Standard Regression

1. Model Setup

Assume a sample of n observations. Regression formula is given as follows:

$$\boxed{\vec{Y} = X\vec{\beta} + \vec{\varepsilon}}$$

- $\vec{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$ represents the vector of target variables for each observation i, $i = 1,2,\dots,n$, where $y_i \in \mathbb{R}$
- $X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,m} \end{bmatrix}$ represents the feature matrix of the sample, where each column represents a feature (typically, the first column has its elements equal to ones to represent the intercept term of the regression model), with at total of m features (including a possible intercept term).
- $\vec{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix}$ represents coefficient values
- $\vec{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$ represents the vector of error variables for each observation i, $i = 1,2,\dots,n$, where $\varepsilon_i \in \mathbb{R}$ and $E[\varepsilon_i] = 0$

Model aims to estimate $\vec{Y}$, given by the vector $\hat{\vec{Y}} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix}$ by estimating $\vec{\beta}$, given by the vector $\hat{\vec{\beta}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_m \end{bmatrix}$ and using the following formula $\hat{\vec{Y}} = X\hat{\vec{\beta}}$. We aim to have $\vec{Y}$ and $\hat{\vec{Y}}$ as close as possible to each other.

2. Parameter Estimation (No Regularization)

2.1 Optimization Function

We choose the following loss function to minimize:

$$L\left(\hat{\vec{\beta}}\right) = \left(\vec{Y} - \hat{\vec{Y}}\right)^T \times \left(\vec{Y} - \hat{\vec{Y}}\right)$$

2.2 Closed-form Solution

Assumptions:

- X is full rank (a necessary assumption when inverting $X^T X$)
- No column of X is all zeros (guarantees a global minimum as we will show)

We use the traditional derivative approach to find the values of $\vec{\beta}$ (which will become our $\hat{\vec{\beta}}$) that optimize the loss function. If the second derivative test yields a positive value, then we guarantee a local minimum. If the second derivative is a constant, then we guarantee a global minimum. If the second derivative test is zero or negative, then we will need to use another approach.

We begin by taking the first derivative and second derivatives with respect to $\hat{\beta}_j$ for $j = 1,2,\dots,m$ and following through with the algebra:

$$\begin{aligned}
\frac{\partial L\left(\hat{\vec{\beta}}\right)}{\partial \hat{\beta}_j} &= \frac{\partial\left(\left(\vec{Y} - \hat{\vec{Y}}\right)^T \times \left(\vec{Y} - \hat{\vec{Y}}\right)\right)}{\partial \hat{\beta}_j} \\
&= \frac{\partial\left(\left(\vec{Y} - X\hat{\vec{\beta}}\right)^T \times \left(\vec{Y} - X\hat{\vec{\beta}}\right)\right)}{\partial \hat{\beta}_j} \\
&= \frac{\partial\left(\sum_{i=1}^{n}\left(y_i - \sum_{j'=1}^{m} x_{i,j'}\hat{\beta}_{j'}\right)^2\right)}{\partial \hat{\beta}_j} \\
&= \frac{\partial\left(\sum_{i=1}^{n}\left(y_i - \sum_{j'=1}^{m} x_{i,j'}\hat{\beta}_{j'}\right)^2\right)}{\partial \hat{\beta}_j} \\
&= \sum_{i=1}^{n} -2\left(y_i - \sum_{j'=1}^{m} x_{i,j'}\hat{\beta}_{j'}\right)x_{i,j} \\
&= -2\sum_{i=1}^{n}\left(y_i - \sum_{j'=1}^{m} x_{i,j'}\hat{\beta}_{j'}\right)x_{i,j}
\end{aligned}$$

$$\Longrightarrow$$

$$\frac{\partial L^2\left(\hat{\vec{\beta}}\right)}{\partial \hat{\beta}_j^2} = \frac{\partial}{\partial \hat{\beta}_j}\left(\frac{\partial L\left(\hat{\vec{\beta}}\right)}{\partial \hat{\beta}_j}\right)$$

$$= \frac{\partial}{\partial \hat{\beta}_j}\left(-2\sum_{i=1}^{n}\left(y_i x_{i,j} - \sum_{j'=1}^{m} x_{i,j'}\hat{\beta}_{j'}x_{i,j}\right)\right)$$

