Simple/Multiple Linear Regression

© Malay K. Das, 210 Southern Lab, ph-7359, mkdas@iitk.ac.in

Office hours: W 1030-1130, SL-210

Previously:

Simple Linear Regression

Today:

1. Linear regression: simple to multiple
2. Error estimation
3. Example
4. Vector-matrix notation

HW due:

August 16, 2024

1

Malay K. Das, mkdas@iitk.ac.in

Simple Linear Regression

(x_i, y_i)

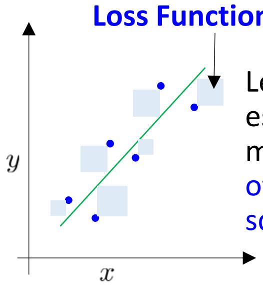
For training data $\mathcal{T} = \{(x_i, y_i)\}_{i=1}^{n}$ estimate $\widehat{y}_{\star}(x_{\star})$

in simple linear regression $x_i \in \mathbb{R}, y_i \in \mathbb{R}$

residual (not the noise from experiments)

we assume a hypothesis $\widehat{y} = w_0 + w_1 x$ $\quad y = \widehat{y} + e \quad L_i = e_i^2 \quad E(w_0, w_1) = \frac{1}{n}\sum_{i=1}^{n} L_i$

Loss Function $L_i = e_i^2$ Least square estimate minimizes **cost function**



Least square estimate minimizes sum of areas of the squares

$$\boxed{w_0 = \bar{y} - \bar{x} w_1 \qquad w_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - (\bar{x})^2}}$$

Alternate forms $w_1 = \frac{S_{xy}}{S_{xx}}$ $\quad S_{xy} = \sum_{i=1}^{n} [(x_i - \bar{x})\, y_i]$

$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$ or $S_{xy} = \sum_{i=1}^{n} [(x_i - \bar{x})(y_i - \bar{y})]$

Also $\sum_{i=1}^{n} e_i = 0$ (prove)

2

Tools for evaluating linear regression quality

$$y_i - \overline{y} = (\widehat{y}_i - \overline{y}) + (y_i - \widehat{y}_i) \Rightarrow (y_i - \overline{y})^2 = (\widehat{y}_i - \overline{y})^2 + (y_i - \widehat{y}_i)^2 + 2(\widehat{y}_i - \overline{y})(y_i - \widehat{y}_i)$$

$$(\widehat{y}_i - \overline{y})(y_i - \widehat{y}_i) = (x_i - \overline{x})\, w_1 \left[(y_i - \overline{y}) - (x_i - \overline{x})\, w_1\right]$$

$$= \left[(x_i - \overline{x})(y_i - \overline{y}) - (x_i - \overline{x})^2 w_1\right] w_1$$

Thus $\sum_{i=1}^{n} \left[(\widehat{y}_i - \overline{y})(y_i - \widehat{y}_i)\right] = (S_{xy} - S_{xx} w_1)\, w_1 = 0$

Therefore, $\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2 + \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$

$$SS_T = SS_R + SS_{res}$$

SS_R : sum of square (regression)

SS_{res} : sum of square (residual)

SS_T : sum of square (total)

$$w_0 = \overline{y} - \overline{x} w_1$$

$$\widehat{y} = w_0 + w_1 x = \overline{y} + (x - \overline{x})\, w_1$$

$$\widehat{y} - \overline{y} = (x - \overline{x})\, w_1$$

$$y - \widehat{y} = (y - \overline{y}) - (x - \overline{x})\, w_1$$

$$w_1 = \frac{S_{xy}}{S_{xx}} \quad S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$$

$$S_{xy} = \sum_{i=1}^{n} \left[(x_i - \overline{x})(y_i - \overline{y})\right]$$

Malay K. Das, mkdas@iitk.ac.in

$$SS_T = SS_R + SS_{res} \quad \sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2 + \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

$$SS_R = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2 = \sum_{i=1}^{n} \left[(x - \overline{x})\, w_1\right]^2 = S_{xx} w_1^2 = S_{xy} w_1$$

