Machine Learning & Signals Learning

2 Uni-Variate Linear Least-Squares

The goal of this chapter is to define and discuss uni-variate linear least-squares (LLS or LS) for a linear model.

Note that LS is also termed *linear regression*. Uni-variate LS is also referred to as a *linear trend-line*.

2.1 Uni-variate Linear LS

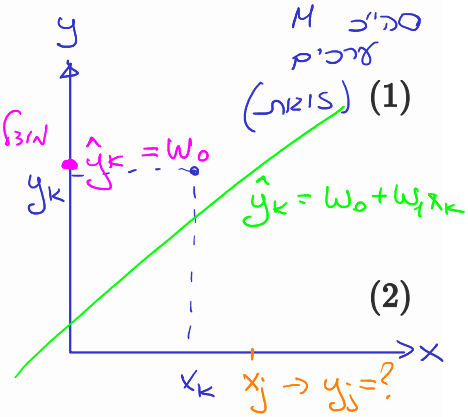
2.1.1 Definitions

Dataset: A random experiment produces a *dataset* of M paired observations $\{x_k, y_k\}_{k=1}^{M}$.

ML Model: The assumed *model* underlying the dataset is

$$y = f(x) + \epsilon, \tag{1}$$

where $f(x)$ is a deterministic function and ϵ is zero-mean noise.



The goal is to find and apply the model,

$$\hat{y}_k = f(x_k) \tag{2}$$

that provide the most appropriate results with relation to y_k.

Error: The model *error* is

$$e_k = y_k - \hat{y}_k \tag{3}$$

Loss function: Loss (or cost) function $\mathcal{L}(y_k, \hat{y}_k)$ is some distance metric between y_k and $\hat{y}_k$. The optimal model parameters are found by minimizing a loss function. Common examples of loss functions include:

- the sum of squared errors (SSE),

$$\mathrm{SSE} = \sum_{k=1}^{M} e_k^2, \tag{4}$$

- mean-square error (MSE) or (biased) error variance

$$\mathrm{MSE} = \frac{1}{M}\sum_{k=1}^{M} e_k^2 = \frac{1}{M}\mathrm{SSE} = s_e^2 \tag{5}$$

- and root-mean-square error (RMSE),

$$\text{RMSE} = \sqrt{\text{MSE}}. \tag{6}$$

Because constant multiplication factors and monotonic transforms do not affect the location of the minimum, all these loss functions share the same optimal parameters:

$$\begin{aligned} w_0, w_1 &= \arg\min_{w_0,w_1} \text{SSE}(w_0, w_1) \\ &= \arg\min_{w_0,w_1} \text{MSE}(w_0, w_1) \\ &= \arg\min_{w_0,w_1} \text{RMSE}(w_0, w_1). \end{aligned} \tag{7}$$

Metric: A quantifiable measure used to evaluate how well a model is performing.

It is easy to confuse metrics with loss functions, but they serve different purposes:

- *Loss* function is used by the **algorithm** during training to learn optimal model parameters.
- *Metric* is used by **humans** to evaluate the model's performance.

2.1.2 Mean and Variance

The goal of this section is to provide an *interpretation* of the mean and variance within the context of LS.

A special case of the linear model is one of the form:

$$\hat{y} = w_0 \tag{8}$$

The corresponding MSE loss (2.5) function is

$$\mathcal{L}(w_0) = \frac{1}{M}\sum_{k=1}^{M}(y_k - w_0)^2, \tag{9}$$

To find the minimum,

$$\frac{d}{dw_0}\mathcal{L}(w_0) = 2\frac{1}{M}\sum_{k=1}^{M}(y_k - w_0)(-1) = 0 \tag{10}$$

with the related minimum at the mean of y_k values,

$$w_0 = \frac{1}{M}\sum_{k=1}^{M} y_k = \bar{y} \tag{11}$$

The corresponding MSE of the model (substituting (2.11) into (2.9)) is

$$\text{MSE} = \frac{1}{M}\sum_{k=1}^{M}(y_k - \bar{y})^2 = \sigma_y^2 \tag{12}$$

which is the *sample variance* of the y_k values [1].

