Linear Regression Model
for
Time Series Prediction

① Consider $x = \{x_1, x_2, \ldots, x_t, \ldots, x_T\}$ the sequence $x_t \in \mathbb{R}$, $1 \leq t \leq T$, of stock-related price values.

② Our aim is to construct a linear regression model that will provide an estimation / prediction for the future price value $\hat{x}_t$ based on a window of past price values of size $K \in \mathbb{N}_+^*$ such that:

$$\boxed{\hat{x}_t = f(x_{t-1}, x_{t-2}, \ldots, x_{t-K}) \quad (1)}$$

③ Since we assume a linear functional form, we may write that:

$$\hat{x}_t = w_1 \cdot x_{t-1} + w_2 \cdot x_{t-2} + \dots + w_k \cdot x_{t-k} + w_0 \quad (2)$$

Eq. (2) may be equivalently be written as:

$$\hat{x}_t = \sum_{z=1}^{z=k} w_z \cdot x_{t-z} + w_0 \quad (3)$$

④ Eq. (3) suggests that the original sequence $\mathcal{X}$ needs to be partitioned into overlapping windows of size k as follows:

$$\begin{aligned}
\underline{x}_1^T &= [\, x_k \;\; x_{k-1} \;\; \dots \;\; x_2 \;\; x_1 \,] \in \mathbb{R}^k \\
\underline{x}_2^T &= [\, x_{k+1} \;\; x_k \;\; \dots \;\; x_3 \;\; x_2 \,] \in \mathbb{R}^k \\
&\;\;\vdots \qquad\qquad\qquad \vdots \qquad\qquad\qquad \vdots \qquad\qquad (4) \\
\underline{x}_t^T &= [\, x_{t+k-1} \;\; x_{t+k-2} \;\; \dots \;\; x_{t+1} \;\; x_t \,] \in \mathbb{R}^k \\
&\;\;\vdots \\
\underline{x}_{T-k}^T &= [\, x_{T-1} \;\; x_{T-2} \;\; \dots \;\; x_{T-k+1} \;\; x_{T-k} \,] \in \mathbb{R}^k
\end{aligned}$$

⑤ Each vector $\underline{x}_t$ of past price values will be associated with a corresponding target regression variable:

$$\boxed{y_t = x_{t+k}, \quad 1 \le t \le T-k \quad (5)}$$

⑥ In this context, our dataset will be composed by row-wise concatenating vectors $\underline{x}_1, \underline{x}_2, \dots, \underline{x}_{T-k}$ such that:

$$\boxed{\underline{\underline{X}} = \begin{bmatrix} \underline{x}_1^T \\ \hline \underline{x}_2^T \\ \hline \vdots \\ \underline{x}_t^T \\ \hline \vdots \\ \underline{x}_{T-k}^T \end{bmatrix} \in \mathbb{R}^{(T-k) \times k} \quad (6)}$$

7 The corresponding vector of target regression values will be composed as:

$$\boxed{\underline{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_t \\ \vdots \\ y_{T-K} \end{bmatrix} \in \mathbb{R}^{T-K} \quad (7)}$$