$$= \frac{\partial}{\partial \hat{\beta}_j}\left(-2\sum_{i=1}^{n} y_i x_{i,j} + 2\sum_{i=1}^{n}\sum_{j'=1}^{m} x_{i,j'}\hat{\beta}_{j'}x_{i,j}\right)$$

$$= \frac{\partial}{\partial \hat{\beta}_j}\left(-2\sum_{i=1}^{n} y_i x_{i,j}\right) + \frac{\partial}{\partial \hat{\beta}_j}\left(2\sum_{i=1}^{n}\sum_{j'=1}^{m} x_{i,j'}\hat{\beta}_{j'}x_{i,j}\right)$$

$$= \left(-2\frac{\partial}{\partial \hat{\beta}_j}\sum_{i=1}^{n} y_i x_{i,j}\right) + \left(2\frac{\partial}{\partial \hat{\beta}_j}\sum_{i=1}^{n}\sum_{j'=1}^{m} x_{i,j'}\hat{\beta}_{j'}x_{i,j}\right)$$

$$= \left(-2\sum_{i=1}^{n}\frac{\partial}{\partial \hat{\beta}_j}\left(y_i x_{i,j}\right)\right) + \left(2\sum_{i=1}^{n}\sum_{j'=1}^{m}\frac{\partial}{\partial \hat{\beta}_j}\left(x_{i,j'}\hat{\beta}_{j'}x_{i,j}\right)\right)$$

$$= \left(-2\sum_{i=1}^{n} 0\right) + \left(2\sum_{i=1}^{n} x_{i,j}\sum_{j'=1}^{m}\frac{\partial}{\partial \hat{\beta}_j}\left(x_{i,j'}\hat{\beta}_{j'}\right)\right)$$

$$= 0 + \left(2\sum_{i=1}^{n} x_{i,j}\left(x_{i,j}\right)\right)$$

$$= 2\sum_{i=1}^{n} x_{i,j}^2$$

For every j, if there exists an $x_{i,j}$ that is non-zero, then $2\sum_{i=1}^{n} x_{i,j}^2 > 0$ implying that we have a global minimum. Since we assume this above, we therefore guarantee a global minimum.

Therefore, the loss function can be globally minimized using the first derivative method. We now apply the first derivative in matrix form. Let $\vec{x}_{i,.}$ be a row matrix that represents the ith row of the feature matrix X (note: when differentiating by vector, we use the so-called numerator layout notation; source: https://en.wikipedia.org/wiki/Matrix_calculus#Layout_conventions):

$$\frac{\partial L\left(\hat{\vec{\beta}}\right)}{\partial \hat{\vec{\beta}}} = \frac{\partial\left(\left(\vec{Y} - \hat{\vec{Y}}\right)^T \times \left(\vec{Y} - \hat{\vec{Y}}\right)\right)}{\partial \hat{\vec{\beta}}}$$

$$= \frac{\partial\left(\left(\vec{Y} - X\hat{\vec{\beta}}\right)^T \times \left(\vec{Y} - X\hat{\vec{\beta}}\right)\right)}{\partial \hat{\vec{\beta}}}$$

$$= \frac{\partial\left(\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{m} x_{i,j}\hat{\beta}_j\right)^2\right)}{\partial \hat{\vec{\beta}}}$$

$$= \left[\frac{\partial \left(\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} x_{i,j} \hat{\beta}_j \right)^2 \right)}{\partial \hat{\beta}_1} \quad \frac{\partial \left(\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} x_{i,j} \hat{\beta}_j \right)^2 \right)}{\partial \hat{\beta}_2} \quad \cdots \quad \frac{\partial \left(\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} x_{i,j} \hat{\beta}_j \right)^2 \right)}{\partial \hat{\beta}_m} \right]$$

$$= \left[\sum_{i=1}^{n} -2 \left(y_i - \sum_{j=1}^{m} x_{i,j} \hat{\beta}_j \right) x_{i,1} \quad \sum_{i=1}^{n} -2 \left(y_i - \sum_{j=1}^{m} x_{i,j} \hat{\beta}_j \right) x_{i,2} \quad \cdots \quad \sum_{i=1}^{n} -2 \left(y_i - \sum_{j=1}^{m} x_{i,j} \hat{\beta}_j \right) x_{i,m} \right]$$