Coefficient of Determination (Goodness of Fit) R^2

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{res}}{SS_T} \qquad 0 \le R^2 \le 1$$

$$R^2 = \frac{SS_R}{SS_T} = \frac{S_{xy} w_1}{S_{yy}} = \frac{(S_{xy})^2}{S_{xx} S_{yy}}$$

Correlation Coefficient

$$0 \le \frac{(S_{xy})^2}{S_{xx} S_{yy}} \le 1 \quad \Rightarrow -1 \le \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \le 1 \quad \Rightarrow -1 \le R \le 1$$

$$R = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

$$S_{xy} = \sum_{i=1}^{n} \left[(x_i - \overline{x})(y_i - \overline{y})\right]$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

$$w_1 = \frac{S_{xy}}{S_{xx}}$$

$$w_0 = \bar{y} - \bar{x} w_1$$

$$\widehat{y} = w_0 + w_1 x$$

$$\widehat{y} - \overline{y} = (x - \overline{x})\, w_1$$

$$y - \widehat{y} = (y - \bar{y}) - (x - \bar{x})\, w_1$$

$$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2 \,;\, SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \,;\, SS_{res} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$R^2 = 1 - \frac{SS_{res}}{SS_T} \qquad 0 \le R^2 \le 1 \qquad\qquad R = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \qquad -1 \le R \le 1$$

$R^2 \to 1:$ regression line runs close to all data points
variation in *y* is captured well by the regression line

$R^2 \to 0:$ regression line fails to capture the variation in *y*

$R > 0:$ +ve correlation, *y* increases with *x*

$R < 0:$ -ve correlation, *y* decreases with *x*

$R = 0:$ no correlation, *y* and *x* are not linearly dependent

$$S_{xy} = \sum_{i=1}^{n} \left[(x_i - \overline{x}) (y_i - \overline{y}) \right]$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

$$w_0 = \bar{y} - \bar{x} w_1$$

$$\hat{y} = w_0 + w_1 x$$

High value of R^2 is not necessarily good, may indicate overfitting; model may not work well with unseen data

Rocket propellant problem

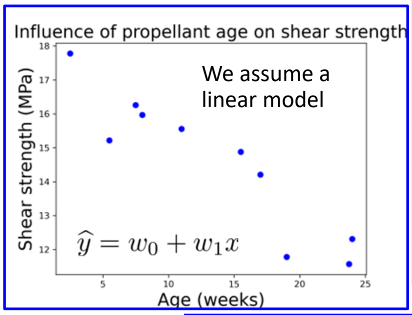
$$w_0, w_1 \leftarrow \underset{w_0, w_1}{\arg\min} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \overbrace{(y_i - w_0 - w_1 x_i)^2}^{\text{loss function}}}_{\text{cost function}}$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = 13.375$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = 14.554$$

$$S_{xy} = \sum_{i=1}^{n} \left[(x_i - \bar{x}) \, y_i \right] = -131.31$$

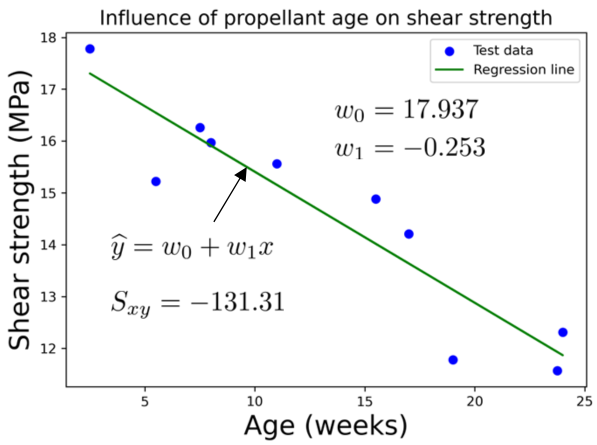
$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = 519.15625$$



Shear strength of rocket propellant

Test	Propellant age (Weeks) x_i	Shear strength (MPa) y_i
1	15.5	14.88
2	23.75	11.57
3	8	15.97
4	17	14.21
5	5.5	15.22
6	19	11.78
7	24	12.31
8	2.5	17.78
9	7.5	16.26
10	11	15.56

$$w_1 = \frac{S_{xy}}{S_{xx}} = -0.253 \qquad w_0 = \bar{y} - \bar{x} w_1 = 17.937$$