To summarize, $\hat{y} = \bar{y}$ with $\mathbf{MSE} = \mathsf{S}_y^2$.

[1] This variance expression is called *biased* and is used throughout this chapter (see Sec. 1.1)

2.1.3 Linear Model

The example of *linear model* is

$$\hat{y}_k = f(x_k; w_0, w_1) = w_0 + w_1 x_k, \tag{13}$$

where w_0 and w_1 are the model parameters.

The corresponding linear model error is

$$e_k = y_k - (w_0 - w_1 x_k). \tag{14}$$

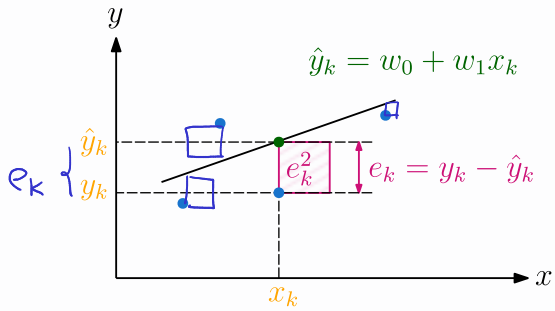


Figure 2.1: Linear regression visualization. The goal is to minimize SSE, which represents the total area $\sum_k e_k^2$ of the rectangles.

The minimization of the SSE loss function ((2.4) and Fig. 2.1),

$$\mathcal{L}(w_0, w_1) = \sum_{k=1}^{M} (y_k - w_0 - w_1 x_k)^2, \tag{15}$$

is performed by setting the partial derivatives of $\mathcal{L}$ to zero:

$$\begin{cases} \dfrac{\partial}{\partial w_0} \mathcal{L}(w_0, w_1) = 0 \\ \dfrac{\partial}{\partial w_1} \mathcal{L}(w_0, w_1) = 0 \end{cases} \tag{16}$$

which yields:

$$\begin{cases} 2\sum_{k=1}^{M}(y_k - w_0 - w_1 x_k)\cdot(-1) = 0 \\ 2\sum_{k=1}^{M}(y_k - w_0 - w_1 x_k)\cdot(-x_k) = 0. \end{cases} \tag{17}$$

Finally, with some basic algebra:

$$\begin{cases} w_0 M + w_1 \sum_{k=1}^{M} x_k = \sum_{k=1}^{M} y_k, \\ w_0 \sum_{k=1}^{M} x_k + w_1 \sum_{k=1}^{M} x_k^2 = \sum_{k=1}^{M} x_k y_k. \end{cases} \tag{18}$$

This system of equations is termed the **normal equations**.

2.2 Normal Equations with Statistical Terms

> **Goal:** To rewrite the normal equations (2.18) using statistical terms.

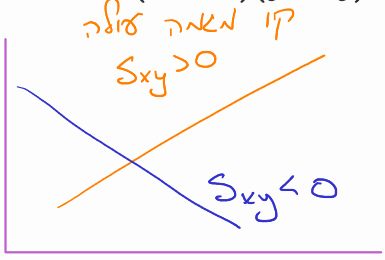
2.2.1 Sample Covariance

The (biased) sample covariance between $x_1, \ldots, x_M$ and $y_1, \ldots, y_M$ is given by

$$s_{xy} = \frac{1}{M}\sum_{k=1}^{M}(x_k - \bar{x})(y_k - \bar{y}) \tag{19}$$

The sign and magnitude of the sample covariance s_{xy} indicate the direction and strength of the linear relationship between x and y:

- $s_{xy} > 0$: As x increases, y tends to increase (most products $(x_i - \bar{x})(y_i - \bar{y})$ are positive).
- $s_{xy} < 0$: As x increases, y tends to decrease.
- $s_{xy} \approx 0$: No linear association.



