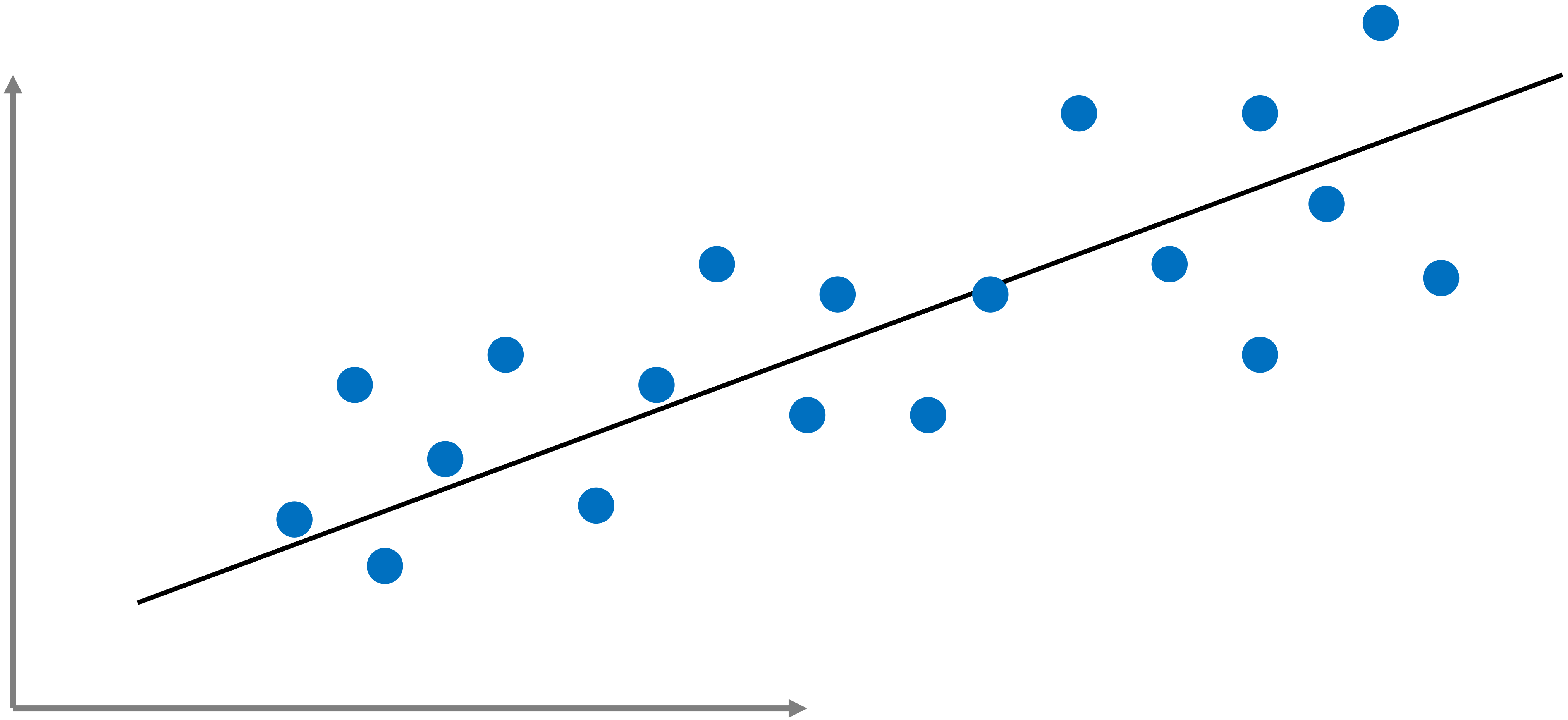