Shear strength of rocket propellant

Test	Propellant age (Weeks) x_i	Shear strength (MPa) y_i
1	15.5	14.88
2	23.75	11.57
3	8	15.97
4	17	14.21
5	5.5	15.22
6	19	11.78
7	24	12.31
8	2.5	17.78
9	7.5	16.26
10	11	15.56

Correlation coefficient $R = \dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = -0.927$

Coefficient of determination $R^2 = 0.86$

x and y are negatively correlated, and the regression line runs close enough to capture the variation

Simple Linear Regression fits $\widehat{y} = w_0 + w_1 x$ over data $\mathcal{T} = \{(x_i, y_i)\}_{i=1}^n$

Most regression problems includes multiple features

where training dataset: $\mathcal{T} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ feature: $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_k \end{bmatrix}^T$

Multiple Linear Regression: learning a linear model with vector feature

Let us take a simple case where $\mathbf{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^T$ training data: $\mathcal{T} = \{(x_{i1}, x_{i2}, y_i)\}_{i=1}^n$

We assume a linear model $\widehat{y} = w_0 + w_1 x_1 + w_2 x_2$

We wish to minimize the **Least Square Cost Function**

$$E(w_0, w_1, w_2) = \frac{1}{n}\sum_{i=1}^{n}(\widehat{y}_i - y_i)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(w_0 + w_1 x_{i1} + w_2 x_{i2} - y_i)^2$$

Multiple Linear Regression $\widehat{y} = w_0 + w_1 x_1 + w_2 x_2$ data: $\mathcal{T} = \{(x_{i1}, x_{i2}, y_i)\}_{i=1}^{n}$

Least Square Cost Function $E(w_0, w_1, w_2) = \frac{1}{n}\sum_{i=1}^{n}(w_0 + w_1 x_{i1} + w_2 x_{i2} - y_i)^2$

$$\frac{\partial E}{\partial w_0} = 0 = \frac{2}{n}\sum_{i=1}^{n}(w_0 + w_1 x_{i1} + w_2 x_{i2} - y_i) \Rightarrow n w_0 + w_1 \sum_{i=1}^{n} x_{i1} + w_2 \sum_{i=1}^{n} x_{i2} = \sum_{i=1}^{n} y_i$$

$$\frac{\partial E}{\partial w_1} = 0 = \frac{2}{n}\sum_{i=1}^{n}(w_0 + w_1 x_{i1} + w_2 x_{i2} - y_i)\, x_{i1} \Rightarrow w_0 \sum_{i=1}^{n} x_{i1} + w_1 \sum_{i=1}^{n} x_{i1}^2 + w_2 \sum_{i=1}^{n} x_{i1} x_{i2} = \sum_{i=1}^{n} x_{i1} y_i$$

$$\frac{\partial E}{\partial w_2} = 0 = \frac{2}{n}\sum_{i=1}^{n}(w_0 + w_1 x_{i1} + w_2 x_{i2} - y_i)\, x_{i2} \Rightarrow w_0 \sum_{i=1}^{n} x_{i2} + w_1 \sum_{i=1}^{n} x_{i2} x_{i1} + w_2 \sum_{i=1}^{n} x_{i2}^2 = \sum_{i=1}^{n} x_{i2} y_i$$

Solving the above three normal equations, we find w_0, w_1, w_2

Setting $x_1 = x, x_2 = x^2$ we can fit polynomial $\widehat{y} = w_0 + w_1 x + w_2 x^2$

Though counter-intuitive, polynomial fitting is also a **linear** regression problem

Let's now generalize for linear egression with $\mathbf{x} = [x_1 \quad x_2 \quad \cdots \quad x_k]^T$

Linear model:

$$\widehat{y} = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_k x_k$$

for $i = 1, 2, \cdots, n$

$$\widehat{y}_i = w_0 + w_1 x_{i1} + w_2 x_{i2} + \cdots + w_k x_{ik} = w_0 + \sum_{j=1}^{k} w_j x_{ij}$$

