$$SSE = \underbrace{\vec{y}^T}_{1\times n}\underbrace{\vec{y}}_{n\times 1} - 2\underbrace{\vec{w}^T}_{1\times(p+1)}\underbrace{X^T}_{(p+1)\times n}\underbrace{\vec{y}}_{n\times 1} + \underbrace{\vec{w}^T}_{1\times(p+1)}\underbrace{X^T}_{(p+1)\times n}\underbrace{X}_{n\times(p+1)}\underbrace{\vec{w}}_{(p+1)\times 1}$$

$$\frac{\partial SSE}{\partial \vec{w}} := \begin{bmatrix} \frac{\partial SSE}{\partial w_0} \\ \frac{\partial SSE}{\partial w_1} \\ \vdots \\ \frac{\partial SSE}{\partial w_p} \end{bmatrix} \overset{set}{=} \vec{0}_{p+1}$$ and solve for b_0, b_1,..., b_p

let $\vec{x} \in \mathbb{R}^n$. Let $a \in \mathbb{R}$ be a constant wrt $\vec{x}$. $\Rightarrow \frac{\partial}{\partial \vec{x}}[a] = \vec{0}_n$ (0)

let $\vec{a} \in \mathbb{R}^n$ constant wrt $\vec{x}$

(1) $$\frac{\partial}{\partial \vec{x}}\Big[\underset{\vec{x}^T\vec{a}}{\overset{}{\vec{a}^T\vec{x}}}\Big] = \begin{bmatrix} \partial/\partial x_1 [a_1x_1 + a_2x_2 + \dots + a_nx_n] \\ \vdots \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \vec{a} \neq \vec{a}^T$$

let $a, b \in \mathbb{R}$ constants wrt $\vec{x}$

(2) $$\frac{\partial}{\partial \vec{x}}\left[a f(\vec{x}) + b g(\vec{x})\right] = \begin{bmatrix} \partial/\partial x_1 [a f(\vec{x}) + b g(\vec{x})] \\ \vdots \end{bmatrix} = \begin{bmatrix} a \frac{\partial}{\partial x_1}[f(\vec{x})] + b\frac{\partial}{\partial x_1}[g(\vec{x})] \\ \vdots \end{bmatrix}$$

$$= a\frac{\partial}{\partial \vec{x}}[f(\vec{x})] + b\frac{\partial}{\partial \vec{x}}[g(\vec{x})]$$

let $A \in \mathbb{R}^{n\times n}$, symmetric, constant wrt $\vec{x}$

(3) $$\frac{\partial}{\partial \vec{x}}\left[\vec{x}^T A \vec{x}\right], \quad A\vec{x} = \begin{bmatrix} \leftarrow \vec{a}_1 \rightarrow \\ \leftarrow \vec{a}_2 \rightarrow \\ \vdots \\ \leftarrow \vec{a}_n \rightarrow \end{bmatrix}\begin{bmatrix} \uparrow \\ \vec{x} \\ \downarrow \end{bmatrix} = \begin{bmatrix} \vec{a}_1\cdot\vec{x} \\ \vec{a}_2\cdot\vec{x} \\ \vdots \\ \vec{a}_n\cdot\vec{x} \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n \\ \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nn}x_n \end{bmatrix}$$

This scalar expression, x-tr A x is called a "quadratic form" and it's a common expression and very well-studied.

$$\vec{x}^T(A\vec{x}) = [x_1\ x_2\ \cdots\ x_n]\begin{bmatrix} \vec{a}_1\cdot\vec{x} \\ \vec{a}_2\cdot\vec{x} \\ \vdots \\ \vec{a}_n\cdot\vec{x} \end{bmatrix} = x_1\vec{a}_1\cdot\vec{x} + x_2\vec{a}_2\cdot\vec{x} + \dots + x_n\vec{a}_n\cdot\vec{x}$$

$$= x_1(a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n) + x_2(a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n) + \dots\dots + x_n(a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nn}x_n)$$

$$\frac{\partial}{\partial x_1}[\ ] = 2a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n + a_{21}x_2 + \dots\dots + a_{n1}x_n = 2a_{11}x_1 + 2a_{12}x_2 + \dots + 2a_{1n}x_n$$

