ニューラルネットワークと回帰分析

021120,030526

1　回帰分析と最小２乗解

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

$\boldsymbol{y}= n \times 1$　　基準変数
$\boldsymbol{X}= n \times (q+1)$　　説明変数 $+\ \mathbf{1}$
$\boldsymbol{\beta}= (q+1) \times 1$　　回帰係数（切片を含む）

$$\begin{pmatrix} \widehat{y}_1 \\ \widehat{y}_{21} \\ \vdots \\ \widehat{y}_n \end{pmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1q} \\ 1 & x_{21} & x_{22} & \dots & x_{2q} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nq} \end{bmatrix} \times \begin{pmatrix} \mu \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_q \end{pmatrix} \tag{2}$$

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3}$$

$$\hat{y}_i = \sum_{j=0}^{q} x_{ij}\beta_j \tag{4}$$

$$\frac{\partial\,\mathrm{RSS}}{\partial\,\beta_k} = 2\sum_{i=1}^{n}(y_i - \hat{y}_i) \times \frac{\partial\,(y_i - \hat{y}_i)}{\partial\,\beta_k} \tag{5}$$

$$= 2\sum_{i=1}^{n}(y_i - \hat{y}_i) \times \frac{\partial \sum_{j=0}^{q} x_{ij}\beta_j}{\partial\,\beta_k} \tag{6}$$

$$= 2\sum_{i=1}^{n}(y_i - \hat{y}_i) \times x_{ik} \tag{7}$$

$$= 2\sum_{i=1}^{n}(y_i - \sum_{j=0}^{q} x_{ij}\beta_j) \times x_{ik} \tag{8}$$

$$= 2\sum_{i=1}^{n} y_i x_{ik} - 2\sum_{i=1}^{n}(\sum_{j=0}^{q} x_{ij}\beta_j)x_{ik} \tag{9}$$

$$= 2\sum_{i=1}^{n} y_i x_{ik} - 2\sum_{i=1}^{n}(\sum_{j=0}^{q} x_{ij}x_{ik}\beta_j) \tag{10}$$

$$= 2\sum_{i=1}^{n} y_i x_{ik} - 2\sum_{j=0}^{q}(\sum_{i=1}^{n} x_{ij}x_{ik})\beta_j \tag{11}$$

$\sum_{i=1}^{n} y_i x_{ik}$ は $\boldsymbol{X}'\boldsymbol{y}$ の第 k 要素、$\sum_{i=1}^{n} x_{ij}x_{ik}$ は $\boldsymbol{X}'\boldsymbol{X}$ の第 jk 要素
したがって、方程式 $\frac{\partial\,\mathrm{RSS}}{\partial\,\beta_k} = 0, k = 0, 1, 2, \dots, q$ は、まとめて

$$\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \mathbf{0} \tag{12}$$

と書ける。従って、解は

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{13}$$

となる。

2　バリエーション

1. 単回帰分析
   基準変数 1 個、説明変数 1 個

$$\boldsymbol{y} = \beta_0\mathbf{1} + \beta\boldsymbol{x} + \boldsymbol{\varepsilon} \tag{14}$$

2. 重回帰分析
   基準変数 1 個、説明変数 q 個

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{15}$$

3. 多変量回帰分析
   基準変数 p 個、説明変数 q 個

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}' + \boldsymbol{\mathcal{E}}, \quad \boldsymbol{\beta} = (p \times q) \tag{16}$$

$$\widehat{\boldsymbol{\beta}'} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \tag{17}$$

4. 冗長性分析

$$\boldsymbol{Y} = \mathbf{1}\boldsymbol{\mu}' + \boldsymbol{X}\boldsymbol{A}\boldsymbol{B}' + \boldsymbol{\mathcal{E}}, \quad \boldsymbol{A} = (q \times r), \quad \boldsymbol{B} = (p \times r) \quad r < p, q \tag{18}$$

5. 多項式回帰分析

$$\widehat{y_i} = \sum_{m=0}^{M} \beta_m x_i^m \tag{19}$$

$$\boldsymbol{Y} = \boldsymbol{G}^{(M)}(\boldsymbol{x})\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{G}^{M}(\boldsymbol{x}) = (n \times M + 1) \tag{20}$$

$$\begin{aligned} \boldsymbol{G}(\boldsymbol{x}) &= [\boldsymbol{g}_m(\boldsymbol{x})], \quad m = 0, 1, \ldots, M & (21)\\ \boldsymbol{g}_m(\boldsymbol{x}) &= [x_i^m] & (22) \end{aligned}$$