Observations i	Label y	Features			
		x_1	x_2	$\cdots$	x_k
1	y_1	x_{11}	x_{12}	$\cdots$	x_{1k}
2	y_2	x_{21}	x_{22}	$\cdots$	x_{2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_n	x_{n1}	x_{n2}	$\cdots$	x_{nk}

We wish to minimize the **least Square Cost Function**

$$E(w_0, w_1, \cdots, w_k) = \frac{1}{n}\sum_{i=1}^{n}\left(w_0 + \sum_{j=1}^{k} w_j x_{ij} - y_i\right)^2 \qquad \frac{\partial E}{\partial w_0} = 0 = \frac{2}{n}\sum_{i=1}^{n}\left(w_0 + \sum_{j=1}^{k} w_j x_{ij} - y_i\right)$$

$$\frac{\partial E}{\partial w_p} = 0 = \frac{2}{n}\sum_{i=1}^{n}\left(w_0 + \sum_{j=1}^{k} w_j x_{ij} - y_i\right) x_{ip} \qquad p = 1, 2, \cdots, k$$

$\rightarrow$ $k+1$ equations for $w_0, w_1, w_2, \cdots, w_k$

$$\frac{\partial E}{\partial w_0} = 0 = \frac{2}{n}\sum_{i=1}^{n}\left(w_0 + \sum_{j=1}^{k} w_j x_{ij} - y_i\right) \qquad \frac{\partial E}{\partial w_p} = 0 = \frac{2}{n}\sum_{i=1}^{n}\left(w_0 + \sum_{j=1}^{k} w_j x_{ij} - y_i\right) x_{ip}$$

$$p = 1, 2, \cdots, k$$

$$\boxed{\begin{aligned}
& n w_0 + w_1 \sum_{i=1}^{n} x_{i1} + w_2 \sum_{i=1}^{n} x_{i2} + \cdots + w_k \sum_{i=1}^{n} x_{ik} = \sum_{i=1}^{n} y_i \\
& w_0 \sum_{i=1}^{n} x_{i1} + w_1 \sum_{i=1}^{n} x_{i1}^2 + w_2 \sum_{i=1}^{n} x_{i2}x_{i1} + \cdots + w_k \sum_{i=1}^{n} x_{ik}x_{i1} = \sum_{i=1}^{n} x_{i1} y_i \\
& w_0 \sum_{i=1}^{n} x_{i2} + w_1 \sum_{i=1}^{n} x_{i1}x_{i2} + w_2 \sum_{i=1}^{n} x_{i2}^2 + \cdots + w_k \sum_{i=1}^{n} x_{ik}x_{i2} = \sum_{i=1}^{n} x_{i2} y_i \\
& \vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
& w_0 \sum_{i=1}^{n} x_{ik} + w_1 \sum_{i=1}^{n} x_{i1}x_{ik} + w_2 \sum_{i=1}^{n} x_{i2}x_{ik} + \cdots + w_k \sum_{i=1}^{n} x_{ik}^2 = \sum_{i=1}^{n} x_{ik} y_i
\end{aligned}}$$

$k + 1$ equations for $w_0, w_1, w_2, \cdots, w_k$

Some compact notation could be more useful

Malay K. Das, mkdas@iitk.ac.in

$$\frac{\partial E}{\partial w_0} = 0 = \frac{2}{n}\sum_{i=1}^{n}\left(w_0 + \sum_{j=1}^{k} w_j x_{ij} - y_i\right) = 2\left(w_0 + \sum_{j=1}^{k} w_j \overline{x_j} - \overline{y}\right) \qquad p = 1, 2, \cdots, k$$