2.2.2 Normal Equations

It is numerically convenient to rewrite the normal equations (2.18) in terms of sample moments

$$\begin{aligned} w_1 &= \frac{s_{xy}}{s_x^2} = \frac{\sum_k (x_k - \bar{x})(y_k - \bar{y})}{\sum_k (x_k - \bar{x})^2}, \\ w_0 &= \bar{y} - w_1 \bar{x} = \bar{y} - \frac{s_{xy}}{s_x^2}\bar{x}, \end{aligned} \tag{20}$$

and the predictive form

$$\hat{y} = \bar{y} + w_1(x - \bar{x}). \tag{21}$$

Note that these expressions are independent of whether biased or unbiased definitions are used for s_{xy} and s_x, since the corresponding normalization factors cancel out.

Remarks.

- A valid solution requires $s_x \neq 0$, i.e. variability in x_i.
- If both variables are centered ($\bar{x} = \bar{y} = 0$), then (2.20) reduces to

$$w_0 = 0,$$

$$w_1 = \frac{\sum_k x_k y_k}{\sum_k x_k^2} \tag{22}$$

2.2.3 Correlation Coefficient

The *sample Pearson correlation coefficient* between $\mathbf{x}$ and $\mathbf{y}$ is the normalized (dimensionless) covariance:

$$r_{xy} = \frac{s_{xy}}{s_x\, s_y} \tag{23}$$

$$= \frac{\sum_{k=1}^{M} (x_k - \bar{x})\,(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{M} (x_k - \bar{x})^2}\,\sqrt{\sum_{k=1}^{M} (y_k - \bar{y})^2}} \tag{24}$$

and the slope in (2.20) can be re-written as

$$w_1 = r_{xy}\,\frac{s_y}{s_x}. \tag{25}$$

Note that there is *no difference between biased and unbiased* definitions for r_{xy}, since the corresponding coefficients cancel out.

2.3 Metrics

> The performance of a model is quantified by a performance metric, which is not necessarily the same as the loss function.

2.3.1 Coefficient of Determination (R^2)

To emphasize the difference between loss and metrics, the following example of an LR metric is provided. The coefficient of determination, denoted R^2 (R-squared), is based on

the relation:

$$\underbrace{\sum_{k=1}^{M}(y_k-\bar{y})^2}_{\text{SST}}=\underbrace{\sum_{k=1}^{M}(\hat{y}_k-\bar{y})^2}_{\text{SSR}}+\underbrace{\sum_{k=1}^{M}e_k^2}_{\text{SSE}}, \tag{26}$$

where:

- **SST**: Total sum of squares.
- **SSR**: Sum of squares due to regression.
- **SSE**: Sum of squared errors (or residual sum of squares).

Proof. Observe that

$$y_k-\bar{y} \;=\; (\hat{y}_k-\bar{y})+(y_k-\hat{y}_k) \;=\; (\hat{y}_k-\bar{y})+e_k.$$

Hence,

$$\begin{aligned}\sum_{k=1}^{M}(y_k-\bar{y})^2 &= \sum_{k=1}^{M}\big[(\hat{y}_k-\bar{y})+e_k\big]^2 \\ &= \sum_{k=1}^{M}(\hat{y}_k-\bar{y})^2 \;+\; \sum_{k=1}^{M}e_k^2 \;+\; 2\sum_{k=1}^{M}(\hat{y}_k-\bar{y})\,e_k.\end{aligned}$$

It remains to show the cross-term vanishes:

$$\sum_{k=1}^{M}(\hat{y}_k-\bar{y})\,e_k=\sum_{k=1}^{M}\hat{y}_k\,e_k \;-\; \bar{y}\sum_{k=1}^{M}e_k=0-0=0,$$

since in LS, $\sum_{k=1}^{M}e_k=0$ and $\sum_{k=1}^{M}\hat{y}_k\,e_k=0$.