6. スプライン回帰分析（多項式 + 切断冪関数）

$$\widehat{y_i} = \sum_{m=0}^{M} \beta_m x_i^m + \sum_{\ell=1}^{K} \beta_\ell (x_i - c_\ell)_+^M \tag{23}$$

$$(x_i - c_\ell)_+^M = \begin{cases} 0 & (x_i < c_\ell) \\ (x - c_\ell)^M & (x_i \geq c_\ell) \end{cases} \tag{24}$$

$$\boldsymbol{Y} = \boldsymbol{G}^{(K,M)}(\boldsymbol{x})\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \boldsymbol{G}^{M}(\boldsymbol{x}) = (n \times M + K + 1) \tag{25}$$

ただし、 K は節点の数、c_ℓ は節点

$$y_i \quad = \quad \beta_0 + \sum_{j=1}^{Q} \psi(a_j(x_i - b_j))\beta_j + \varepsilon_i, \quad i = 1, 2, \cdots, n \tag{26}$$

ただし、a_j, b_j は定数であり、

$$\psi(x) = \frac{1}{(1 + \exp(-x))} \tag{27}$$

はロジスティク関数である。従って

$$\boldsymbol{y} = \boldsymbol{G}(\boldsymbol{x})\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{28}$$

の形で書くと

$$\boldsymbol{G}(\boldsymbol{x}) = [\mathbf{1},\ \psi(a_1(\boldsymbol{x} - b_1)),\ \psi(a_2(\boldsymbol{x} - b_2)),\ \ldots\ \psi(a_Q(\boldsymbol{x} - b_Q))] \tag{29}$$

となる。
なお、ニューラルネットワークモデルでは、$w_{ij}^{(k)}$ を第 $k-1$ 層の第 i 番目の変数から第 k 層の第 j 番目の変数を作るための重み、また、その際用いられる定数項を $\theta_j^{(k)}$ として、以下のように記すことが多い。

$$y_i \quad = \quad \theta^{(3)} + \sum_{j=1}^{Q} \psi(w_{j1}^{(2)} x_i + \theta_j^{(2)}) w_{j1}^{(3)} + \varepsilon_i, \quad i = 1, 2, \cdots, n \tag{30}$$

上記のモデルは、x を第 1 層の一つしかないニューロンの出力とし、第 2 層の第 j 番目のニューロンの入力が $w_{j1}^{(2)} x_i + \theta_j^{(2)}$、その出力が $\psi(w_{j1}^{(2)} x_i + \theta_j^{(2)})$ となっていると考えられる。したがって、これまでの係数が既知の場合は、ニューラルネットワークモデルは、n_2 個の第 2 層の出力を説明変数、y を基準変数とした重回帰分析モデルと考えることが出来る。
スプライン回帰や多項式と似ている点は、両者とも、本来の説明変数を G という形に展開し、それらの線形結合で $\boldsymbol{y}$ を予測しようと試みることである。
両者の違いは、スプラインの場合は、展開に際して用いられる接点 $c_\ell, \ell = 1, 2, \ldots, K$ は定数であるのに対し、ニューラルネットワークにおいては、その係数 $(a_j, b_j), j = 1, 2, \ldots, Q$ はパラメタとして推定されるものである点である。
また、多項式回帰の場合、中間層を増やすことは、結局、最初に展開された説明変数の線形回帰を作ることになり、意味がないが、スプラインや、特にニューラルネットワークの場合は、意味がある。

3　ニューラルネットワークについて

Q 個の説明変数 $\boldsymbol{x}$ を第 1 層、Q 個の基準変数 $\boldsymbol{y}$ を第 n 層とするネットワークを考える。各層のニューロンの数は $n_k, k = 1, 2, \ldots, n$ とする。従って、$n_1 = Q, n_n = Q$ である。また、第 k 層の第 j 番目のニューロンへの入力を i_j^k, 同じく出力を o_j^k と書く。各個体を表す添え字は $a, a = 1, 2, \ldots, N$ である。各ニューロンはそれぞれ上位のニューロンに向かって同一の出力を発する。また、各ニューロンは下位のニューロンの発する出力を重み付けした形でそのニューロンへの入力を形成する。すなわち、第 k-1 層の第 i 番目のニューロンの出力 o_i^{k-1} から第 k 層の第 j 番目のニューロンの入力 i_j^k を作る際の重みをw_{ij}^k とすれば、