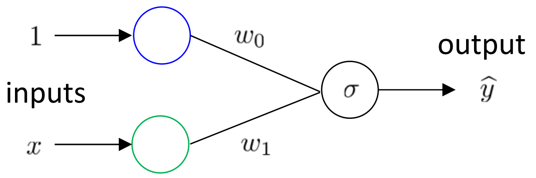
$$\frac{\partial E}{\partial w_p} = 0 = \frac{2}{n}\sum_{i=1}^{n}\left(w_0 + \sum_{j=1}^{k} w_j x_{ij} - y_i\right) x_{ip} = 2\left(\overline{x_p} w_0 + \sum_{j=1}^{k} w_j \overline{x_j x_p} - \overline{x_p y}\right)$$

$$\boxed{\begin{aligned}
& w_0 + w_1 \overline{x_1} + w_2 \overline{x_2} + \cdots + w_k \overline{x_k} = \overline{y} \\
& \overline{x_1} w_0 + \overline{x_1^2} w_1 + \overline{x_1 x_2} w_2 + \cdots + \overline{x_1 x_k} w_k = \overline{x_1 y} \\
& \overline{x_2} w_0 + \overline{x_2 x_1} w_1 + \overline{x_2^2} w_2 + \cdots + \overline{x_2 x_k} w_k = \overline{x_2 y} \\
& \vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
& \overline{x_k} w_0 + \overline{x_k x_1} w_1 + \overline{x_k x_2} w_2 + \cdots + \overline{x_k^2} w_k = \overline{x_k y}
\end{aligned}}$$

Formulation as well as solution procedure may be greatly improved by using a vector-matrix notation

Simple Linear Regression: $y = \widehat{y} + e \quad \widehat{y} = w_0 + w_1 x \quad \widehat{y}:$ least square estimation of y

Training dataset: $\mathcal{T} = \{(x_i, y_i)\}_{i=1}^{n}$



To find w_0, w_1

$$w_0, w_1 \leftarrow \underbrace{\arg\min_{w_0, w_1} \frac{1}{n} \sum_{i=1}^{n} \overbrace{(\widehat{y}_i - y_i)^2}^{\text{loss function}}}_{\text{cost function}}$$

---

let us write our input as a vector $\mathbf{x} = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} x_0 & x_1 \end{bmatrix}^T = \begin{bmatrix} 1 & x_1 \end{bmatrix}^T$

for simple linear regression $\quad x_0 = 1 \quad x_1 = x$

regression coefficients may also be treated as a vector $\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} \underset{\text{bias}}{w_0} & \underset{\text{weight}}{w_1} \end{bmatrix}^T$

Thus $\widehat{y} = w_0 + w_1 x = \mathbf{x}^T \mathbf{w}$ (inner/dot product)

Simple Linear Regression: $y = \widehat{y} + e \quad \widehat{y} = w_0 + w_1 x \quad \widehat{y}:$ least square estimation of y

Training dataset: $\mathcal{T} = \{(x_i, y_i)\}_{i=1}^{n}$

$\mathbf{x} = \begin{bmatrix} 1 & x_1 \end{bmatrix}^T \quad \mathbf{w} = \begin{bmatrix} w_0 & w_1 \end{bmatrix}^T \quad \widehat{y} = \mathbf{x}^T \mathbf{w}$

$$\mathbf{w} \leftarrow \underbrace{\arg\min_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} \overbrace{\left(\mathbf{x}_i^T \mathbf{w} - y_i\right)^2}^{\text{loss function}}}_{\text{cost function}}$$

now $\quad \widehat{y}_i = \mathbf{x}_i^T \mathbf{w} \qquad \mathbf{x}_i = \begin{bmatrix} 1 & x_{i1} \end{bmatrix}^T \quad i = 1, 2, \cdots, n$

$\widehat{y}_1 = \mathbf{x}_1^T \mathbf{w} \qquad \mathbf{x}_1 = \begin{bmatrix} 1 & x_{11} \end{bmatrix}^T$

$\widehat{y}_2 = \mathbf{x}_2^T \mathbf{w} \qquad \mathbf{x}_2 = \begin{bmatrix} 1 & x_{21} \end{bmatrix}^T$ etc.

the label vector

$\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix}^T$

least square estimate of $\mathbf{y}$ is a vector as well

$$\widehat{\mathbf{y}} = \begin{bmatrix} \widehat{y}_1 \\ \widehat{y}_2 \\ \vdots \\ \widehat{y}_n \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^T \mathbf{w} \\ \mathbf{x}_2^T \mathbf{w} \\ \vdots \\ \mathbf{x}_n^T \mathbf{w} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \vdots & \vdots \\ 1 & x_{n1} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \Rightarrow \widehat{\mathbf{y}} = \mathbf{X}\mathbf{w}$$

where $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \vdots & \vdots \\ 1 & x_{n1} \end{bmatrix}$

