Model: $f : \mathbf{x} \rightarrow y, \;\; f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$

Training data: $\{(\mathbf{x_1}, y_1), \ldots, (\mathbf{x_N}, y_N)\}$, or $\mathbf{X} = \begin{bmatrix} -\ \mathbf{x_1}^T\ - \\ \vdots \\ -\ \mathbf{x_N}^T\ - \end{bmatrix}$ and $\mathbf{y} = [y_1, \ldots, y_N]^T$

Evaluation through residual sum of squares (no regularization):
$J(\mathbf{w}) = RSS(\mathbf{w}) = \sum_{n=1}^{N} (y_n - \mathbf{w}^T\mathbf{x_n})^2 = (\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w})$

Evaluation through residual sum of squares ($l2$-norm regularization):

$$\begin{aligned} J(\mathbf{w}) &= RSS(\mathbf{w}) + \lambda \sum_{d=1}^{D} w_d^2 \\ &= RSS(\mathbf{w}) + \lambda \|\mathbf{w}\|_2^2 \\ &= (\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda \|\mathbf{w}\|_2^2 \\ &= \mathbf{y}^T\mathbf{y} - 2\mathbf{w}^T(\mathbf{X}^T\mathbf{y}) + \mathbf{w}^T(\mathbf{X}^T\mathbf{X})\mathbf{w} + \lambda \mathbf{w}^T\mathbf{w} \\ &= \mathbf{y}^T\mathbf{y} - 2\mathbf{w}^T(\mathbf{X}^T\mathbf{y}) + \mathbf{w}^T(\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_{D\times D})\mathbf{w} \end{aligned}$$

We compute the first-order derivative of the above cost and set it to zero:

$$\frac{\theta J(\mathbf{w})}{\theta \mathbf{w}} = 0 \Rightarrow -2(\mathbf{X}^T\mathbf{y}) + 2(\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_{D\times D})\mathbf{w} = 0 \quad \Rightarrow \mathbf{w}^* = (\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_{D\times D})^{-1}\mathbf{X}^T\mathbf{y}$$

The same calculations hold for non-linear regression, where we can substitute $\mathbf{X}$ with $\mathbf{\Phi}$.