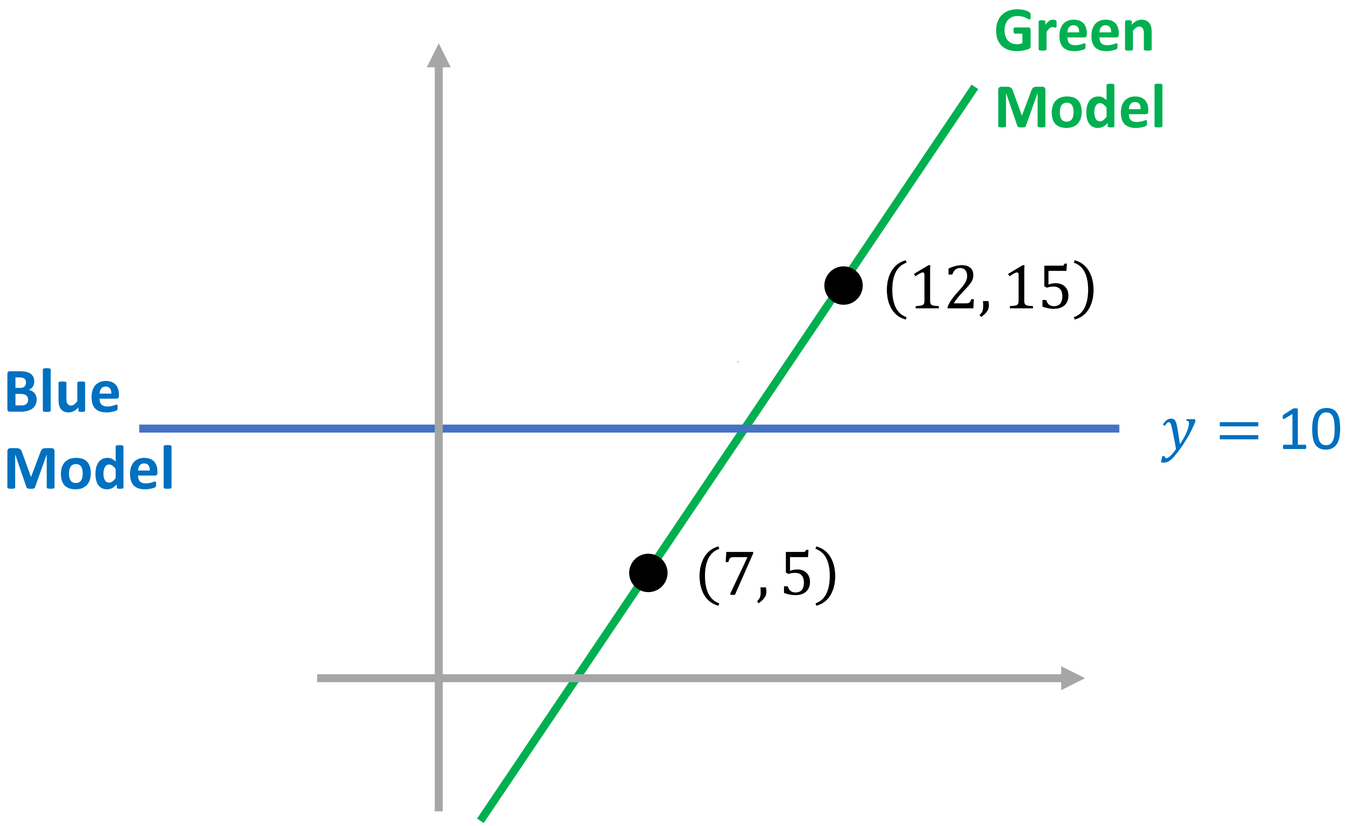
Linear models

