.i*chanest_mu16/32768;
}
```

1.3.1 Iteration solving of the non-linear least square problem(not sure whether correct)

With formula (1-16), one can solve the non-linear least square peroblem by

$$\begin{aligned}\boldsymbol{\beta}^{k} &= \boldsymbol{\beta}^{k-\mathbf{1}} + \Delta\boldsymbol{\beta} \\ &= \boldsymbol{\beta}^{k-\mathbf{1}} + (\boldsymbol{J}^T\boldsymbol{J})^{-\mathbf{1}}\boldsymbol{J}^T\Delta\boldsymbol{y}\end{aligned} \qquad (1\text{-}19)$$

Where $\boldsymbol{J}$ can be calculated by

$$J_{ij} = \frac{\partial f(x_i, \beta)}{\partial \beta_j} \qquad (1\text{-}20)$$