The R^2 metric is defined as:

$$R^2=\frac{\text{SSR}}{\text{SST}}=1-\frac{\text{SSE}}{\text{SST}}, \tag{27}$$

providing a unitless goodness-of-fit measure that shares an intuitive $[0,1]$ range:

- $R^2=1$: Perfect fit.
- $R^2=0$: The model is no better than predicting the mean, $\hat{y}_i=\bar{y}$.
- $R^2<0$: The model performs worse than the mean (possible for machine learning models on test data).

R^2 of Uni-variate Linear LS For the uni-variate LR case, it is the fraction of the sample variance of y explained by the linear fit,

$$R^2=r_{xy}^2. \tag{28}$$

Proof. Starting with $\hat{y}_k = w_0 + w_1 x_k$:

$$\begin{aligned} SSR &= \sum_{k=1}^{M} (\hat{y}_k - \bar{y})^2 \\ &= \sum_{k=1}^{M} (w_0 + w_1 x_k - \bar{y})^2 \\ &= \sum_{k=1}^{M} (\bar{y} - w_1 \bar{x} + w_1 x_k - \bar{y})^2 \\ &= w_1^2 \sum_{k=1}^{M} (x_k - \bar{x})^2 \\ &= w_1^2 M s_x^2 = M \left(\frac{s_{xy}}{s_x^2} \right)^2 s_x^2 = M \frac{s_{xy}^2}{s_x^2}, \\ SST &= \sum_{k=1}^{M} (y_k - \bar{y})^2 = M s_y^2 \end{aligned} \tag{29}$$

Thus, $SSR/SST = s_{xy}^2 / (s_x^2 s_y^2) = r_{xy}^2$.

2.3.2 MSE of Linear LS

The general (biased) MSE of unnormalized values is

$$\begin{aligned} \mathrm{MSE} &= \frac{1}{M} \sum_{k=1}^{M} (y_k - \hat{y}_k)^2 \\ &= s_y^2 (1 - r_{xy}^2) \\ &= \frac{1}{M} \mathrm{SSE} = \frac{1}{M} (\mathrm{SST} - \mathrm{SSR}) = s_e^2 \end{aligned} \tag{30}$$

Note, the unbiased estimator of s_e^2 requires $1/(M-2)$ factor instead $1/M$.

2.4 Normalized (z-score) Linear Regression (*)

The goal of this section is to provide an intuitive interpretation of the correlation coefficient.

Rescaling To make the slope directly interpretable, we rescale both variables to zero mean and unit variance:

$$z_i = \frac{x_i - \bar{x}}{s_x}, \qquad t_i = \frac{y_i - \bar{y}}{s_y}, \qquad i = 1, \ldots, M \tag{31}$$

The properties of z_i and t_i are

$$\bar{z} = \frac{1}{M}\sum z_i = 0,$$
$$\bar{t} = \frac{1}{M}\sum t_i = 0,$$
$$s_z = s_t = 1.$$

Proof. Using biased variance definitions [2]:

$$s_x^2 = \frac{1}{M}\sum_{k=1}^{M}(x_k - \bar{x})^2, \qquad s_y^2 = \frac{1}{M}\sum_{k=1}^{M}(y_k - \bar{y})^2. \tag{32}$$

For the means

$$\begin{aligned}\sum_{k=1}^{M} z_k &= \frac{1}{s_x}\sum_{k=1}^{M}(x_k - \bar{x}) \\ &= \frac{1}{s_x}\left(\sum_{k=1}^{M} x_k - M\bar{x}\right) \\ &= \frac{1}{s_x}(M\bar{x} - M\bar{x}) = 0\end{aligned} \tag{33}$$

hence $\bar{z} = 0$ and similarly $\bar{t} = 0$. For the variances:

$$s_z^2 = \frac{1}{M}\sum_{k=1}^{M} z_k^2 = \frac{1}{M}\sum_{k=1}^{M}\frac{(x_k - \bar{x})^2}{s_x^2} = \frac{1}{s_x^2}\left[\frac{1}{M}\sum_{k=1}^{M}(x_k - \bar{x})^2\right] = \frac{s_x^2}{s_x^2} = 1. \tag{34}$$

The proof for $s_t^2 = 1$ is analogous.