Linear regression



Error definition

y_i = Instance
$\hat{y}_i$ = Model

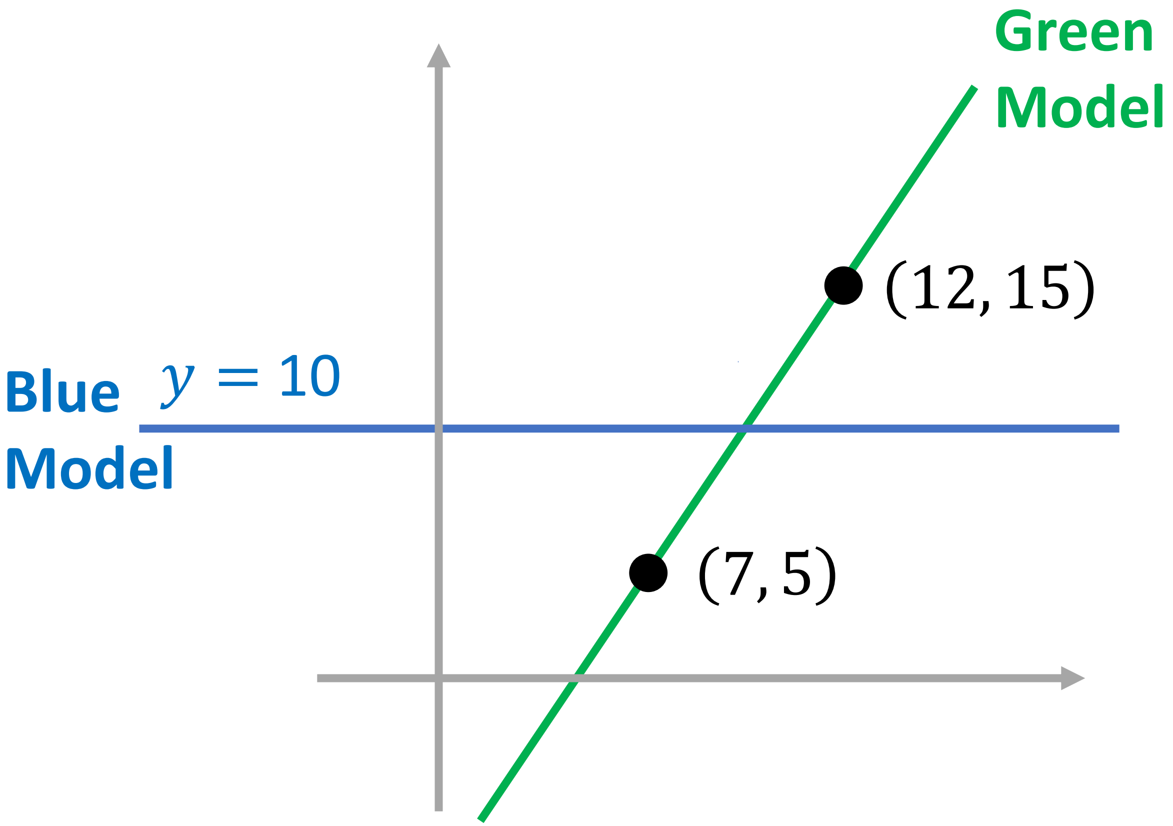


$\varepsilon \;\propto\; \hat{y}_i - y_i$

$\varepsilon = 0$

$\varepsilon = 0$

Error definition



$$\varepsilon \propto \hat{y}_i - y_i$$

$$\varepsilon = 0$$

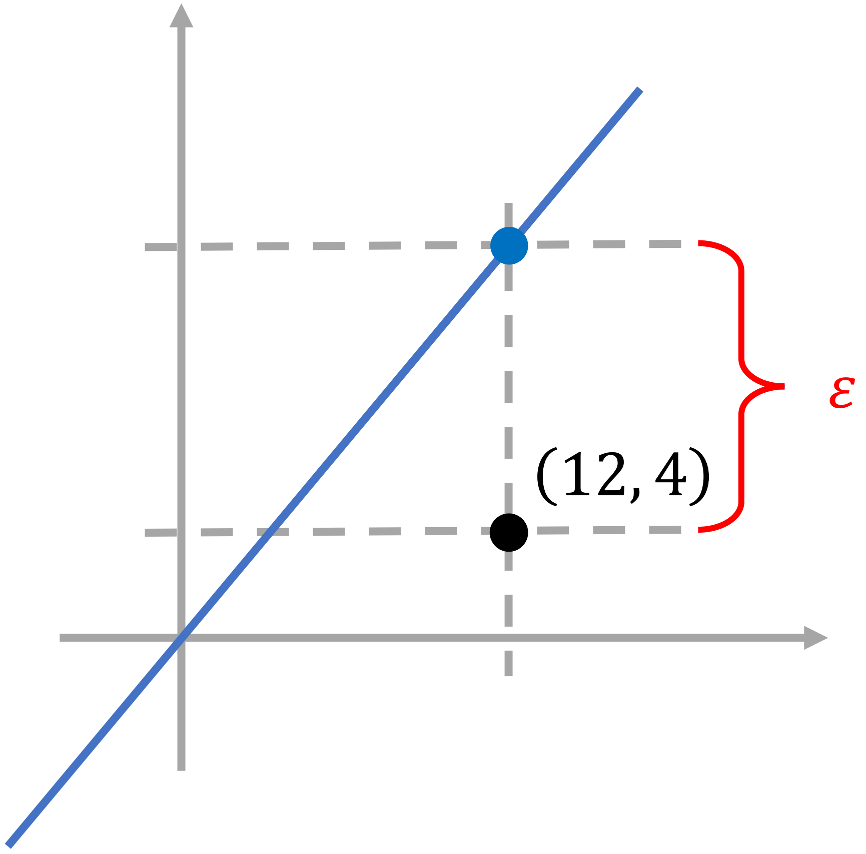
$$\varepsilon = 0$$

Solution:

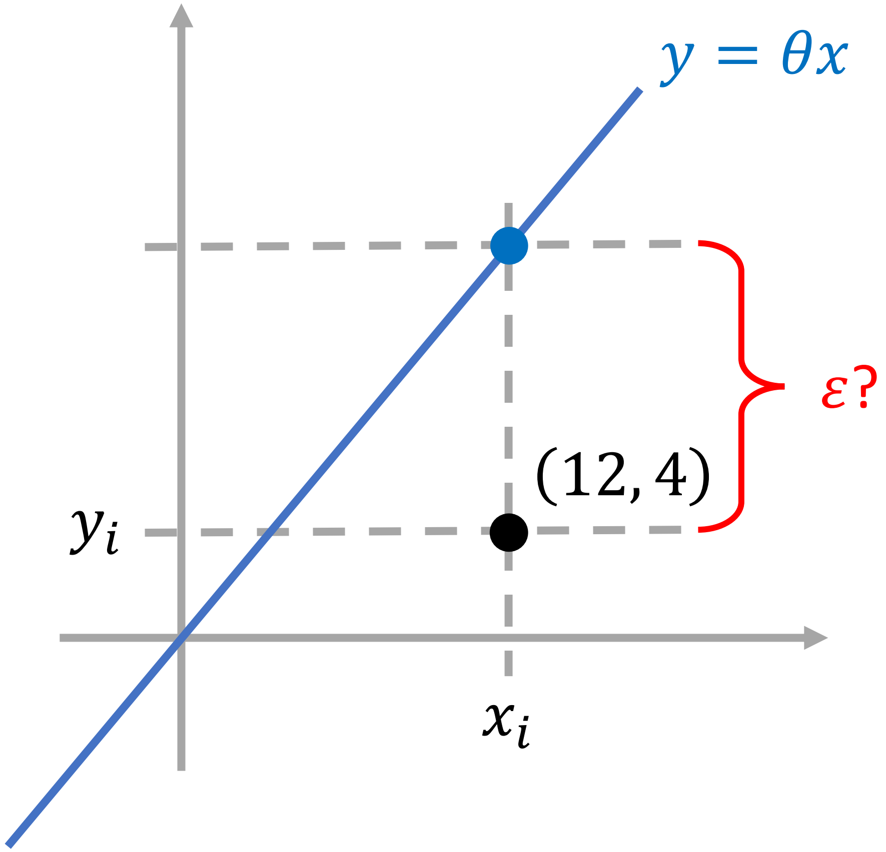
Residual sum of squares (RSS) $$\varepsilon := RSS = \sum_{i=1}^{m} (\hat{y}_i - y_i)^2$$

Page 62, Gareth & Hastie, 2014.