Simple Linear Regression: $\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix}^T$ $\quad \widehat{\mathbf{y}} = \begin{bmatrix} \widehat{y}_1 & \widehat{y}_2 & \cdots & \widehat{y}_n \end{bmatrix}^T$

$\mathbf{x} = \begin{bmatrix} 1 & x_1 \end{bmatrix}^T \quad \mathbf{w} = \begin{bmatrix} w_0 & w_1 \end{bmatrix}^T \quad \widehat{y} = \mathbf{x}^T \mathbf{w}$

Least square estimate $\quad \widehat{\mathbf{y}} = \mathbf{X}\mathbf{w}$

$$\mathbf{w} \leftarrow \underbrace{\arg\min_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} \overbrace{\left(\widehat{y}_i - y_i\right)^2}^{\text{loss function}}}_{\text{cost function}}$$

Cost function

$$E\left(\mathbf{w}\right) = \frac{1}{n} \sum_{i=1}^{n} \left(\widehat{y}_i - y_i\right)^2 = \frac{1}{n} \left(\widehat{\mathbf{y}} - \mathbf{y}\right)^T \left(\widehat{\mathbf{y}} - \mathbf{y}\right)$$

$$= \frac{1}{n} \left(\mathbf{X}\mathbf{w} - \mathbf{y}\right)^T \left(\mathbf{X}\mathbf{w} - \mathbf{y}\right)$$

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \vdots & \vdots \\ 1 & x_{n1} \end{bmatrix}$$

To minimize the cost function, we now enforce

$$\nabla E\left(\mathbf{w}\right) = \frac{\partial E}{\partial \mathbf{w}} = \begin{bmatrix} \frac{\partial E}{\partial w_0} & \frac{\partial E}{\partial w_1} \end{bmatrix}^T = \mathbf{0} \quad \text{(zero gradient)}$$

Malay K. Das, mkdas@iitk.ac.in

Linear regression $\quad \widehat{\mathbf{y}} = \mathbf{X}\mathbf{w}$

$$E\left(\mathbf{w}\right) = \frac{1}{n} \left(\mathbf{X}\mathbf{w} - \mathbf{y}\right)^T \left(\mathbf{X}\mathbf{w} - \mathbf{y}\right)$$

We wish to find

$$\arg\min_{\mathbf{w}} E\left(\mathbf{w}\right)$$

$\widehat{\mathbf{y}}$: nD vector

$\mathbf{X}$: n×2 matrix $\quad n > 2$

$\mathbf{w}$: 2D vector (to be evaluated)

Simple linear regression

$$E = \frac{1}{n} \sum_{i=1}^{n} \left(w_0 \underbrace{x_{i0}}_{=1} + w_1 x_{i1} - y_i\right)^2$$

$$\mathbf{X}^T = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \end{bmatrix}$$

$$\frac{\partial E}{\partial w_0} = \frac{2}{n} \sum_{i=1}^{n} \left(w_0 x_{i0} + w_1 x_{i1} - y_i\right)$$

$$= \frac{2}{n} \sum_{i=1}^{n} x_{i0} \left(\widehat{y}_i - y_i\right)$$

$$\frac{\partial E}{\partial w_1} = \frac{2}{n} \sum_{i=1}^{n} x_{i1} \left(w_0 + w_1 x_{i1} - y_i\right)$$

$$= \frac{2}{n} \sum_{i=1}^{n} x_{i1} \left(\widehat{y}_i - y_i\right)$$

gradient

$$\nabla E\left(\mathbf{w}\right) = \frac{\partial E}{\partial \mathbf{w}} = \begin{bmatrix} \frac{\partial E}{\partial w_0} \\ \frac{\partial E}{\partial w_1} \end{bmatrix} = \frac{2}{n} \mathbf{X}^T \left(\widehat{\mathbf{y}} - \mathbf{y}\right) = \frac{2}{n} \mathbf{X}^T \left(\mathbf{X}\mathbf{w} - \mathbf{y}\right)$$