$$i_{aj}^k = \sum_{i=1}^{n_{k-1}} o_{ai}^{k-1} w_{ij}^k + \theta_j^k \tag{31}$$

である。また、第 k 層の入力はシグモイド関数 ψ により、

$$o_{aj}^k = \psi(i_{aj}^k) = \frac{1}{1+\exp(-\lambda\, i_{aj}^k)}, \qquad (\lambda = 1.7 \quad \text{or} \quad 1) \tag{32}$$

という変換を受けた形で出力される。

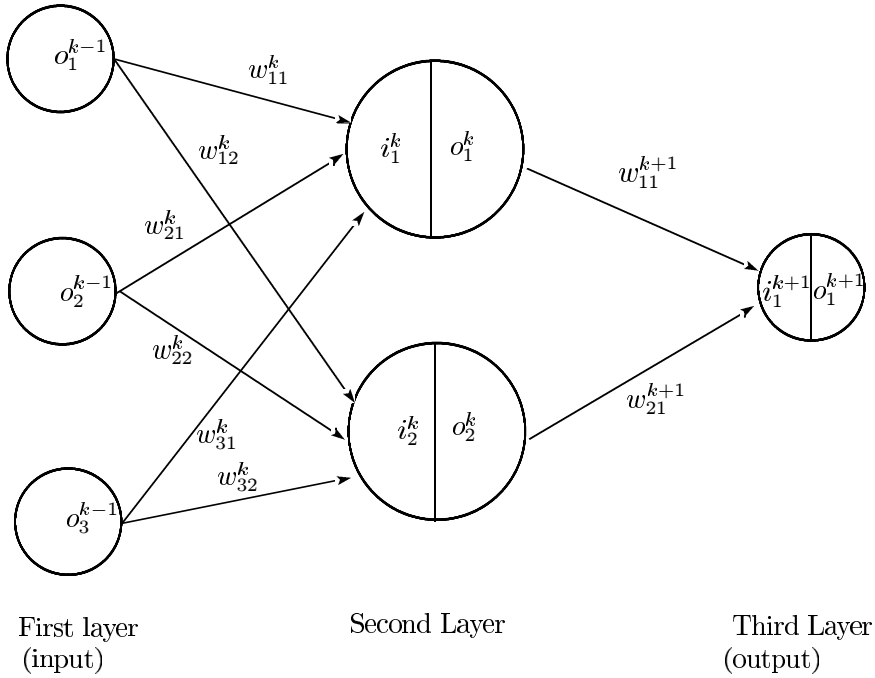

Figure 1: ニューラルネットワークモデル

これを全ての $i_{aj}^k, o_{aj}^k, a = 1, 2, \ldots, N$ に関してまとめて書くと、

$$\begin{aligned}
\boldsymbol{O}^k &= \boldsymbol{\Psi}(\boldsymbol{I}^k) && (33)\\
&= \boldsymbol{\Psi}(\boldsymbol{O}^{k-1}\boldsymbol{W}^k + \mathbf{1}\boldsymbol{\theta}^k) && (34)\\
&= \boldsymbol{\Psi}(\boldsymbol{\Psi}(\boldsymbol{I}^{k-1})\boldsymbol{W}^k + \mathbf{1}\boldsymbol{\theta}^k) && (35)\\
&= \boldsymbol{\Psi}(\boldsymbol{\Psi}(\boldsymbol{O}^{k-2}\boldsymbol{W}^{k-1} + \mathbf{1}\boldsymbol{\theta}^{k-1})\boldsymbol{W}^k + \mathbf{1}\boldsymbol{\theta}^k) && (36)\\
&\ \ \vdots \quad \vdots && (37)
\end{aligned}$$

となるが、($\boldsymbol{O}^k : N \times n_k$, $\boldsymbol{W}^k : n_{k-1} \times n_k$, $\boldsymbol{\theta}_j : 1 \times n_k$) これは

$$\boldsymbol{O}^k = \boldsymbol{\Psi}(\boldsymbol{O}^{k-1}\boldsymbol{W}^k + \mathbf{1}\boldsymbol{\theta}^k) = \boldsymbol{\Gamma}_k(\boldsymbol{O}^{k-1}) \tag{38}$$