$$= \left[-2 \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} x_{i,j} \hat{\beta}_j \right) x_{i,1} \quad -2 \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} x_{i,j} \hat{\beta}_j \right) x_{i,2} \quad \cdots \quad -2 \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} x_{i,j} \hat{\beta}_j \right) x_{i,m} \right]$$

$$= -2 \times \left[\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} x_{i,j} \hat{\beta}_j \right) x_{i,1} \quad \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} x_{i,j} \hat{\beta}_j \right) x_{i,2} \quad \cdots \quad \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} x_{i,j} \hat{\beta}_j \right) x_{i,m} \right]$$

$$= -2 \times \left[\sum_{i=1}^{n} \left(y_i - \vec{x}_{i,.} \hat{\vec{\beta}} \right) x_{i,1} \quad \sum_{i=1}^{n} \left(y_i - \vec{x}_{i,.} \hat{\vec{\beta}} \right) x_{i,2} \quad \cdots \quad \sum_{i=1}^{n} \left(y_i - \vec{x}_{i,.} \hat{\vec{\beta}} \right) x_{i,m} \right]$$

$$= -2 \times \begin{bmatrix} y_1 - \vec{x}_{1,.} \hat{\vec{\beta}} & y_1 - \vec{x}_{1,.} \hat{\vec{\beta}} & \cdots & y_1 - \vec{x}_{1,.} \hat{\vec{\beta}} \\ y_2 - \vec{x}_{2,.} \hat{\vec{\beta}} & y_2 - \vec{x}_{2,.} \hat{\vec{\beta}} & \cdots & y_2 - \vec{x}_{2,.} \hat{\vec{\beta}} \\ \vdots & \vdots & \ddots & \vdots \\ y_n - \vec{x}_{n,.} \hat{\vec{\beta}} & y_n - \vec{x}_{n,.} \hat{\vec{\beta}} & \cdots & y_n - \vec{x}_{n,.} \hat{\vec{\beta}} \end{bmatrix}^T \times \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,m} \end{bmatrix}$$

$$= -2 \times \left(\vec{Y} - X \hat{\vec{\beta}} \right)^T \times \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,m} \end{bmatrix}$$

$$= -2 \times \left(\vec{Y} - X \hat{\vec{\beta}} \right)^T \times X$$

$$= -2 \left(\vec{Y} - X \hat{\vec{\beta}} \right)^T X$$

Now we set the $\frac{\partial L\left(\hat{\vec{\beta}}\right)}{\partial \hat{\vec{\beta}}} = [0 \quad 0 \quad \cdots \quad 0]$ and solve:

$$-2 \left(\vec{Y} - X \hat{\vec{\beta}} \right)^T X = [0 \quad 0 \quad \cdots \quad 0]$$

$$\Leftrightarrow 2 \left(\vec{Y} - X \hat{\vec{\beta}} \right)^T X = [0 \quad 0 \quad \cdots \quad 0]$$

$$\Leftrightarrow 2 \left(\vec{Y}^T - \hat{\vec{\beta}}^T X^T \right) X = [0 \quad 0 \quad \cdots \quad 0]$$

$$\Leftrightarrow 2 \left(\vec{Y}^T X - \hat{\vec{\beta}}^T X^T X \right) = [0 \quad 0 \quad \cdots \quad 0]$$

$$\Leftrightarrow \vec{Y}^T X - \hat{\vec{\beta}}^T X^T X = [0 \quad 0 \quad \cdots \quad 0]$$

$$\Leftrightarrow \hat{\vec{\beta}}^T X^T X - Y^T X = [0 \quad 0 \quad \cdots \quad 0]$$

$$\Leftrightarrow \hat{\vec{\beta}}^T X^T X = [0 \quad 0 \quad \cdots \quad 0] + \vec{Y}^T X$$

$$\Leftrightarrow \hat{\vec{\beta}}^T X^T X = \vec{Y}^T X$$
$$\Leftrightarrow \hat{\vec{\beta}}^T X^T X \times (X^T X)^{-1} = \vec{Y}^T X \times (X^T X)^{-1}$$
$$\Leftrightarrow \hat{\vec{\beta}}^T = \vec{Y}^T X \times (X^T X)^{-1}$$
$$\Leftrightarrow \hat{\vec{\beta}}^T = \vec{Y}^T X (X^T X)^{-1}$$
$$\Leftrightarrow \hat{\vec{\beta}}^{T^T} = \left(\vec{Y}^T X (X^T X)^{-1}\right)^T$$
$$\Leftrightarrow \boxed{\hat{\vec{\beta}} = (X^T X)^{-1} X^T \vec{Y}}$$