$$= 2\vec{a}_1\cdot\vec{x}$$

$$\frac{\partial}{\partial x_2}[\ ] = a_{12}x_1 + a_{21}x_1 + 2a_{22}x_2 + \dots + a_{2n}x_n + \dots + a_{n2}x_n = 2a_{12}x_1 + 2a_{22}x_2 + \dots + 2a_{n2}x_n$$

$$= 2\vec{a}_2\cdot\vec{x}$$

$\vdots$

$$\frac{\partial}{\partial \vec{x}}\left[\vec{x}^T A\vec{x}\right] = \begin{bmatrix} 2\vec{a}_1\cdot\vec{x} \\ 2\vec{a}_2\cdot\vec{x} \\ \vdots \\ 2\vec{a}_n\cdot\vec{x} \end{bmatrix} = 2A\vec{x}$$

$$\frac{\partial}{\partial \vec{w}}\Big[\underbrace{\vec{y}^T\vec{y} - 2\vec{w}^TX^T\vec{y} + \vec{w}^TX^TX\vec{w}}_{SSE}\Big] \overset{\text{rule \#2}}{=} \frac{\partial}{\partial \vec{w}}\Big[\cancel{\vec{y}^T\vec{y}}\Big]^{0 \text{ by rule \#0}} - 2\frac{\partial}{\partial \vec{w}}\left[\vec{w}^T(X^T\vec{y})\right] + \frac{\partial}{\partial \vec{w}}\left[\vec{w}^TX^TX\vec{w}\right]$$

$(X^TX)^T = X^T(X^T)^T = X^TX \Rightarrow$ symm.

$$\overset{\text{rule \#1}}{=} -2X^T\vec{y} + \frac{\partial}{\partial \vec{w}}\left[\vec{w}^T(X^TX)\vec{w}\right] \overset{\text{rule \#3}}{=} -\cancel{2}X^T\vec{y} + \cancel{2}X^TX\vec{w} \overset{set}{=} \vec{0}_{p+1}$$ and solve for $\vec{b}$

$A\vec{x} = \vec{b}$

$$\Rightarrow \underbrace{\cancel{(X^TX)^{-1}X^TX}}_{I_{p+1}}\vec{w} = (X^TX)^{-1}X^T\vec{y} \Rightarrow \boxed{\vec{b} = (X^TX)^{-1}X^T\vec{y}} \Rightarrow \underbrace{\hat{y}_* = g(\vec{x}_*) = \vec{x}_*\vec{b}}_{\text{predictions}}$$

In order to compute the OLS coefficients (vector b), you need X^T X, a (p+1) x (p+1) square matrix, to be invertible. Equivalently, rank[X^T X] = p + 1 i.e. "full rank" i.e. all columns of X^T X are linearly independent. Since there's a thm: rank[X^T X] = rank[X], this means rank[X] = p + 1, i.e. the columns of X are linearly indep.

$$X = \underbrace{\begin{bmatrix} 1 & \uparrow & \uparrow & & \uparrow \\ \vdots & \vec{x}_{\cdot 1} & \vec{x}_{\cdot 2} & \cdots & \vec{x}_{\cdot p} \\ 1 & \downarrow & \downarrow & & \downarrow \end{bmatrix}}_{\text{feature measurements on all n subjects}}$$

If X is full rank that means there is no exact data duplication e.g. x_1: height measured in inches and x_2: height measured in centi-meters. What if you do have a feature that is linearly dependent with the other features in X? You just drop it. Then X will be full rank and you're good to estimate the OLS coefficients.

$$\vec{y} = \overset{= X\vec{b}}{\hat{\vec{y}}} + \vec{e} \Rightarrow \vec{e} = \vec{y} - \hat{\vec{y}}, \quad SSE = \sum_{i=1}^{n} e_i^2 = \vec{e}^T\vec{e}$$

$$MSE = \frac{1}{n-(p+1)}SSE, \quad RMSE = \sqrt{MSE}, \quad R^2 = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

$$= \frac{s_y^2 - s_e^2}{s_y^2} \quad (\text{Same}).$$

you sometimes say the model has p+1 "degrees of freedom" (i.e. the number of parameters, w_0, w_1, ..., w_p, is p + 1) and p + 1 = dim[colsp[X]].