minimization of E requires $\quad \nabla E\left(\mathbf{w}\right) = \mathbf{0}$

$$\Rightarrow \frac{2}{n} \mathbf{X}^T \left(\mathbf{X}\mathbf{w} - \mathbf{y}\right) = \mathbf{0} \quad \Rightarrow \mathbf{X}^T \mathbf{X}\mathbf{w} = \mathbf{X}^T \mathbf{y}$$

$$\boxed{\mathbf{w} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y}} \quad \text{and} \quad \boxed{\widehat{\mathbf{y}} = \mathbf{X} \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y}}$$

Consider multiple linear regression with k features: $1 < k < n$

Model: $\widehat{y} = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_k x_k$

Observations i	Response y	Regressors			
		x_1	x_2	$\cdots$	x_k
1	y_1	x_{11}	x_{12}	$\cdots$	x_{1k}
2	y_2	x_{21}	x_{22}	$\cdots$	x_{2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_n	x_{n1}	x_{n2}	$\cdots$	x_{nk}

Regressor as a **vector**

$$\mathbf{x} = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix}$$

$$\boxed{\widehat{y} = \mathbf{x}^T \mathbf{w}}$$

Weight as a **vector**

$$\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_k \end{bmatrix}$$

also, each observation $\widehat{y}_i = w_0 + \sum_{j=1}^{k} w_j x_{ij} \qquad i = 1, 2, \cdots, n$

features at each observation form a **vector** $\mathbf{x}_i = \begin{bmatrix} 1 \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{ik} \end{bmatrix}$

$$\boxed{\widehat{y}_i = \mathbf{x}_i^T \mathbf{w}}$$

transposes of these vectors are often important

$$\mathbf{x}^T = \begin{bmatrix} 1 & x_1 & x_2 & \ldots & x_k \end{bmatrix}$$
$$\mathbf{x}_i^T = \begin{bmatrix} 1 & x_{i1} & x_{i2} & \ldots & x_{ik} \end{bmatrix}$$
$$\mathbf{w}^T = \begin{bmatrix} w_0 & w_1 & w_2 & \ldots & w_k \end{bmatrix}$$

Malay K. Das, mkdas@iitk.ac.in

Consider multiple linear regression with k features: $1 < k < n$

Model: $\widehat{y} = \mathbf{w}^T \mathbf{x}$

$\widehat{y}_i = \mathbf{w}^T \mathbf{x}_i$

$i = 1, 2, \cdots, n$

Observations i	Response y	Regressors			
		x_1	x_2	$\cdots$	x_k
1	y_1	x_{11}	x_{12}	$\cdots$	x_{1k}
2	y_2	x_{21}	x_{22}	$\cdots$	x_{2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_n	x_{n1}	x_{n2}	$\cdots$	x_{nk}

$$\mathbf{x}^T = \begin{bmatrix} 1 & x_1 & x_2 & \ldots & x_k \end{bmatrix}$$
$$\mathbf{x}_i^T = \begin{bmatrix} 1 & x_{i1} & x_{i2} & \ldots & x_{ik} \end{bmatrix}$$
$$\mathbf{w}^T = \begin{bmatrix} w_0 & w_1 & w_2 & \ldots & w_k \end{bmatrix}$$

The response is also a **vector**

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \widehat{\mathbf{y}} = \begin{bmatrix} \widehat{y}_1 \\ \widehat{y}_2 \\ \vdots \\ \widehat{y}_n \end{bmatrix}$$

$\widehat{y}_i = \mathbf{x}_i^T \mathbf{w} \qquad i = 1, 2, \cdots, n \qquad n > k$

$$\boxed{\widehat{\mathbf{y}} = \mathbf{X}\mathbf{w}}$$

Regressors together forms a **matrix**

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix}$$

This regression problem finds the least square estimates

$$\arg\min_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{x}_i^T \mathbf{w} - y_i \right)^2$$