[2] For unbiased evaluation the factor $M - 1$ is to be used.

Normal Equations The linear model in normalized space is

$$\hat{t}_i = r_{xy}\, z_i. \tag{35}$$

where

$$r_{xy} = \frac{s_{xy}}{s_x\, s_y} \tag{36}$$

is the correlation coefficient.

Proof. Applying (2.22) to

$$\hat{t}_i = w_0^* + w_1^* z_i \tag{37}$$

yields

$$w_0^* = 0, \tag{38}$$

$$w_1^* = \frac{s_{zt}}{s_z^2} = \frac{1}{M}\sum_{i=1}^{M} z_i t_i = \frac{1}{M}\sum_{i=1}^{M} z_i t_i = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = r_{xy} \tag{39}$$

Regression error as a function of r_{xy}

With the optimal normalized model $\hat{t} = r_{xy} z$ the residuals are

$$e_i = t_i - \hat{t}_i = t_i - r_{xy} z_i \tag{40}$$

with

$$\text{MSE} = 1 - r_{xy}^2 = s_e^2. \tag{41}$$

Proof.

$$\begin{aligned} s_e^2 &= \frac{1}{M}\sum_{i=1}^{M} e_i^2 = \frac{1}{M}\sum_{i=1}^{M} (t_i - r_{xy} z_i)^2 \\ &= \frac{1}{M}\sum_{i=1}^{M} t_i^2 - 2 r_{xy} \frac{1}{M}\sum_{i=1}^{M} t_i z_i + r_{xy}^2 \frac{1}{M}\sum_{i=1}^{M} z_i^2 \\ &= s_t^2 - 2 r_{xy}\,(r_{xy}) + r_{xy}^2 s_z^2 \\ &= 1 - r_{xy}^2 \end{aligned} \tag{42}$$

Geometrical interpretation of r_{xy}

• Units of x- and y-axes are standard deviation from the origin since t_i and z_i are centered and standardized.

• In the normalized space, r_{xy} *is* the regression slope. A one-standard-deviation increase in x (z changes by 1) changes y by r_{xy} standard deviations on average ($\hat{t}$ changes by r_{xy}).

• Let's define two vectors, $\mathbf{z}$ and $\mathbf{t}$, both in $\mathbb{R}^M$. The angle between them is θ and $r_{xy} = \cos\theta$ (Fig. 2.2).

Proof. The corresponding dot product is

$$\mathbf{z} \cdot \mathbf{t} = \|\mathbf{z}\|\,\|\mathbf{t}\|\cos(\theta) \tag{43}$$

Using (2.39),

$$\mathbf{z} \cdot \mathbf{t} = \sum_{i=1}^{M} z_i t_i = M r_{xy}. \tag{44}$$

Since $\|\mathbf{z}\| = \sqrt{\sum z_i^2} = \sqrt{M s_z^2} = \sqrt{M}$, we have

$$\|\mathbf{z}\| \, \|\mathbf{t}\| = \sqrt{M} \cdot \sqrt{M} = M \tag{45}$$

Consequences:

- Range: $-1 \leq r_{xy} \leq +1$.
- Special cases:
  * $r_{xy} = +1$ indicates perfect positive linear association, $\theta = 0°$.
  * $r_{xy} = -1$ indicates perfect negative linear association, $\theta = 180°$.
  * $r_{xy} = 0$ indicates no linear association, $\theta = 90°$.