と定義をすれば、最終層の出力が

$$\begin{aligned}
\boldsymbol{O}^n &= \boldsymbol{\Gamma}_n(\boldsymbol{O}^{n-1}) && (39)\\
&= \boldsymbol{\Gamma}_n(\boldsymbol{\Gamma}_{n-1}(\boldsymbol{O}^{n-2})) && (40)\\
&\ \ \vdots \quad \vdots && (41)\\
&= \boldsymbol{\Gamma}_n \circ \boldsymbol{\Gamma}_{n-1} \circ \cdots \circ \boldsymbol{\Gamma}_2(\boldsymbol{O}^1) && (42)\\
&= \boldsymbol{\Gamma}(\boldsymbol{O}^1) && (43)
\end{aligned}$$

と書けることを示している。ただし、

$$\boldsymbol{\Gamma}(.) = \boldsymbol{\Gamma}_n \circ \boldsymbol{\Gamma}_{n-1} \circ \cdots \circ \boldsymbol{\Gamma}_2(.) \tag{44}$$

である。従って、ニューラルネットワークのモデルは、$\boldsymbol{Y}=\boldsymbol{O}$ とおくことにより、

$$\boldsymbol{Y}=\boldsymbol{\Gamma}(\boldsymbol{X})+\boldsymbol{\mathcal{E}} \qquad \text{or} \qquad \widehat{\boldsymbol{Y}}=\boldsymbol{\Gamma}(\boldsymbol{X}) \tag{45}$$

と書くことができる。

4　最小２乗法によるパラメタの推定

データ $\boldsymbol{Y}$ とモデルの値 $\widehat{\boldsymbol{Y}}$ の差の２乗和

$$\mathrm{RSS}=\sum_{a=1}^{N}\sum_{j=1}^{Q}(y_{aj}-\hat{y}_{aj})^2 \tag{46}$$

を最小にするようにモデルのパラメタ

$$w_{ij}^k,\theta_j^k,k=2,3,\ldots,n,i=1,2,\ldots,n_{k-1},j=1,2,\ldots,n_k$$

を求めるためには以下に示す $\widehat{y_{aj}}$ を各パラメタで偏微分したものが必要となる。ただし、

$$\hat{y}_{aj}=o_{as}^n \tag{47}$$

であるが、個体を表す添字 a は落としてある。また、θ に関する微分も同時に示してある。

$$\frac{\partial\, o_s^n}{\partial\, w_{ij}^k} = \frac{\partial\, o_s^n}{\partial\, i_j^k}\times\frac{\partial\, i_j^k}{\partial\, w_{ij}^k} \tag{48}$$

$$\frac{\partial\, o_s^n}{\partial\, i_j^k} = \frac{\partial\, o_s^n}{\partial\, o_j^k}\times\frac{\partial\, o_j^k}{\partial\, i_j^k} \quad \text{and} \quad \frac{\partial\, o_s^n}{\partial\, o_j^k}=\sum_{q=1}^{n_{k+1}}\frac{\partial\, o_s^n}{\partial\, i_q^{k+1}}\times\frac{\partial\, i_q^{k+1}}{\partial\, o_j^k} \tag{49}$$

$$\frac{\partial\, i_q^{k+1}}{\partial\, o_j^k} = w_{jq}^{k+1} \quad \text{and} \quad \frac{\partial\, o_j^k}{\partial\, i_j^k}=o_j^k(1-o_j^k) \tag{50}$$

従って、$\delta_j^{k,s}=\dfrac{\partial\, o_s^n}{\partial\, i_j^k}$ と書くと、

$$\delta_j^{k,s} = \frac{\partial\, o_s^n}{\partial\, i_j^k} \tag{51}$$

$$= \sum_{q=1}^{n_{k+1}}\frac{\partial\, o_s^n}{\partial\, i_q^{k+1}}\,w_{jq}^{k+1}\times\frac{\partial\, o_j^k}{\partial\, i_j^k}=\sum_{q=1}^{n_{k+1}}\delta_q^{k+1,s}\,w_{jq}^{k+1}\times\frac{\partial\, o_j^k}{\partial\, i_j^k} \tag{52}$$