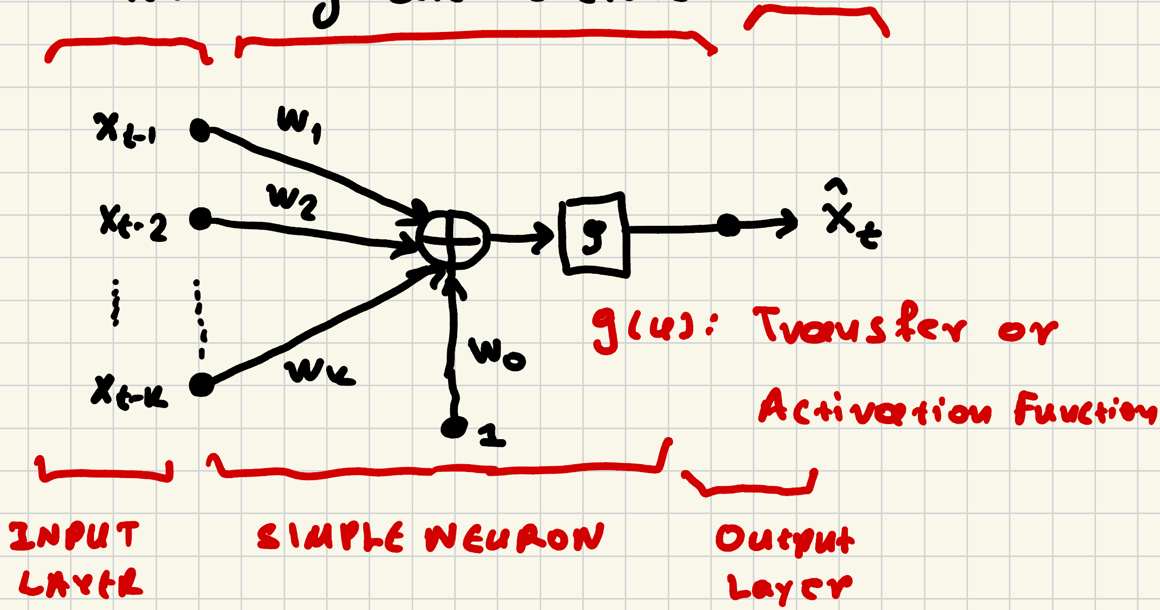
8 Thus, the model output $\hat{y}_t$ provides its prediction for the true target value y_t based on the input pattern $\underline{x}_t$ as:

$$\boxed{\hat{y}_t = f(\underline{x}_t) = \underline{w}^T \cdot \underline{x}_t + w_0 \quad (8)}$$

where

$$\boxed{\underline{w} = \begin{bmatrix} w_K \\ w_{K-1} \\ \vdots \\ w_2 \\ w_1 \end{bmatrix} \in \mathbb{R}^K \quad (9) \qquad w_0 \in \mathbb{R}}$$

(9) Eqs. (8), (9) provide the functionality of a Perceptron Neural Network of the following architecture:



For a Linear Neural Network as the one above, the transfer function is purely linear, i.e. $\boxed{g(u) = u \quad (10)}$

(10) We need to obtain the optimal configuration of weight parameters $\underline{w}^*$ and w_0^* based on the training patterns $\underline{\underline{X}}$ and corresponding target values $\underline{Y}$.

(11) We may define the error for the t-th data point as:

$$\boxed{e_t = y_t - \hat{y}_t \,, \quad 1 \leq t \leq T-K \quad (11)}$$

which may be also written as:

$$\boxed{e_t = y_t - (\underline{w}^T \cdot \underline{x}_t + w_0) \quad (12)}$$

(12) In this context, the overall loss function (objective function) may be defined as:

$$\boxed{J(\underline{w}, w_0) = \sum_{t=1}^{t=T-K} e_t^2 \quad (13)}$$

Letting $N = T-K$ to denote the total number of training patterns, we may write that:

$$\boxed{J(\underline{w}, w_0) = \sum_{t=1}^{N} (y_t - (\underline{w}^T \cdot \underline{x}_t + w_0))^2 \quad (14)}$$

(13) Acquiring the optimal values for $\underline{w}^*$ and w_0^* requires addressing the following optimization problem:

$$\boxed{(\underline{w}^*, w_0^*) = \arg\min_{(\underline{w}, w_0)} J(\underline{w}, w_0) \quad (15)}$$

(14) Given the quadratic form of the loss function, obtaining the optimal values for $\underline{w}^*$ and w_0^* can be achieved by imposing the First-Order Conditions (FOCs):

$$\boxed{\begin{cases} \dfrac{\partial J}{\partial \underline{w}} = \underline{0} \in \mathbb{R}^k \\[2ex] \dfrac{\partial J}{\partial w_0} = 0 \in \mathbb{R} \end{cases} \quad (16)}$$

(15) Gradient with Respect to $\underline{w}$:

$$\boxed{\frac{\partial J}{\partial \underline{w}} = \frac{\partial}{\partial \underline{w}} \left\{ \sum_{t=1}^{N} e_t^2 \right\} = \sum_{t=1}^{N} \frac{\partial}{\partial \underline{w}} e_t^2 \quad (17)}$$