2.3 Gradient Descent Solution

Let $\hat{\vec{\beta}}^{\{n\}}$ represent the estimates of $\vec{\beta}$ at the nth iteration. Given initial values at $n = 0$, a learning rate η, we have the following expression (note: the transpose expression below is included as we are conforming to the numerator layout notation when differentiating vectors):

$$\hat{\vec{\beta}}^{\{n+1\}} = \hat{\vec{\beta}}^{\{n\}} - \eta \times \left(\frac{\partial L\left(\hat{\vec{\beta}}^{\{n\}}\right)}{\partial \hat{\vec{\beta}}^{\{n\}}} \right)^T$$

Using the results in 2.2, we have $\frac{\partial L\left(\hat{\vec{\beta}}^{\{n\}}\right)}{\partial \hat{\vec{\beta}}^{\{n\}}} = -2\left(\vec{Y} - X\hat{\vec{\beta}}^{\{n\}}\right)^T X$

Thus, we have the following result:

$$\hat{\vec{\beta}}^{\{n+1\}} = \hat{\vec{\beta}}^{\{n\}} + \eta \times 2 \times X^T \left(\vec{Y} - X\hat{\vec{\beta}}^{\{n\}}\right)$$

We can let the learning rate, η, absorb the constant 2. Thus, we have the following results:

$$\boxed{\hat{\vec{\beta}}^{\{n+1\}} = \hat{\vec{\beta}}^{\{n\}} + \eta \times X^T \times \vec{Y} - X\hat{\vec{\beta}}^{\{n\}}}$$

We can keep iterating until convergence is met or a maximum number of iterations is hit.

3. Parameter Estimation (Elastic Net Regularization)

3.1 Optimization Function

Let $\left\|\hat{\vec{\beta}}\right\|_1 = \sum_{j=1}^m |\hat{\beta}_j|$ and $\left\|\hat{\vec{\beta}}\right\|_2 = \sum_{j=1}^m \hat{\beta}_j^2$. Given the two non-negative hyperparameters λ_1 and λ_2, we choose the following loss function to minimize:

$$L\left(\hat{\vec{\beta}}\right) = \left(\vec{Y} - \hat{\vec{Y}}\right)^T \times \left(\vec{Y} - \hat{\vec{Y}}\right) + \lambda_1 \left\|\hat{\vec{\beta}}\right\|_1 + \lambda_2 \left\|\hat{\vec{\beta}}\right\|_2$$

3.2 Coordinate Descent

Assumptions:

- X is full rank (a necessary assumption when inverting $X^T X$)
- No column of X is all zeros (guarantees a global minimum as we will show)

We use the traditional derivative approach to find the values of $\vec{\beta}$ (which will become our $\hat{\vec{\beta}}$) that optimize the loss function. If the second derivative test yields a positive value, then we guarantee a local minimum. If the second derivative is a constant, then we guarantee a global minimum. If the second derivative test is zero or negative, then we will need to use another approach.

We begin by taking the first derivative and second derivatives with respect to $\hat{\beta}_j$ for $j = 1,2,\dots,m$ and following through with the algebra (note: we leverage the results from 2.2):

$$\frac{\partial L\left(\hat{\vec{\beta}}\right)}{\partial \hat{\beta}_j} = \frac{\partial\left(\left(\vec{Y} - \hat{\vec{Y}}\right)^T \times \left(\vec{Y} - \hat{\vec{Y}}\right) + \lambda_1 \left\|\hat{\vec{\beta}}\right\|_1 + \lambda_2 \left\|\hat{\vec{\beta}}\right\|_2\right)}{\partial \hat{\beta}_j}$$

$$= \frac{\partial\left(\left(\vec{Y} - \hat{\vec{Y}}\right)^T \times \left(\vec{Y} - \hat{\vec{Y}}\right) + \lambda_1 \sum_{j'=1}^{m} |\hat{\beta}_{j'}| + \lambda_2 \sum_{j=1}^{m} \hat{\beta}_{j'}^2\right)}{\partial \hat{\beta}_j}$$

