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

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: $RSS(\mathbf{w}) = \sum_{n=1}^{N} (y_n - \mathbf{w}^T\mathbf{x_n})^2 = (\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w})$ (residual sum of squares)

Optimization through analytic solution (ordinary least squares solution):
$\frac{\vartheta RSS(\mathbf{w})}{\vartheta \mathbf{w}} = 0 \Rightarrow \mathbf{w}^{\mathbf{OLS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

Rule for adding a constant to a Normal random variable
If $X \sim \mathcal{N}(\mu, \tau^2)$, then $X + c \sim \mathcal{N}(\mu + c, \tau^2)$

Noisy observation model (assuming Gaussian noise)
Model: $y = \mathbf{w}^T\mathbf{x} + \epsilon, \;\; \epsilon \sim \mathcal{N}(0, \sigma^2)$

In the above rule for adding a constant to a Normal random variable, if we substitute $\mu := 0$, $\tau^2 := \sigma^2$, and $c := \mathbf{w}^T\mathbf{x}$, we get $y \sim \mathcal{N}(\mathbf{w}^T\mathbf{x}, \sigma^2)$

Data likelihood, assuming training samples are independent and identically distributed (i.i.d):

$$\begin{aligned}
\mathcal{L}(\mathcal{D}) &= \prod_{n=1}^{N} p(y_n|\mathbf{x_n}, \mathbf{w}) \\
&= \prod_{n=1}^{N} \mathcal{N}(\mathbf{w}^T\mathbf{x_n}, \sigma^2) \\
&= \prod_{n=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(y_n - \mathbf{w}^T\mathbf{x_n})^2}{2\sigma^2}\right]
\end{aligned}$$

Log-likelihood:

$$\begin{aligned}
l(\mathcal{D}) &= \log\mathcal{L}(\mathcal{D}) \\
&= \log\left\{\prod_{n=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(y_n - \mathbf{w}^T\mathbf{x_n})^2}{2\sigma^2}\right]\right\} \\
&= \sum_{n=1}^{N} \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \sum_{n=1}^{N} \frac{(y_n - \mathbf{w}^T\mathbf{x_n})^2}{2\sigma^2} \\
&= -N\log\sigma - N\log(2\pi)^{1/2} - \sum_{n=1}^{N} \frac{(y_n - \mathbf{w}^T\mathbf{x_n})^2}{2\sigma^2} \\
&= -\frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{n=1}^{N} (y_n - \mathbf{w}^T\mathbf{x_n})^2 + \text{const} \\
&- \frac{1}{2}\left[N\log\sigma^2 + \frac{1}{\sigma^2}\sum_{n=1}^{N} (y_n - \mathbf{w}^T\mathbf{x_n})^2\right] + \text{const}
\end{aligned}$$

By substituting $RSS(\mathbf{w}) = \sum_{n=1}^{N} (y_n - \mathbf{w}^T \mathbf{x_n})^2$ and $s = \sigma^2$, we get:

$$l(\mathcal{D}) = -\frac{1}{2}\left[N\log s + \frac{1}{s}RSS(\mathbf{w})\right] + \text{const}$$

Maximum likelihood estimation (MLE) is equivalent to minimizing RSS:

$$\max l(\mathcal{D}) \Leftrightarrow \min RSS(\mathbf{w})$$

Estimation of the noise variance:
We maximize the data log-likelihood with respect to $s = \sigma^2$:

$$\begin{aligned}
\frac{\vartheta l(\mathcal{D})}{\vartheta s} &= \frac{\vartheta}{\vartheta s}\left\{-\frac{1}{2}\left[N\log s + \frac{1}{s}RSS(\mathbf{w})\right]\right\} \\
&= -\frac{1}{2}\left[\frac{N}{s} - \frac{1}{s^2}RSS(\mathbf{w})\right] \\
&= -\frac{1}{2s}\left[N - \frac{1}{s}RSS(\mathbf{w})\right]
\end{aligned}$$

$\frac{\vartheta l(\mathcal{D})}{\vartheta s} = 0 \Leftrightarrow s = \frac{1}{N}RSS(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N} (y_n - \mathbf{w}^T \mathbf{x_n})^2$

The MLE of noise variance coincides with the average RSS:

$$\sigma^2_{\text{MLE}} = \frac{1}{N}RSS(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N} (y_n - \mathbf{w}^T \mathbf{x_n})^2$$