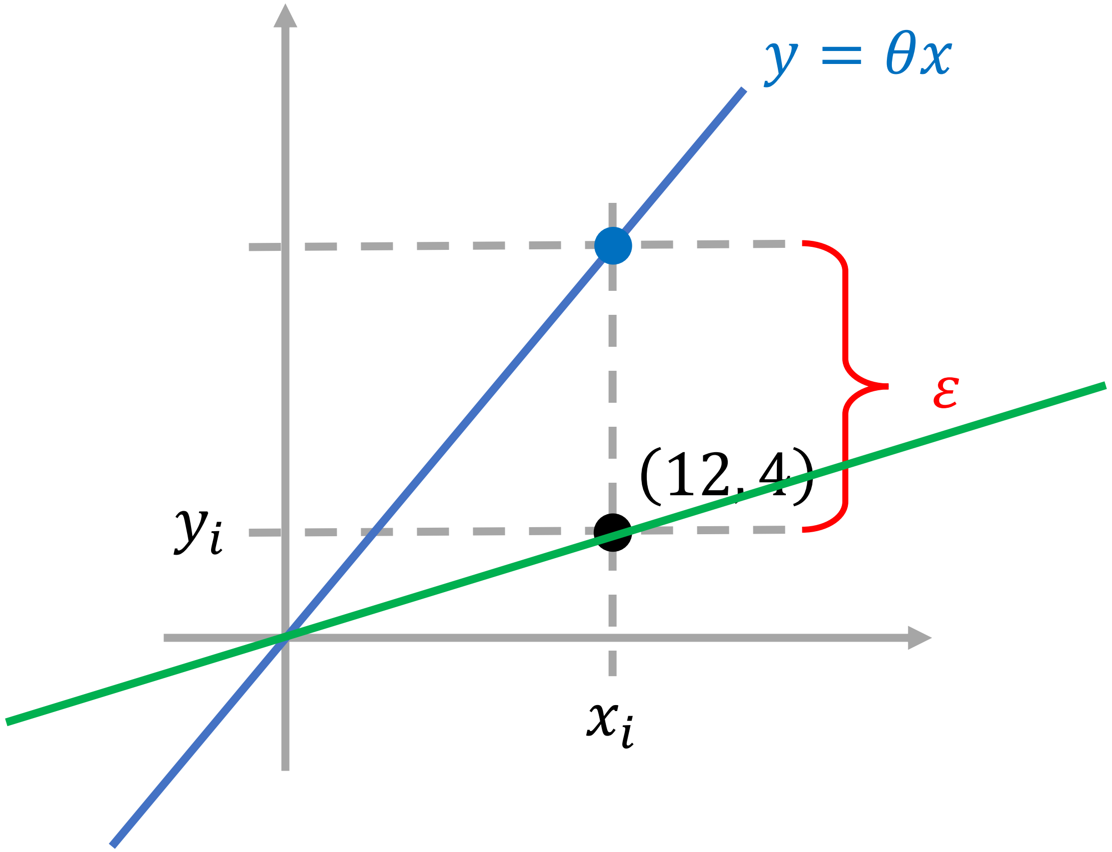
$y = \theta x \qquad m = 1$



$$y = \theta x \qquad m = 1$$



$$y = \theta x \qquad m = 1$$