$$= \frac{\partial\left(\left(\vec{Y} - \hat{\vec{Y}}\right)^T \times \left(\vec{Y} - \hat{\vec{Y}}\right)\right)}{\partial \hat{\beta}_j} + \frac{\partial\left(\lambda_1 \sum_{j'=1}^{m} |\hat{\beta}_{j'}|\right)}{\partial \hat{\beta}_j} + \frac{\partial\left(\lambda_2 \sum_{j=1}^{m} \hat{\beta}_{j'}^2\right)}{\partial \hat{\beta}_j}$$

$$= -2\sum_{i=1}^{n}\left(y_i - \sum_{j'=1}^{m} x_{i,j'}\hat{\beta}_{j'}\right)x_{i,j} + \frac{\partial\left(\lambda_1 |\hat{\beta}_j|\right)}{\partial \hat{\beta}_j} + \frac{\partial\left(\lambda_2 \hat{\beta}_j^2\right)}{\partial \hat{\beta}_j}$$

$$= -2\sum_{i=1}^{n}\left(y_i - \sum_{j'=1}^{m} x_{i,j'}\hat{\beta}_{j'}\right)x_{i,j} + \lambda_1 \frac{\partial |\hat{\beta}_j|}{\partial \hat{\beta}_j} + \lambda_2 \frac{\partial \hat{\beta}_j^2}{\partial \hat{\beta}_j}$$

$$= -2\sum_{i=1}^{n}\left(y_i - \sum_{j'=1}^{m} x_{i,j'}\hat{\beta}_{j'}\right)x_{i,j} + \lambda_1 \begin{cases} -1 \; if \hat{\beta}_j < 0 \\ [-1,1] \; if \; \hat{\beta}_j = 0 + 2\lambda_2\hat{\beta}_j \\ 1 \; if \; \hat{\beta}_j > 0 \end{cases}$$

$$= -2\sum_{i=1}^{n}\left(y_i - \sum_{j'=1}^{m} x_{i,j'}\hat{\beta}_{j'}\right)x_{i,j} + 2\lambda_1 \begin{cases} -\frac{1}{2} \; if \hat{\beta}_j < 0 \\ \left[-\frac{1}{2}, \frac{1}{2}\right] \; if \; \hat{\beta}_j = 0 + 2\lambda_2\hat{\beta}_j \\ \frac{1}{2} \; if \; \hat{\beta}_j > 0 \end{cases}$$

$$\Rightarrow$$

$$\frac{\partial L^2\left(\hat{\vec{\beta}}\right)}{\partial \hat{\beta}_j^2} = \frac{\partial}{\partial \hat{\beta}_j}\left(\frac{\partial L\left(\hat{\vec{\beta}}\right)}{\partial \hat{\beta}_j}\right)$$

$$= \frac{\partial}{\partial \hat{\beta}_j}\left(-2\sum_{i=1}^{n}\left(y_i - \sum_{j'=1}^{m} x_{i,j'}\hat{\beta}_{j'} \right) x_{i,j} + 2\lambda_1 \begin{cases} -\frac{1}{2}\ if \hat{\beta}_j < 0 \\ \left[-\frac{1}{2}, \frac{1}{2}\right]\ if\ \hat{\beta}_j = 0 + 2\lambda_2\hat{\beta}_j \\ \frac{1}{2}\ if\ \hat{\beta}_j > 0 \end{cases} \right)$$

$$\frac{\partial}{\partial \hat{\beta}_j}\left(-2\sum_{i=1}^{n}\left(y_i - \sum_{j'=1}^{m} x_{i,j'}\hat{\beta}_{j'} \right) x_{i,j} \right) + \frac{\partial}{\partial \hat{\beta}_j}\left(2\lambda_1 \begin{cases} -\frac{1}{2}\ if \hat{\beta}_j < 0 \\ \left[-\frac{1}{2}, \frac{1}{2}\right]\ if\ \hat{\beta}_j = 0 \\ \frac{1}{2}\ if\ \hat{\beta}_j > 0 \end{cases} \right) + \frac{\partial}{\partial \hat{\beta}_j}\left(2\lambda_2\hat{\beta}_j\right)$$

$$= 2\sum_{i=1}^{n} x_{i,j}^2 + 0 + 2\lambda_2$$

Thus, assuming that $\sum_{i=1}^{n} x_{i,j}^2 > 0$, we guarantee a global minimum.