Linear regression $\widehat{\mathbf{y}} = \mathbf{X}\mathbf{w}$

The problem cannot have a unique solution, in general

$\widehat{\mathbf{y}}$: n-dimensional vector

$\mathbf{X}$: n×k matrix $n > k$

$\mathbf{w}$: k-dimensional vector (to be evaluated)

we minimize the Least Square cost function

$$E(\mathbf{w}) = \frac{1}{n}(\mathbf{X}\mathbf{w} - \mathbf{y})^T(\mathbf{X}\mathbf{w} - \mathbf{y})$$

$$\nabla E(\mathbf{w}) = \frac{2}{n}\mathbf{X}^T(\mathbf{X}\mathbf{w} - \mathbf{y})$$

$$\nabla E(\mathbf{w}) = \mathbf{0} \Rightarrow \mathbf{X}^T\mathbf{X}\mathbf{w} = \mathbf{X}^T\mathbf{y}$$

$$\boxed{\mathbf{w} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}}$$

We wish to find

$$\arg\min_{\mathbf{w}} E(\mathbf{w})$$

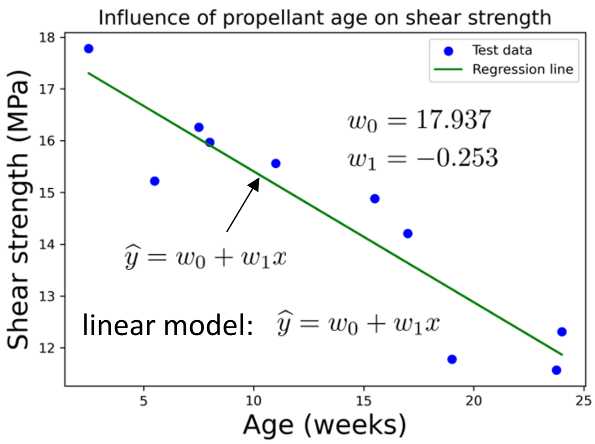
$$\boxed{\widehat{\mathbf{y}} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}}$$

least square estimation

in short, linear regression approximately solves $\mathbf{X}\mathbf{w} = \mathbf{y}$ where $\mathbf{X}$ is a rectangular $m \times n$ matrix with $m > n$

minimizing $\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 \Rightarrow \mathbf{X}^T\mathbf{X}\mathbf{w} = \mathbf{X}^T\mathbf{y}$ existence of closed form of solution is very useful

Let’s revisit the example problem using **vector-matrix notation**



Shear strength of rocket propellant

Test	Propellant age (Weeks) x_i	Shear strength (MPa) y_i
1	15.5	14.88
2	23.75	11.57
3	8	15.97
4	17	14.21
5	5.5	15.22
6	19	11.78
7	24	12.31
8	2.5	17.78
9	7.5	16.26
10	11	15.56

$$\widehat{y} = w_0 + w_1 x \quad = \mathbf{x}^T \mathbf{w} \qquad \mathbf{x} = \begin{bmatrix} 1 \\ x \end{bmatrix} \quad \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} 1 & 15.5 \\ 1 & 23.75 \\ \vdots & \vdots \\ 1 & 11 \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{10} \end{bmatrix} = \begin{bmatrix} 14.88 \\ 11.57 \\ \vdots \\ 15.56 \end{bmatrix}$$

$$\mathbf{X}^T = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 15.5 & 23.75 & \dots & 11 \end{bmatrix}$$

$$\mathbf{w} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y} \;= \begin{bmatrix} 17.937 \\ -0.253 \end{bmatrix}$$

we usually don't invert $\mathbf{X}^T\mathbf{X}$ directly

instead we solve $\left(\mathbf{X}^T\mathbf{X}\right) \mathbf{w} = \mathbf{X}^T \mathbf{y}$

unfortunately, the matrix $\mathbf{X}^T\mathbf{X}$ is not always well-behaved

Shear strength of rocket propellant

Test	Propellant age (Weeks) x_i	Shear strength (MPa) y_i
1	15.5	14.88
2	23.75	11.57
3	8	15.97
4	17	14.21
5	5.5	15.22
6	19	11.78
7	24	12.31
8	2.5	17.78
9	7.5	16.26
10	11	15.56

Malay K. Das, mkdas@iitk.ac.in