と再帰的にかける。（これが誤差逆伝搬の由来である。）ただし

$$\frac{\partial\, i_j^k}{\partial\, w_{ij}^k} = \begin{cases} o_i^{k-1} & \text{if not}\quad \theta \\ 1 & \text{if}\quad \theta \end{cases} \tag{53}$$

である。また、$k=n$ の場合には

$$\delta_j^{n,s} = \frac{\partial\, o_s^n}{\partial\, i_j^n}=\begin{cases} o_s^n(1-o_s^n) & \text{if}\quad j=s \\ 0 & \text{otherwise} \end{cases} \tag{54}$$

となる。

最適化の具体的方法に関しては次節以降に述べるが、ニューラルネットワークの重みの推定には、全ての $\boldsymbol{W}^K$ もしくはその部分を $R\times 1$ のベクトルの形にしたものを $\boldsymbol{\xi}$ とし、ヤコビアン行列 $\boldsymbol{F}_{(j)}(\boldsymbol{\xi})$ の要素としてこの $\dfrac{\partial\, o_j^n}{\partial\, w_{IJ}^K}$ を用いることになる。また、個体を表す添え字 a は、次節では i となっている。

5　非線形モデルの最小２乗解：　ガウス・ニュートン法

モデル $\widehat{\boldsymbol{y}}(\boldsymbol{\xi})$ がパラメタ $\boldsymbol{\xi}$ $(R\times\ 1)$ に関して非線形の場合，最小２乗基準

$$\mathrm{RSS}(\boldsymbol{\xi}) = \|\boldsymbol{y} - \widehat{\boldsymbol{y}}(\boldsymbol{\xi})\|^2 = \sum_{i=1}^{N}(y_i - \hat{y}_i(\boldsymbol{\xi}))^2 \tag{55}$$

を最小にするようなパラメタの推定値を解析的に求めることは困難である場合が多い．しかし，そのような場合にも，何らかの数値的最適化の方法を用いれば，最小２乗解を求めることが可能である．

一般に，ある関数 $\mathrm{r}(\boldsymbol{\xi})$ の極値（極小値または極大値のことであるが，本節では極小値とする）を与えるパラメタ $\boldsymbol{\xi}$ の値を数値的に求めることを関数の最適化と呼ぶが，その方法は，何らかの方法で求めた $\boldsymbol{\xi}$ の初期値から，$\mathrm{r}(\boldsymbol{\xi})$ が減少する方向 $\boldsymbol{d}$ へある分量 α だけ修正する，

$$\boldsymbol{\xi}_{(\ell+1)} = \boldsymbol{\xi}_{(\boldsymbol{\ell})} + \alpha\boldsymbol{d}_{(\boldsymbol{\ell})} \tag{56}$$

という形の繰り返し計算でパラメタの値を更新していくという形をとる．数値的最適化の方法には様々なものがあるが，それらの主な違いは，更新の方向 $\boldsymbol{d}_{(\boldsymbol{\ell})}$ をどの様にして定めるかによる．

最も単純なものは最急降下法と呼ばれる方法であり，修正の方向を $\mathrm{r}(\boldsymbol{\xi})$ のパラメタの各要素に関する一次偏微分係数行列の逆方向

$$\boldsymbol{d}_{(\boldsymbol{\ell})} = -\boldsymbol{g}(\boldsymbol{\xi}_{(\boldsymbol{\ell})}) = -\left.\frac{\partial\,\mathrm{r}(\boldsymbol{\xi})}{\partial\,\boldsymbol{\xi}}\right|_{\boldsymbol{\xi}=\boldsymbol{\xi}_{(\boldsymbol{\ell})}} \tag{57}$$

に取るものである．この，$R\ \times\ 1$ のベクトル $\boldsymbol{g}$ はグレイディエントと呼ばれることがあるが，これは $\boldsymbol{\xi}_{(\boldsymbol{\ell})}$ における $\mathrm{r}(\boldsymbol{\xi})$ のパラメタの各要素に関する傾きの逆方向を示すもので，Eq.(56) を用いることにより，

$$\mathrm{r}(\boldsymbol{\xi}_{(\ell+1)}) \leq\ \mathrm{r}(\boldsymbol{\xi}_{(\boldsymbol{\ell})}) \tag{58}$$