We now proceed to deriving an optimization expression for $\hat{\beta}_j$ (unlike 2.2, we do not use matrices due to the existence of cases because of the $\lambda_1 \left\| \hat{\vec{\beta}} \right\|_1$ component:

$$\frac{\partial L\left(\hat{\vec{\beta}}\right)}{\partial \hat{\beta}_j} = -2\sum_{i=1}^{n}\left(y_i - \sum_{j'=1}^{m} x_{i,j'}\hat{\beta}_{j'} \right) x_{i,j} + 2\lambda_1 \begin{cases} -\frac{1}{2}\ if \hat{\beta}_j < 0 \\ \left[-\frac{1}{2}, \frac{1}{2}\right]\ if\ \hat{\beta}_j = 0 + 2\lambda_2\hat{\beta}_j \\ \frac{1}{2}\ if\ \hat{\beta}_j > 0 \end{cases}$$

$$= \begin{cases} -2\sum_{i=1}^{n}\left(y_i - \sum_{j'=1}^{m} x_{i,j'}\hat{\beta}_{j'} \right) x_{i,j} - \lambda_1 + 2\lambda_2\hat{\beta}_j & if \hat{\beta}_j < 0 \\ \left[-2\sum_{i=1}^{n}\left(y_i - \sum_{j'=1}^{m} x_{i,j'}\hat{\beta}_{j'} \right) x_{i,j} - \lambda_1 + 2\lambda_2\hat{\beta}_j, 2\sum_{i=1}^{n}\left(y_i - \sum_{j'=1}^{m} x_{i,j'}\hat{\beta}_{j'} \right) x_{i,j} + \lambda_1 + 2\lambda_2\hat{\beta}_j \right] & if\ \hat{\beta}_j = 0 \\ -2\sum_{i=1}^{n}\left(y_i - \sum_{j'=1}^{m} x_{i,j'}\hat{\beta}_{j'} \right) x_{i,j} + \lambda_1 + 2\lambda_2\hat{\beta}_j & if \hat{\beta}_j > 0 \end{cases}$$

$$= \begin{cases} -2\sum_{i=1}^{n}\left(y_i - \sum_{j'=1}^{m} x_{i,j'}\hat{\beta}_{j'}\right)x_{i,j} - \lambda_1 + 2\lambda_2\hat{\beta}_j & if \hat{\beta}_j < 0 \\ \left[-2\sum_{i=1}^{n}\left(y_i - \sum_{j'=1}^{m} x_{i,j'}\hat{\beta}_{j'}\right)x_{i,j} - \lambda_1, -2\sum_{i=1}^{n}\left(y_i - \sum_{j'=1}^{m} x_{i,j'}\hat{\beta}_{j'}\right)x_{i,j} + \lambda_1\right] & if\ \hat{\beta}_j = 0 \\ -2\sum_{i=1}^{n}\left(y_i - \sum_{j'=1}^{m} x_{i,j'}\hat{\beta}_{j'}\right)x_{i,j} + \lambda_1 + 2\lambda_2\hat{\beta}_j & if \hat{\beta}_j > 0 \end{cases}$$

We break up $\sum_{i=1}^{n}\left(y_i - \sum_{j'=1}^{m} x_{i,j'}\hat{\beta}_{j'}\right)x_{i,j}$ as follows:

$$\begin{aligned} \sum_{i=1}^{n}\left(y_i - \sum_{j'=1}^{m} x_{i,j'}\hat{\beta}_{j'}\right)x_{i,j} &= \sum_{i=1}^{n}\left(y_i - \sum_{j'=1,j'\neq j}^{m} x_{i,j'}\hat{\beta}_{j'} - x_{i,j}\hat{\beta}_j\right)x_{i,j} \\ &= \sum_{i=1}^{n}\left(y_i - \sum_{j'=1,j'\neq j}^{m} x_{i,j'}\hat{\beta}_{j'}\right)x_{i,j} - x_{i,j}^2\hat{\beta}_j \\ &= \sum_{i=1}^{n}\left(y_i - \sum_{j'=1,j'\neq j}^{m} x_{i,j'}\hat{\beta}_{j'}\right)x_{i,j} - \sum_{i=1}^{n} x_{i,j}^2\hat{\beta}_j \\ &= \sum_{i=1}^{n}\left(y_i - \sum_{j'=1,j'\neq j}^{m} x_{i,j'}\hat{\beta}_{j'}\right)x_{i,j} - \hat{\beta}_j\sum_{i=1}^{n} x_{i,j}^2 \end{aligned}$$