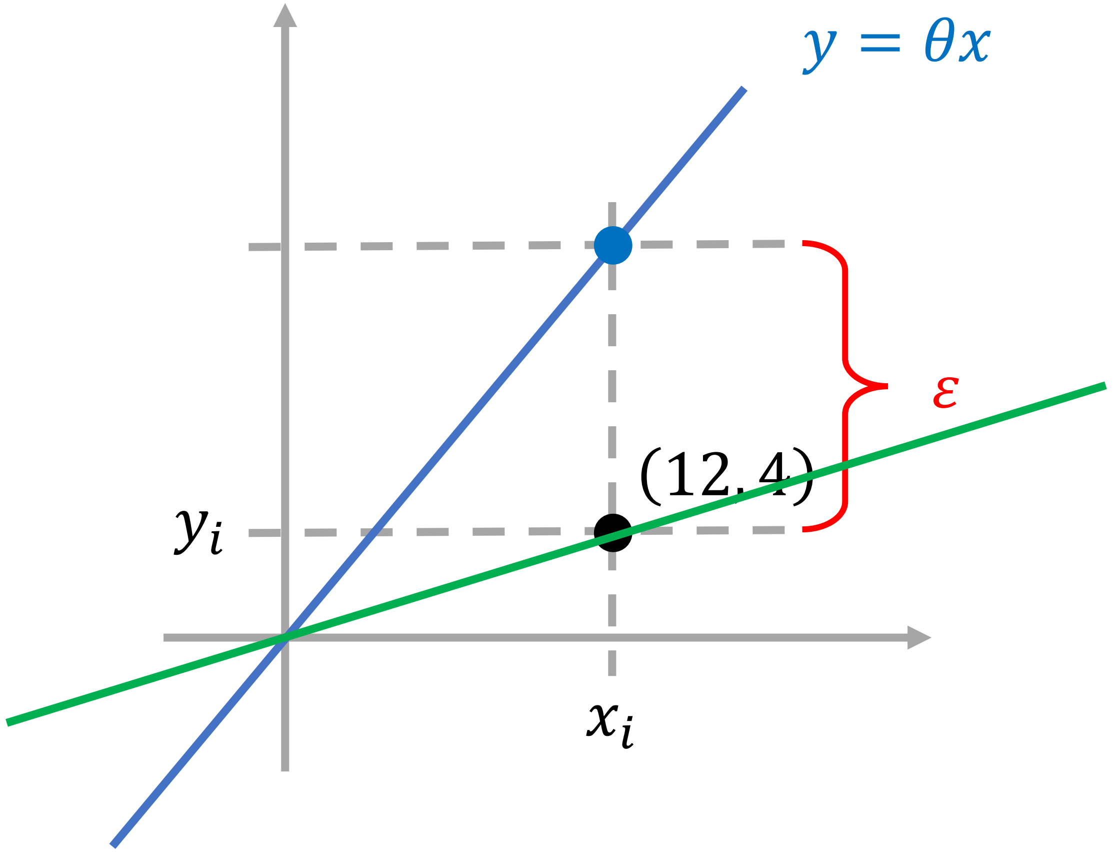
$$\begin{aligned} \varepsilon &= (\;y\; - 4)^2 \\ &= (12\theta - 4)^2 \\ &= (x_i\theta - y_i)^2 \end{aligned}$$