を満たす α が存在することが知られている．

$\mathrm{r}(\boldsymbol{\xi})$ を $\boldsymbol{\xi}_{(\boldsymbol{\ell})}$ のまわりで２次の項まで Taylor 展開すると

$$\mathrm{r}(\boldsymbol{\xi}) \approx \mathrm{r}(\boldsymbol{\xi}_{(\boldsymbol{\ell})}) + \boldsymbol{g}(\boldsymbol{\xi}_{(\boldsymbol{\ell})})'(\boldsymbol{\xi} - \boldsymbol{\xi}_{(\boldsymbol{\ell})}) + \tfrac{1}{2}(\boldsymbol{\xi} - \boldsymbol{\xi}_{(\boldsymbol{\ell})})'\boldsymbol{H}(\boldsymbol{\xi}_{(\boldsymbol{\ell})})(\boldsymbol{\xi} - \boldsymbol{\xi}_{(\boldsymbol{\ell})}) \tag{59}$$

となるが，この $R\ \times\ R$ の２次微分行列（ r をパラメタの各要素で２回偏微分したものを要素として持つ行列）

$$\boldsymbol{H}(\boldsymbol{\xi}_{(\boldsymbol{\ell})}) = \left[\left.\frac{\partial^2 r(\boldsymbol{\xi})}{\partial\boldsymbol{\xi}\ \partial\boldsymbol{\xi}'}\right|_{\boldsymbol{\xi}=\boldsymbol{\xi}_{(\boldsymbol{\ell})}}\right] \tag{60}$$

はヘシアン行列と呼ばれるものである．これを考慮すると、Eq.(56) において

$$\boldsymbol{d}_{(\boldsymbol{\ell})} = -\boldsymbol{H}(\boldsymbol{\xi}_{(\boldsymbol{\ell})})^{-1}\boldsymbol{g}(\boldsymbol{\xi}_{(\boldsymbol{\ell})}) \tag{61}$$

とおいたものが Newton-Raphson 法の更新式として得られる．

最適化されるべき関数が最小２乗基準の場合（ $\mathrm{r}(\boldsymbol{\xi}) = \mathrm{RSS}(\boldsymbol{\xi})$ ）求めるべきパラメタの仮の値を $\boldsymbol{\xi}_{(\boldsymbol{\ell})}$ とする．モデルの各要素を $\boldsymbol{\xi}_{(\boldsymbol{\ell})}$ のまわりで 1 次の項まで Taylor 展開したものをまとめれば，

$$\widehat{\boldsymbol{y}}(\boldsymbol{\xi}) \approx \widehat{\boldsymbol{y}}(\boldsymbol{\xi}_{(\boldsymbol{\ell})}) + \boldsymbol{F}(\boldsymbol{\xi}_{(\boldsymbol{\ell})})(\boldsymbol{\xi} - \boldsymbol{\xi}_{(\boldsymbol{\ell})}) \tag{62}$$

となるが，ここに，

$$\boldsymbol{F}(\boldsymbol{\xi}) = \frac{\partial\,\widehat{\boldsymbol{y}}(\boldsymbol{\xi})}{\partial\,\boldsymbol{\xi}} = \left[\frac{\partial\,\hat{y}_i(\boldsymbol{\xi})}{\partial\,\boldsymbol{\xi}_k}\right],\quad i = 1,2,\ \ldots,N,\ k = 1,2,\ \ldots,R \tag{63}$$

は，ヤコビアン行列と呼ばれる $N \times R$ の 1 次微分行列，また，$\boldsymbol{F}(\boldsymbol{\xi}_{(\ell)})$ はそれを $\boldsymbol{\xi}_{(\ell)}$ で評価した行列

$$\boldsymbol{F}(\boldsymbol{\xi}_{(\ell)}) = \left.\frac{\partial\, \widehat{\boldsymbol{y}}(\boldsymbol{\xi})}{\partial\, \boldsymbol{\xi}}\right|_{\boldsymbol{\xi}=\boldsymbol{\xi}_{(\ell)}} \tag{64}$$

である．このことを考慮して、Eq.(56) において

$$\boldsymbol{d}_{(\ell)} = -\tfrac{1}{2}\big(\boldsymbol{F}(\boldsymbol{\xi}_{(\ell)})'\boldsymbol{F}(\boldsymbol{\xi}_{(\ell)})\big)^{-1}\boldsymbol{g}(\boldsymbol{\xi}_{(\ell)}) \tag{65}$$

