Linear regression

- We use a linear model to approximate an unknown function which depends on several variables, to predict a continuous target variable.

- Objective: to find the optimal parameters of such a model.

$$\hat{y}_i = \boxed{\beta_0} + \sum_{j=1}^{p} \beta_j x_{ij}$$

p → total number of features in the model.

$\hat{y}_i$ → predicted value of the target variable for the i^{th} data point

β_0 → intercept

β_j → weight of the j^{th} feature in the model

x_{ij} → value of the j^{th} feature in the i^{th} data point

$$\hat{y}_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

$$\underbrace{\begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \vdots \\ \hat{y}_n \end{pmatrix}}_{\hat{Y}} = \underbrace{\begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ 1 & & \ddots & & \\ \vdots & & & \ddots & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}}_{\beta}$$

$$\hat{Y} = X \qquad \beta$$

$\longrightarrow \boxed{\hat{Y} = X\beta} \longrightarrow$ in matrix notation.

Mean-squared error (MSE), also called the residual sum of squares:

$$MSE(\beta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

true value of the target variable

predicted value of the target

$$\boxed{= \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \left\{ \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right\} \right)^2}$$

To find the optimal set of parameters, $\frac{\partial MSE(\beta)}{\partial \beta_j} = 0.$

for any extremum

To ensure a minimum: $\frac{\partial^2 MSE}{\partial \beta_j} > 0.$

$$SSE = n \times MSE$$

$$SSE(\beta) = (Y - X\beta)^T (Y - X\beta)$$

$$= (Y^T - (X\beta)^T)(Y - X\beta)$$

$$= (Y^T - \beta^T X^T)(Y - X\beta)$$

$$\boxed{SSE(\beta) = Y^T Y - Y^T X\beta - \beta^T X^T Y + \beta^T X^T X\beta}$$

$$\frac{\partial (SSE)}{\partial \beta} = 0 \Rightarrow \boxed{\frac{\partial}{\partial \beta}(SSE) = -2X^T Y + 2X^T X\beta}$$

$$\frac{\partial^2 (SSE)}{\partial \beta^2} = 2X^T X \longrightarrow \text{positive semi-definite matrix}$$

$\downarrow$

ensures a minimum in most cases

For the optimal β, denoted as $\hat{\beta}$

$$-2X^TY + 2X^TX\hat{\beta} = 0$$

$$2X^TX\hat{\beta} = 2X^TY$$

$$\boxed{\hat{\beta} = (X^TX)^{-1}X^TY}$$

expression for the optimal set of parameters "best-fit linear model".

If X or Y have complex entries,

$$\hat{\beta} = (X^HX)^{-1}X^HY$$

where $X^H = \overline{X}^T$ → complex conjugate

X^H → Hermitian transpose: complex conjugate transpose.

The matrix $\boxed{X^{+} = (X^{H}X)^{-1}X^{H}}$ is referred to as a pseudo-inverse or a Moore-Penrose inverse of the matrix X and is the generalization of the concept of an inverse to a rectangular matrix.

$$Ax = b \longrightarrow x = A \backslash b$$

$$x = (A^{H}A)^{-1}A^{H}b$$

Coefficient of determination (R^2):

It is a measure of the goodness of fit of a model.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

SS_{res} → residual sum of squares

SS_{tot} → total sum of squares

$$SS_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \equiv SSE$$

y_i: true value

$\hat{y}_i$: predicted value

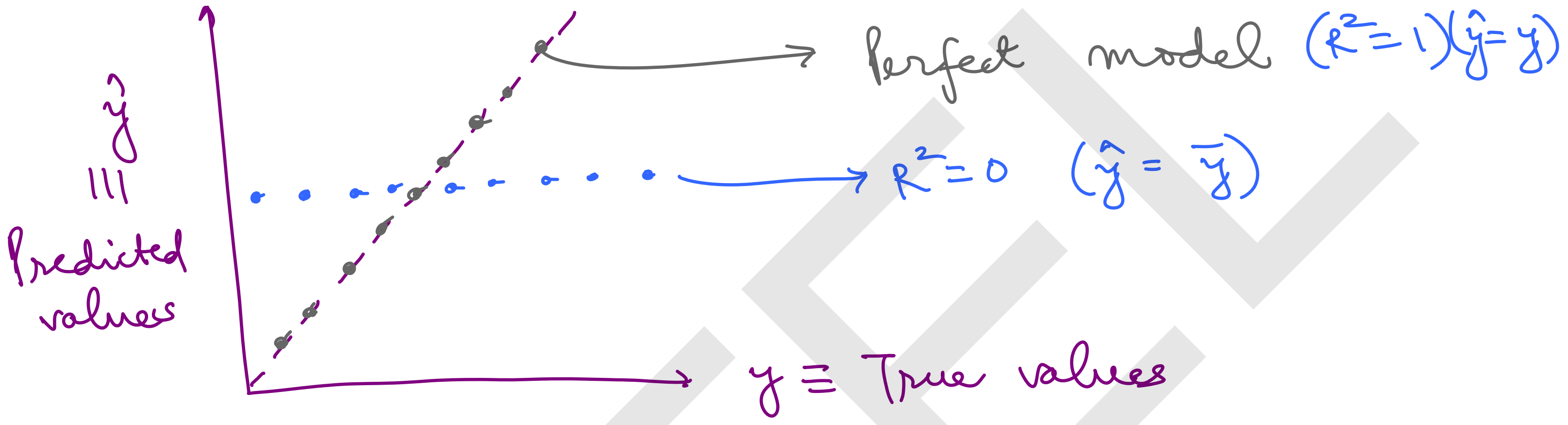
$$SS_{tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$\bar{y}$: mean of the true values

$$\boxed{R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

* For a perfect model: $\hat{y}_i = y_i \Rightarrow R^2 = 1 - 0 = \boxed{1}$

Parity plot



* For imperfect models: $R^2 < 1$

<u>eg:</u> $R^2 = 0$

When $\hat{y}_i = \bar{y}$, then $R^2 = 1 - 1 = \boxed{0}$

$\longrightarrow$ the model always predicts the mean value.

Note that R^2 can be less than zero for very inferior model.