Terminology recap

- Define the **design matrix** $X \in \mathbb{R}^{n \times 2}$, **observation vector** $\vec{y} \in \mathbb{R}^n$, and **parameter vector** $\vec{w} \in \mathbb{R}^2$ as follows:

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \qquad \vec{y} = \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{\text{vector of actual } y \text{ values}} \qquad \vec{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

$y_i = w_0 (1) + w_1 (x_i)$

Rewriting mean squared error

- The mean squared error of the predictions in $\vec{h}$ is:

$$R_{\text{sq}}(w_0, w_1) = \frac{1}{n} \sum_{i=1}^{n} (y_i - (w_0 + w_1 x_i))^2$$

$\Updownarrow$ Equivalent!

- Equivalently, we have:

$$R_{\text{sq}}(\vec{w}) = \frac{1}{n} \|\vec{y} - \underbrace{X\vec{w}}_{\vec{h}}\|^2$$

Minimizing mean squared error

- To minimize mean squared error, $R_{\text{sq}}(\vec{w}) = \frac{1}{n}\|\vec{y} - X\vec{w}\|^2$, we must choose the $\vec{w}^*$ such that the error vector, $\vec{e} = \vec{y} - X\vec{w}^*$, is orthogonal to the columns of X.

- Equivalently, we have:

$$X^T\vec{e} = 0$$

$$\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}_{n \times 2} \qquad \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}_{n \times 1} = \begin{bmatrix} y_1 - H(x_1) \\ y_2 - H(x_2) \\ \vdots \end{bmatrix}$$

What are the dimensions of $X^T\vec{e}$?

$\hookrightarrow (2 \times n) \times (n \times 1)$

$= (2 \times 1)$

Minimizing mean squared error

- To minimize mean squared error, $R_{\text{sq}}(\vec{w}) = \frac{1}{n}\|\vec{y} - X\vec{w}\|^2$, we must choose the $\vec{w}^*$ such that the error vector, $\vec{e} = \vec{y} - X\vec{w}^*$, is orthogonal to the columns of X.

- Equivalently, we have:

$$X^T\vec{e} = 0$$

What are the dimensions of $X^T\vec{e}$?

$$\begin{bmatrix} | & | & \cdots & | \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

- Expanding, we have:

$$\vec{e} = \vec{y} - X\vec{w}$$

$$X^T(\vec{y} - X\vec{w}^*) = 0$$

$$X^T\vec{y} - X^TX\vec{w}^* = 0$$

$$X^TX\vec{w}^* = X^T\vec{y}$$

$$R_{\text{sq}}(w_0, w_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - (w_0 + w_1 x_i))^2$$

- We found, using calculus, that:
  - $\boxed{w_1^* = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = r\frac{\sigma_y}{\sigma_x}}$.
  - $\boxed{w_0^* = \bar{y} - w_1^* \bar{x}}$.

Equivalent!

HW problem?

- Another way of finding optimal model parameters for simple linear regression is to find the $\vec{w}^*$ that minimizes $R_{\text{sq}}(\vec{w}) = \frac{1}{n}\|\vec{y} - X\vec{w}\|^2$.

- The minimizer, if X^TX is invertible, is the vector $\boxed{\vec{w}^* = (X^TX)^{-1}X^T\vec{y}}$.

  If not, solve the **normal equations** from the previous slide.

The hypothesis vector

- When our hypothesis function is of the form:

$$H(\text{departure hour}, \text{day of month}) = w_0 + w_1 \cdot \text{departure hour} + w_2 \cdot \text{day of month}$$

the hypothesis vector $\vec{h} \in \mathbb{R}^n$ can be written as:

$$\vec{h} = \begin{bmatrix} H(\text{departure hour}_1, \text{day}_1) \\ H(\text{departure hour}_2, \text{day}_2) \\ \dots \\ H(\text{departure hour}_n, \text{day}_n) \end{bmatrix} = \underbrace{\overset{X}{\begin{bmatrix} 1 & \text{departure hour}_1 & \text{day}_1 \\ 1 & \text{departure hour}_2 & \text{day}_2 \\ \dots & \dots & \dots \\ 1 & \text{departure hour}_n & \text{day}_n \end{bmatrix}}}_{\text{3 columns!}} \overset{\vec{w}}{\begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix}}$$

The general solution

- Define the **design matrix** $X \in \mathbb{R}^{n\times(d+1)}$ and **observation vector** $\vec{y} \in \mathbb{R}^n$:

$$X = \underbrace{\begin{bmatrix} 1 & x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(d)} \\ 1 & x_2^{(1)} & x_2^{(2)} & \dots & x_2^{(d)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n^{(1)} & x_n^{(2)} & \dots & x_n^{(d)} \end{bmatrix}}_{n\times(d+1)} = \begin{bmatrix} \text{Aug}(\vec{x_1})^T \\ \text{Aug}(\vec{x_2})^T \\ \vdots \\ \text{Aug}(\vec{x_n})^T \end{bmatrix} \qquad \vec{y} = \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{n\times 1}$$

23 . 1