$$\frac{d\varepsilon}{d\theta} = 0 \quad \Rightarrow \quad \theta = 4/12 = y_i/x_i$$

$$\therefore \; y = \frac{4}{12} x$$

$$\begin{aligned} y &= f(\theta) \\ \varepsilon &= g(y) \end{aligned} \quad \Rightarrow \varepsilon = h(\theta)$$

$y = \theta x \qquad m = 1$

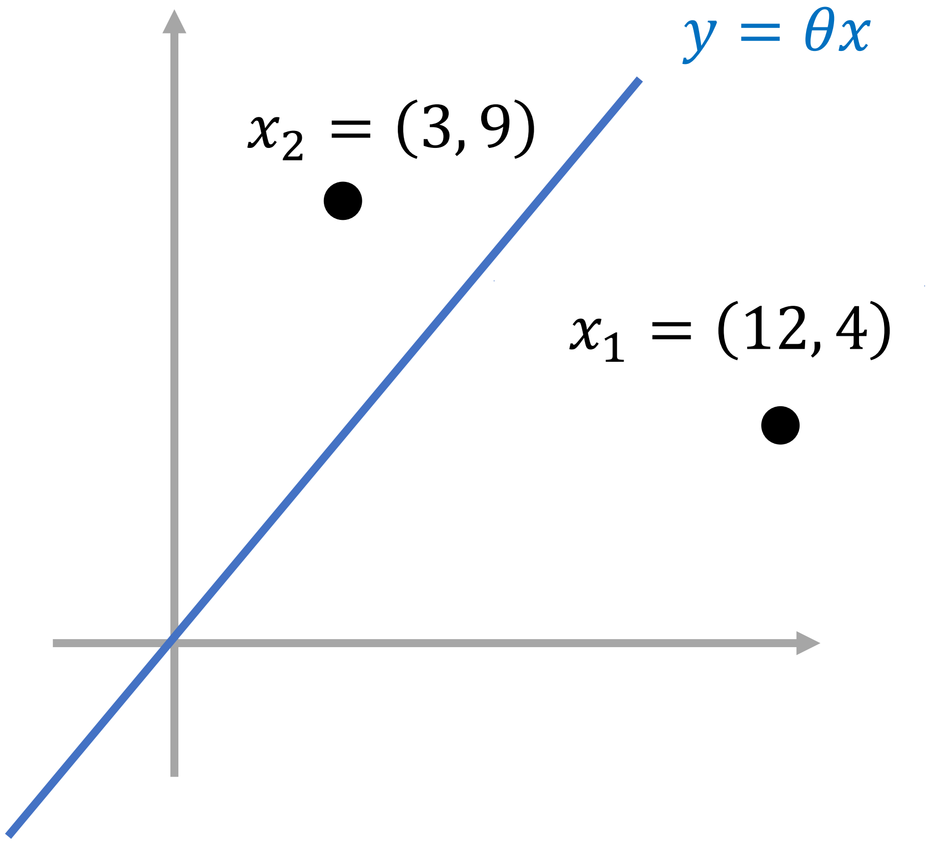


>_ Code

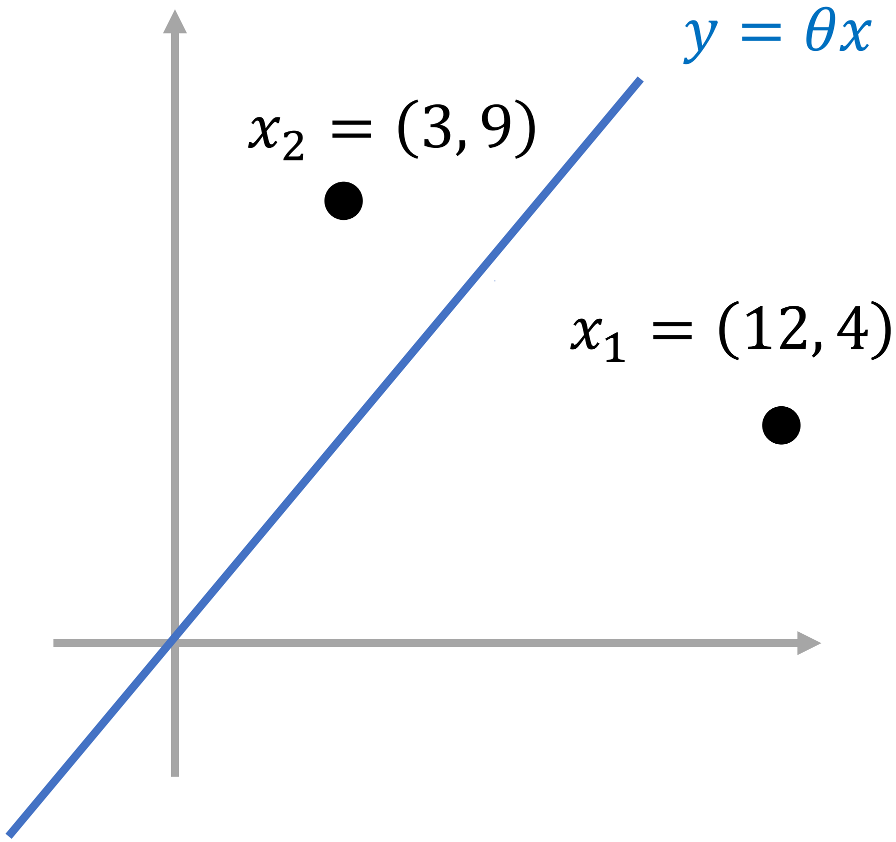
$\therefore \; y = \frac{4}{12}x$

```
import matplotlib.pyplot as plt
xi = 12 # data points o o o o o
yi = 4
theta = yi/xi
plt.scatter(xi,yi)
# model
x = np.linspace(10, 14, num = 5)
y = theta * x
plt.plot(x, y)
plt.axis('equal')
plt.show()
```