Utilizing the chain rule, we may get that:

$$\boxed{\frac{\partial}{\partial \underline{w}} e_t^2 = 2 e_t \cdot \frac{\partial e_t}{\partial \underline{w}} \quad (18)}$$

Accordingly, we may write that:

$$\frac{\partial e_t}{\partial \underline{w}} = \frac{\partial}{\partial \underline{w}} \left\{ y_t - (\underline{w}^T \cdot \underline{x}_t + w_0) \right\} \Rightarrow$$

$$\frac{\partial e_t}{\partial \underline{w}} = - \frac{\partial}{\partial \underline{w}} \left\{ \underline{w}^T \cdot \underline{x}_t + w_0 \right\} \Rightarrow$$

$$\boxed{\frac{\partial e_t}{\partial \underline{w}} = - \underline{x}_t \quad (19)}$$

Substituting Eq. (19) into Eq. (18), yields:

$$\boxed{\frac{\partial e_t^2}{\partial \underline{w}} = -2\, e_t \cdot \underline{x}_t \quad (20)}$$

Thus, Eq. (17) may be written as:

$$\boxed{\frac{\partial J}{\partial \underline{w}} = -2 \sum_{t=1}^{N} \left(y_t - \left(\underline{w}^T \cdot \underline{x}_t + w_0 \right) \right) \cdot \underline{x}_t \quad (21)}$$

Enforcing the respective First order Condition:

$$\frac{\partial J}{\partial \underline{w}} = \underline{0} \Rightarrow$$

$$\boxed{\sum_{t=1}^{N} \left(y_t - \left(\underline{w}^T \cdot \underline{x}_t + w_0 \right) \right) \underline{x}_t = \underline{0} \quad (22)}$$

or

$$\boxed{\sum_{t=1}^{N} e_t \cdot \underline{x}_t = \underline{0} \quad (22')}$$

(16) Gradient with Respect to w_0:

$$\boxed{\frac{\partial J}{\partial w_0} = \frac{\partial}{\partial w_0}\left\{\sum_{t=1}^{N} e_t^2\right\} = \sum_{t=1}^{N} \frac{\partial}{\partial w_0} e_t^2 \quad (23)}$$

Utilizing the chain rule again:

$$\boxed{\frac{\partial e_t^2}{\partial w_0} = 2\, e_t \frac{\partial e_t}{\partial w_0} \quad (24)}$$

Accordingly, we may write that:

$$\frac{\partial e_t}{\partial w_0} = \frac{\partial}{\partial w_0}\left\{ y_t - (\underline{w}^T \cdot \underline{x}_t + w_0) \right\} \Rightarrow$$

$$\frac{\partial e_t}{\partial w_0} = -\frac{\partial}{\partial w_0}\left\{ \underline{w}^T \cdot \underline{x}_t + w_0 \right\} \Rightarrow$$

$$\boxed{\frac{\partial e_t}{\partial w_0} = -1 \quad (25)}$$

Substituting Eq. (25) into Eq. (24) yields:

$$\boxed{\frac{\partial e_t^2}{\partial w_0} = -2 \cdot e_t \quad (26)}$$

Thus, Eq. (23) may be written as:

$$\boxed{\frac{\partial J}{\partial w_0} = -2 \sum_{t=1}^{N} (y_t - (\underline{w}^T \cdot \underline{x}_t + w_0)) \quad (27)}$$

Enforcing the respective First Order Condition:

$$\frac{\partial J}{\partial w_0} = 0 \Rightarrow$$

$$\boxed{\sum_{t=1}^{N} (y_t - (\underline{w}^T \cdot \underline{x}_t + w_0)) = 0 \quad (28)}$$

