Example 1: A Least Squares Approach

Linear Regression

$$\hat{\omega}_0 = \bar{t} - \bar{\omega}_1 \bar{x} \qquad \hat{\omega}_1 = \frac{\overline{xt} - \bar{x}\bar{t}}{\overline{x^2} - (\bar{x})^2}$$

N=3 data points

n	x_n	t_n	$x_n t_n$	x_n^2
1	1	4.8	4.8	1
2	3	11.3	33.9	9
3	5	17.2	86	25
$\frac{1}{N}\sum_{n=1}^{N}$	3	11.1	41.57	11.67

$x_n : \frac{1}{3}(1+3+5) = \frac{9}{3} = 3$

$t_n : \frac{1}{3}(4.8+11.3+17.2) = \frac{33.3}{3} = 11.1$

$n=1 : x_1 t_1 = 1 \cdot 4.8 = 4.8$

$n=2 : x_2 t_2 = 3 \cdot 11.3 = 33.9$

$n=3 : x_3 t_3 = 5 \cdot 17.2 = 86$

$\frac{1}{3}\sum_{n=1}^{3} \frac{4.8+33.9+86}{3} = 41.57$

$\sum_{n=1}^{3} \frac{(x_1)^2+(x_2)^2+(x_3)^2}{3}$

$\frac{1^2+3^2+5^2}{3} = 11.67$

$\hat{\omega}_1 = \frac{41.57-(3)(11.1)}{11.67-(3)^2} = \frac{8.27}{2.67} = 3.1$

$\hat{\omega}_0 = 11.1 - 3.1(3) = 1.8$

$f(x; w_0, w_1) = w_0 + w_1 x$

$f(x; w_0, w_1) = 1.8 + 3.1x$

Example 2: Matrix Approach to Multiple Regression

We are given the following data

- Q1 - Write the feature matrix & the equations for coefficients
  - Find the coefficients β
- Q2 If the initial $x_{i1} = 8$ and $x_{i2} = 15$, target y_i ?

Sample Number	Target Y_i	Feature 1 X_{i1}	Feature 2 X_{i2}
1	3.0	0	1
2	5.0	2	3
3	9.0	4	8
4	10.0	6	10

$$\beta = (X^T X)^{-1} X^T y$$

Step 1: Define the feature Matrix (X) & Target Vector (y)

* The first column of all 1's in the feature matrix → bias term

$y = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$;
- β_0 is intercept (or bias)
- β_1, β_2 are coefficients for features x_{i1} & x_{i2}

$$X = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 2 & 3 \\ 1 & 4 & 8 \\ 1 & 6 & 10 \end{bmatrix}, \quad y = \begin{bmatrix} 3.0 \\ 5.0 \\ 9.0 \\ 10.0 \end{bmatrix}$$

Step 2: Compute $(X^T X)^{-1} \Rightarrow X^T X$

$$\underset{3\times4}{\begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 2 & 4 & 6 \\ 1 & 3 & 8 & 10 \end{bmatrix}} \cdot \underset{4\times3}{\begin{bmatrix} 1 & 0 & 1 \\ 1 & 2 & 3 \\ 1 & 4 & 8 \\ 1 & 6 & 10 \end{bmatrix}} = \underset{3\times3}{\begin{bmatrix} 4 & 12 & 22 \\ 12 & 56 & 98 \\ 22 & 98 & 174 \end{bmatrix}}$$

$$\det A = \begin{bmatrix} 4 & 12 & 22 \\ 12 & 56 & 98 \\ 22 & 98 & 174 \end{bmatrix}\begin{matrix} 4 & 12 \\ 12 & 56 \\ 22 & 98 \end{matrix} = (174 \cdot 56 \cdot 4) + 25872 + 25872 - (27104 + 38416 + 25056)$$

$\det A = 144$

$$\begin{bmatrix} 4 & 12 & 22 \\ 12 & 56 & 98 \\ 22 & 98 & 174 \end{bmatrix}\begin{matrix} 4 & 12 \\ 12 & 56 \\ 22 & 98 \end{matrix}$$
$$\begin{matrix} 4 & 12 & 22 & 4 & 12 \\ 12 & 56 & 98 & 12 & 56 \end{matrix}$$

$$\underset{3\times3}{\Rightarrow} \begin{bmatrix} 140 & 68 & -56 \\ 68 & 212 & -128 \\ -56 & -128 & 80 \end{bmatrix}^T \Rightarrow A^{-1} = \frac{1}{\det A} \cdot \mathrm{Adj}(A) = \frac{1}{144}\begin{bmatrix} 140 & 68 & -56 \\ 68 & 212 & -128 \\ -56 & -128 & 80 \end{bmatrix}$$

Step 3: Compute $X^T y$

$$X^T y = \underset{3\times3}{\begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 2 & 4 & 6 \\ 1 & 3 & 8 & 10 \end{bmatrix}} \cdot \underset{4\times1}{\begin{bmatrix} 3.0 \\ 5.0 \\ 9.0 \\ 10.0 \end{bmatrix}} = \underset{3\times1}{\begin{bmatrix} 27 \\ 106 \\ 190 \end{bmatrix}}$$