Let $\rho_j = \sum_{i=1}^{n}\left(y_i - \sum_{j'=1,j'\neq j}^{m} x_{i,j'}\hat{\beta}_{j'}\right)x_{i,j}$ and $z_j = \sum_{i=1}^{n} x_{i,j}^2$

Thus,

$$\frac{\partial L\left(\hat{\beta}\right)}{\partial\hat{\beta}_j} = \begin{cases} -2\left(\rho_j - \hat{\beta}_j z_j\right) - \lambda_1 + 2\lambda_2\hat{\beta}_j & if \hat{\beta}_j < 0 \\ \left[-2\left(\rho_j - \hat{\beta}_j z_j\right) - \lambda_1, -2\left(\rho_j + \hat{\beta}_j z_j\right) + \lambda_1\right] & if\ \hat{\beta}_j = 0 \\ -2\left(\rho_j - \hat{\beta}_j z_j\right) + \lambda_1 + 2\lambda_2\hat{\beta}_j & if \hat{\beta}_j > 0 \end{cases}$$

$$= \begin{cases} -2\rho_j + 2\hat{\beta}_j z_j - \lambda_1 + 2\lambda_2\hat{\beta}_j & if \hat{\beta}_j < 0 \\ \left[-2\rho_j - \lambda_1, -2\rho_j + \lambda_1\right] & if\ \hat{\beta}_j = 0 \\ -2\rho_j + 2\hat{\beta}_j z_j + \lambda_1 + 2\lambda_2\hat{\beta}_j & if \hat{\beta}_j > 0 \end{cases}$$

We set the expression for each case and set the derivative of each one separately:
$if \hat{\beta}_j < 0$:

$$-2\rho_j + 2\hat{\beta}_j z_j - \lambda_1 + 2\lambda_2\hat{\beta}_j = 0$$
$$\Leftrightarrow 2\hat{\beta}_j\left(z_j + \lambda_2\right) = 2\rho_j + \lambda_1$$

$$\Leftrightarrow \hat{\beta}_j(z_j + \lambda_2) = \rho_j + \frac{\lambda_1}{2}$$

$$\Leftrightarrow \hat{\beta}_j = \frac{\rho_j + \frac{\lambda_1}{2}}{z_j + \lambda_2}$$

Now $z_j = \sum_{i=1}^{n} x_{i,j}^2 > 0$, $\lambda_1 > 0$, and $\lambda_2 > 0$. Thus, $\hat{\beta}_j < 0 \Leftrightarrow \rho_j + \frac{\lambda_1}{2} < 0 \Leftrightarrow \rho_j < -\frac{\lambda_1}{2}$

$if \hat{\beta}_j = 0$:

$$\left[-2\rho_j - \lambda_1, -2\rho_j + \lambda_1\right] = 0$$

For this to occur, we need
$-2\rho_j - \lambda_1 = -2\rho_j + \lambda_1 \Leftrightarrow 0 = 2\lambda_1 \Leftrightarrow \lambda_1 = 0$, which is a trivial solution.
Thus, we will use a sub-optimal result: Letting zero lie within $\left[-2\rho_j - \lambda_1, -2\rho_j + \lambda_1\right]$

Hence,

$$\left[-2\rho_j - \lambda_1 \geq 0\right] \cup \left[-2\rho_j + \lambda_1 \leq 0\ \right] \Leftrightarrow \left[\rho_j \geq -\frac{\lambda_1}{2}\right] \cup \left[\rho_j \leq \frac{\lambda_1}{2}\right]$$
$$\Leftrightarrow \rho_j \in \left[-\frac{\lambda_1}{2}, \frac{\lambda_1}{2}\right]$$