<u>or</u>

$$\boxed{\sum_{t=1}^{N} e_t = 0 \quad (28)'}$$

(17) **Matrix Formulation:**

Having defined the matrix of input patterns $\underline{X} \in \mathbb{R}^{N \times K}$ where row t is $\underline{x}_t^T$ and the vector of target regression values $\underline{y} \in \mathbb{R}^N$, we may also define the augmented data matrix $\hat{\underline{\underline{X}}}$ to include a column of ones to account for w_0, as:

$$\boxed{\hat{\underline{\underline{X}}} = \begin{bmatrix} \underline{x}_1^T & 1 \\ \vdots & \vdots \\ \underline{x}_t^T & 1 \\ \vdots & \vdots \\ \underline{x}_N^T & 1 \end{bmatrix} \quad (29)} \qquad \hat{\underline{\underline{X}}} \in \mathbb{R}^{N \times (K+1)}$$

Likewise, we may define the augmented weight vector $\boxed{\hat{\underline{w}} = \begin{bmatrix} \underline{w} \\ w_0 \end{bmatrix} \in \mathbb{R}^{K+1} \quad (30)}$

In this context, the vector $\underline{\hat{y}} \in \mathbb{R}^N$ of network predictions may be defined as:

$$\underline{\hat{y}} = \underline{\underline{\hat{X}}} \cdot \underline{\hat{w}} \qquad (31)$$

$[N \times 1] \qquad [N \times (K+1)] \cdot [(K+1) \times 1]$

Therefore, the set of FOCs can be written in matrix form as:

$$\underline{\underline{\hat{X}}}^T \cdot (\underline{y} - \underline{\hat{y}}) = \underline{0} \qquad (32)$$

$[(K+1) \times N] \, [N \times 1] \qquad [(K+1) \times 1]$

Taking into consideration the fact that the error vector $\underline{e} = \underline{y} - \underline{\hat{y}}$ and the form of the augmented data matrix $\underline{\underline{\hat{X}}}$, we may write that:

$$\underline{\underline{\hat{X}}}^T \cdot \underline{e} = \underline{0} \in \mathbb{R}^{K+1} \qquad (33)$$

Moreover,

$$\boxed{\underline{\underline{\hat{X}}}^T = \begin{bmatrix} \underline{x}_1 & \underline{x}_2 & \cdots & \underline{x}_N \\ 1 & 1 & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{(K+1)\times N} \quad (34)}$$

Eqs. (33) and (34) may be combined as:

$$\boxed{\begin{bmatrix} \underline{x}_1 & \underline{x}_2 & \cdots & \underline{x}_N \\ 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix} = \underline{0}_{[(K+1)\times 1]} \quad (35)}$$

Expanding Eq. (35) results in:

$$\begin{cases} \displaystyle\sum_{t=1}^{N} e_t \underline{x}_t = \underline{0}_{[K\times 1]} \\ \displaystyle\sum_{t=1}^{N} e_t = 0 \in \mathbb{R} \end{cases}$$

which are the FOCs that we have already derived.

Thus, by substituting Eq.(31) into Eq.(32), we get that:

$$\hat{\underline{\underline{X}}}^T \cdot \left(\underline{y} - \hat{\underline{\underline{X}}} \cdot \hat{\underline{w}} \right) = \underline{0}_{[K+1 \times 1]} \Rightarrow$$

$$\underbrace{\hat{\underline{\underline{X}}}^T \cdot \underline{y}}_{\substack{[(K+1)\times N]\,[N\times 1] \\ [(K+1)\times 1]}} - \underbrace{\hat{\underline{\underline{X}}}^T \cdot \hat{\underline{\underline{X}}} \cdot \hat{\underline{w}}}_{\substack{[(K+1)\times N]\cdot[N\times(K+1)]\cdot[(K+1)\times 1] \\ [(K+1)\times(K+1)]\cdot[(K+1)\times 1] \\ [(K+1)\times 1]}} = \underbrace{0}_{[(K+1)\times 1]} \Rightarrow$$

$$\hat{\underline{\underline{X}}}^T \cdot \hat{\underline{\underline{X}}} \cdot \hat{\underline{w}} = \hat{\underline{\underline{X}}}^T \cdot \underline{y} \Rightarrow$$

$$\boxed{\hat{\underline{\underline{w}}}^* = \left(\hat{\underline{\underline{X}}}^T \cdot \hat{\underline{\underline{X}}} \right)^{-1} \cdot \hat{\underline{\underline{X}}}^T \cdot \underline{y}}$$

OPTIMAL WEIGHT VECTOR