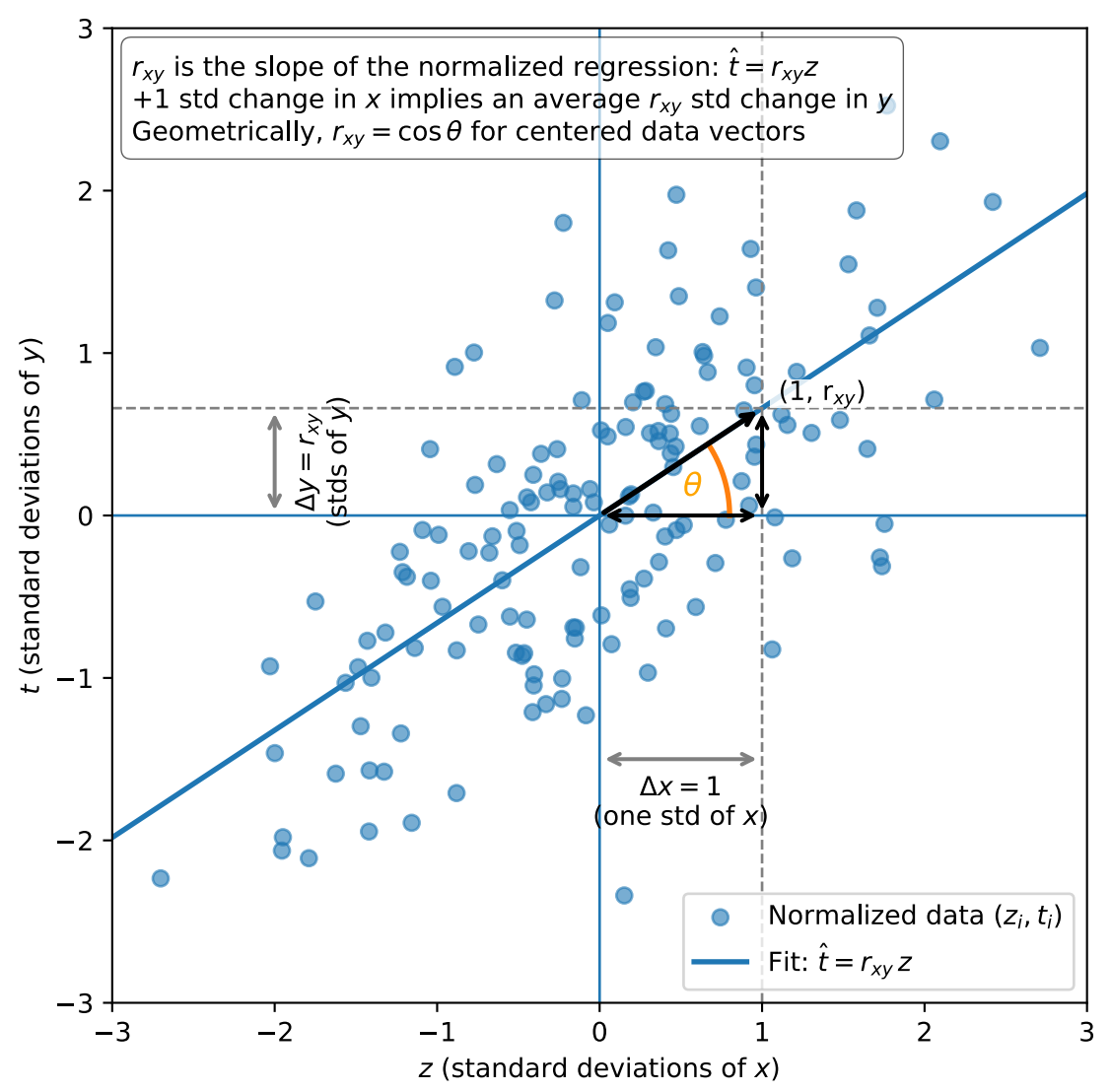


Figure 2.2: Geometric interpretation of r_{xy} in a normalized LR.