Step 4: Compute $\beta = (X^T X)^{-1} X^T y$

$$\beta = \frac{1}{144}\begin{bmatrix} 140 & 68 & -56 \\ 68 & 212 & -128 \\ -56 & -128 & 80 \end{bmatrix} \cdot \begin{bmatrix} 27 \\ 106 \\ 190 \end{bmatrix} = \frac{1}{144}\begin{bmatrix} 348 \\ -12 \\ 120 \end{bmatrix}$$

$$\beta = \begin{bmatrix} 2.42 \\ -0.08 \\ 0.83 \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{bmatrix} \Rightarrow \hat{y} = \beta_0 + \beta_1 x + \dots \beta_K x_K$$

$$\hat{y} = 2.42 - 0.08x_1 + 0.83x_2$$

$x_1 = 8$ $\quad \hat{y} = 2.42 - 0.08(8) + 0.83(15)$

$x_2 = 15$ $\quad \hat{y} = 14.23$

Example 3:

- For each pair, state if the true function is in the model class or not
  - That is, is there under-modeling or not?
  - If true function is in the model class, state the true Parameter

Ex 1
- True Function: $f_0(x) = 2+3x$ Model class: $f(x,\beta) = \beta_0 + \beta_1 x + \beta_2 x^2$
- No under-modeling. True parameter: $\beta = (2,3,0)$

Ex 2
- True Function: $f_0(x) = 2+3x+4x^2$ Model class $f(x,\beta) = \beta_0 + \beta_1 x$
- There is under-modeling. Model class doesn't contain x^2 term

Ex 3
- True Func: $f_0(x) = \sin(2\pi(5x)x+7)$ Model class $f(x,\beta) = \beta_0 \sin(2\pi(5)x) + \beta_1 \cos(2\pi(5)x)$
- No under-modeling $f_0(x) = \sin(2\pi(5x)x+7) = \sin(2\pi(5)x)\cos(7) + \cos(2\pi(5)x)\sin(7)$
- True Parameter $\beta = (\cos 7, \sin 7)$

Ex 4
- True Func $f_0(x) = \sin(2\pi(8)x+7)$ Model class: $f(x,\beta) = \beta_0 \sin(2\pi(5)x) + \beta_1 \cos(2\pi(5)x)$
- There is under-modeling. Model class does not contain $\sin(2\pi 8)$ or $\cos(2\pi 8)$

Finding the determinant of a 4x4 matrix

$$A = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 2 & 1 & 2 \\ 2 & 1 & 0 & 1 \\ 2 & 0 & 1 & 4 \end{bmatrix}$$

Expand row 1

$$\det(A) = 1\begin{bmatrix} 2 & 1 & 2 \\ 1 & 0 & 1 \\ 0 & 1 & 4 \end{bmatrix} + 0 + 0 + 1\begin{bmatrix} 0 & 2 & 1 \\ 2 & 1 & 0 \\ 2 & 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 2 & 1 & 2 \\ 1 & 0 & 1 \\ 0 & 1 & 4 \end{bmatrix}\begin{matrix} 2 & 1 \\ 1 & 0 \\ 0 & 1 \end{matrix} \qquad \begin{bmatrix} 0 & 2 & 1 \\ 2 & 1 & 0 \\ 2 & 0 & 1 \end{bmatrix}\begin{matrix} 0 & 2 \\ 2 & 1 \\ 2 & 0 \end{matrix}$$

$0+0+2-(0+2+4) = -4$ $\qquad$ $0+0+0-(2+0+4) = -6$

$1(-4) = -4$ $\qquad$ $1(-6) = -6$

$\det(A) = -4 + (-6) = -10$

Example 4:

You are given dataset with 100 features & an output variable y. You want to fit a Linear model of the form

$$y = \beta_0 + \sum_{i=1}^{100} \beta_i x_i + \epsilon, \quad \epsilon \text{ represents random noise}$$

1) All coefficients β_i should be close to zero, but not necessarily exactly zero

Soln: Ridge Regression (L2 Regularization)
- Regularization term: $\lambda \sum_{i=1}^{100} \beta_i^2$
- Ridge regression shrinks the values of all coefficients toward zero but not 0
- This approach helps in preventing overfitting

2) Most of coefficients β_i should be exactly zero

Soln: Lasso Regression (L1) Reg. term: $\lambda \sum_{i=1}^{100} |\beta_i|$
- Lasso Reg. applies L1 penalty → forces coefficients to be 0
- This tech. performs automatic feature selection

3) Among the first 50 coefficients, most should be zero, but the remaining coefficients should be unconstrained

Soln: Group lasso or Elastic Net
- Option 1: Group lasso
  - we apply L1 penalty first 50, keep last 50 free
- Option 2: Elastic Net (comb. L1 & L2)

Reg term: $\lambda_1 \sum_{i=1}^{50} |\beta_i| + \lambda_2 \sum_{i=51}^{100} \beta_i^2$

Elastic Net:
- Lasso effect on the first 50 coefficients (sparsity)
- Ridge effect on 50 remaining coefficients (prevent large swings)

Constraint	Regularization type	Regularization term
Coefficients should be small but not exactly zero	Ridge Reg (L2)	$\lambda \sum \beta_i^2$
Most coefficients should be zero	Lasso Reg (L1)	$\lambda \sum$
Among 1'st 50 most should be zero, rest unconstrained	Group Lasso/Elastic Net	$\lambda_1 \sum_{i=1}^{50}$