と置いたものを用いる方法が Gauss-Newton 法と呼ばれるものである。なお、このヤコビアン行列を用いると

$$\boldsymbol{g} = \frac{\partial\, \mathrm{RSS}(\boldsymbol{\xi})}{\partial\, \boldsymbol{\xi}} = -2\boldsymbol{F}(\boldsymbol{\xi})'(\boldsymbol{y} - \widehat{\boldsymbol{y}}(\boldsymbol{\xi})) \tag{66}$$

となる。Gauss-Newton 法は 1 次の情報しか利用しないが効率はよく，Eq.(58) を満たす α も必ず存在することが知られている．

また，基準変数が多変量の場合は，

$$\mathrm{RSS} \;=\; \|\boldsymbol{y} - \widehat{\boldsymbol{y}}(\boldsymbol{\xi})\|^2 \;=\; \sum_{j=1}^{p}\left\|\boldsymbol{y}_{(j)} - \widehat{\boldsymbol{y}}_{(j)}(\boldsymbol{\xi})\right\|^2 \;=\; \sum_{j=1}^{p}\sum_{i=1}^{N}(y_{ij} - \hat{y}_{ij}(\boldsymbol{\xi}))^2 \tag{67}$$

と書けるため，変数ごとのヤコビアン行列を

$$\boldsymbol{F}_{(j)}(\boldsymbol{\xi}) = \frac{\partial\, \widehat{\boldsymbol{y}}_{(j)}(\boldsymbol{\xi})}{\partial\, \boldsymbol{\xi}} = \left[\frac{\partial\, \hat{y}_{ij}(\xi)}{\partial\, \boldsymbol{\xi}_k}\right], \quad i = 1,2,\ldots,N,\; k = 1,2,\ldots,R \tag{68}$$

と定義すれば，グレイディエント並びに更新の方向は

$$\boldsymbol{g} \;=\; \frac{\partial\, \mathrm{RSS}(\boldsymbol{\xi})}{\partial\, \boldsymbol{\xi}} \;=\; -2\sum_{j=1}^{p}\boldsymbol{F}_{(j)}(\boldsymbol{\xi})'(\boldsymbol{y}_{(j)} - \widehat{\boldsymbol{y}}_{(j)}(\boldsymbol{\xi})) \tag{69}$$

$$\boldsymbol{d}_{(\ell)} \;=\; -\tfrac{1}{2}\Big(\textstyle\sum_{j=1}^{p}\boldsymbol{F}_{(j)}(\boldsymbol{\xi}_{(\ell)})'\boldsymbol{F}_{(j)}(\boldsymbol{\xi}_{(\ell)})\Big)^{-1}\boldsymbol{g}(\boldsymbol{\xi}_{(\ell)}) \tag{70}$$

で与えられる．

パラメタを更新すべき方向が定まれば，ステップサイズ α の値を適当に決めることにより，Eq.(58) を満たすような $\boldsymbol{\xi}_{(\ell)}$ を求めることが出来る．ステップサイズを決める最も簡単な方法は，まず $\alpha = 1$ から始めて Eq.(58) が満たされない場合には，それが満たされるまで α の値を半減していくという方法である．より高度な方法としては，RSS の変化を参考にしてステップサイズの減少幅を決める方法や，$\boldsymbol{d}_{(\ell)}$ を既知として RSS を α の関数とみなし，それを３次関数で近似して極値を求める方法がある．

Gauss-Newton 法の一般的な形は以下のようになる．

Gauss-Newton 法

$\boldsymbol{\xi}_1$ に初期値を設定する.
for ℓ = 1 to maxiter until(RSS($\boldsymbol{\xi}_{(\ell)}$) - RSS($\boldsymbol{\xi}_{(\ell+1)}$) が小さい);
  $\boldsymbol{F}_{(j)}(\boldsymbol{\xi}_{(\ell)})$, $\boldsymbol{g}(\boldsymbol{\xi}_{(\ell)})$, $\boldsymbol{d}_{(\ell)}$ を計算する.
  α=1;
  for iters = 1 to maxhalf;
    $\boldsymbol{\xi}_{(\ell+1)} = \boldsymbol{\xi}_{(\ell)} + \alpha\, \boldsymbol{d}_{(\ell)}$;
    if RSS($\boldsymbol{\xi}_{(\ell+1)}$) $\leq$ RSS($\boldsymbol{\xi}_{(\ell)}$) then break;
    α=α/2;
  end;
end;