$if \hat{\beta}_j > 0$:

$$-2\rho_j + 2\hat{\beta}_j z_j + \lambda_1 + 2\lambda_2 \hat{\beta}_j = 0$$
$$\Leftrightarrow 2\hat{\beta}_j(z_j + \lambda_2) = 2\rho_j - \lambda_1$$
$$\Leftrightarrow \hat{\beta}_j(z_j + \lambda_2) = \rho_j - \frac{\lambda_1}{2}$$

$$\Leftrightarrow \hat{\beta}_j = \frac{\rho_j - \frac{\lambda_1}{2}}{z_j + \lambda_2}$$

Now $z_j = \sum_{i=1}^{n} x_{i,j}^2 > 0$, $\lambda_1 > 0$, and $\lambda_2 > 0$. Thus, $\hat{\beta}_j < 0 \Leftrightarrow \rho_j - \frac{\lambda_1}{2} > 0 \Leftrightarrow \rho_j > \frac{\lambda_1}{2}$

Combining all cases, we have the following result:

$$\hat{\beta}_j = \begin{cases} \dfrac{\rho_j + \frac{\lambda_1}{2}}{z_j + \lambda_2} & if \rho_j < -\dfrac{\lambda_1}{2} \\ 0 & if\ \rho_j \in \left[-\dfrac{\lambda_1}{2}, \dfrac{\lambda_1}{2}\right] \\ \dfrac{\rho_j - \frac{\lambda_1}{2}}{z_j + \lambda_2} & if \rho_j > \dfrac{\lambda_1}{2} \end{cases}$$

Let $\hat{\beta}_j^{\{n\}}$ represent the estimate of β_j for $j = 1,2,\dots,m$, at the nth iteration. Given initial values at $n = 0$, we have the following expression:

$$\hat{\beta}_j^{\{n+1\}} = \begin{cases} \dfrac{\rho_j^{\{n\}} + \frac{\lambda_1}{2}}{z_j + \lambda_2} & if \rho_j^{\{n\}} < -\dfrac{\lambda_1}{2} \\ 0 & if\ \rho_j^{\{n\}} \in \left[-\dfrac{\lambda_1}{2}, \dfrac{\lambda_1}{2}\right] \\ \dfrac{\rho_j^{\{n\}} - \frac{\lambda_1}{2}}{z_j + \lambda_2} & if \rho_j^{\{n\}} > \dfrac{\lambda_1}{2} \end{cases}$$

We can keep iterating until convergence is met or a maximum number of iterations is hit.

3.3 Gradient Descent

Let $\hat{\beta}_j^{\{n\}}$ represent the estimate of β_j for $j = 1,2,\dots,m$, at the nth iteration. Given initial values at $n = 0$, a learning rate η, we have the following expression for each j:

$$\hat{\beta}_J^{\{n+1\}} = \hat{\beta}_J^{\{n\}} - \eta \times \frac{\partial L\left(\hat{\hat{\beta}}^{\{n\}}\right)}{\partial \hat{\beta}_J^{\{n\}}}$$

Using the results and notation given in 3.2, we have

$$\frac{\partial L\left(\hat{\hat{\beta}}^{\{n\}}\right)}{\partial \hat{\beta}_j^{\{n\}}} = \begin{cases} -2\rho_j^{\{n\}} + 2\hat{\beta}_j^{\{n\}} z_j - \lambda_1 + 2\lambda_2 \hat{\beta}_j^{\{n\}} & if \hat{\beta}_j^{\{n\}} < 0 \\ \left[-2\rho_j^{\{n\}} - \lambda_1, -2\rho_j^{\{n\}} + \lambda_1\right] & if\ \hat{\beta}_j^{\{n\}} = 0 \\ -2\rho_j^{\{n\}} + 2\hat{\beta}_j^{\{n\}} z_j + \lambda_1 + 2\lambda_2 \hat{\beta}_j^{\{n\}} & if \hat{\beta}_j^{\{n\}} > 0 \end{cases}$$

For the middle expression, if $\hat{\beta}_j^{\{n\}} = 0$, we randomly pick a value between $\left[-2\rho_j^{\{n\}} - \lambda_1, -2\rho_j^{\{n\}} + \lambda_1\right]$ to be the value of $\frac{\partial L\left(\hat{\hat{\beta}}^{\{n\}}\right)}{\partial \hat{\beta}^{\{n\}}}$.

We can keep iterating until convergence is met or a maximum number of iterations is hit.