$y = \theta x \qquad m = 2$

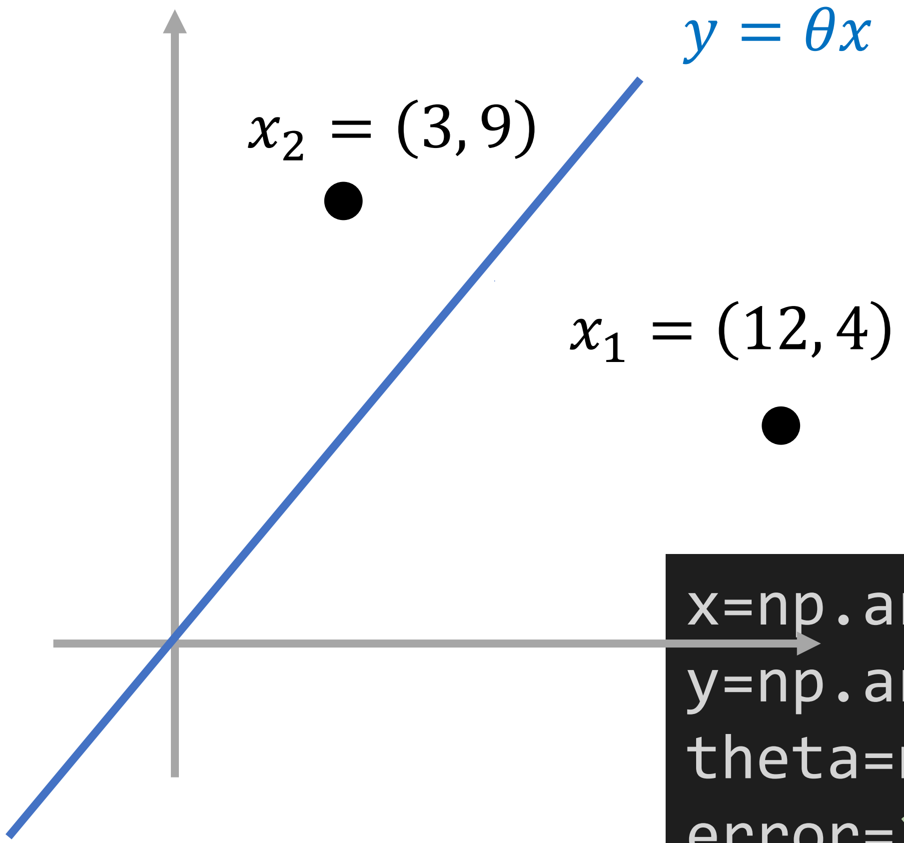


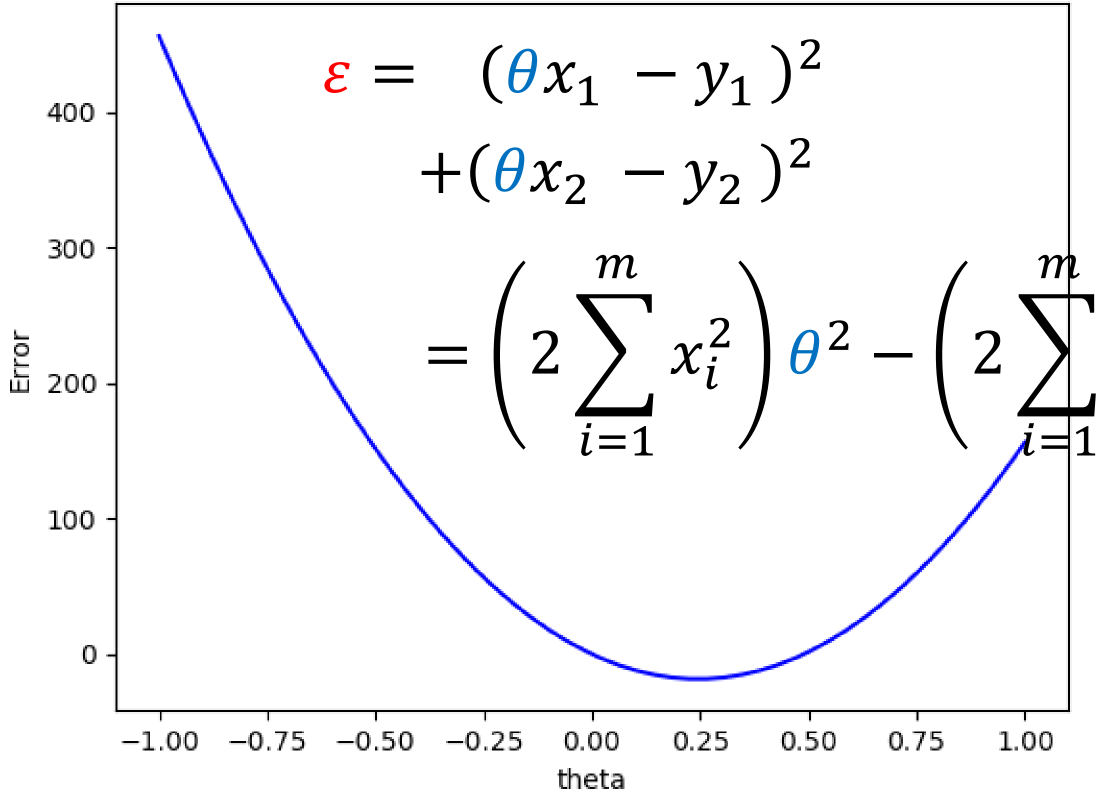
$y = \theta x \qquad m = 2$



$$\varepsilon = (\theta x_1 - y_1)^2 + (\theta x_2 - y_2)^2$$

$$y = \theta x \qquad m = 2$$

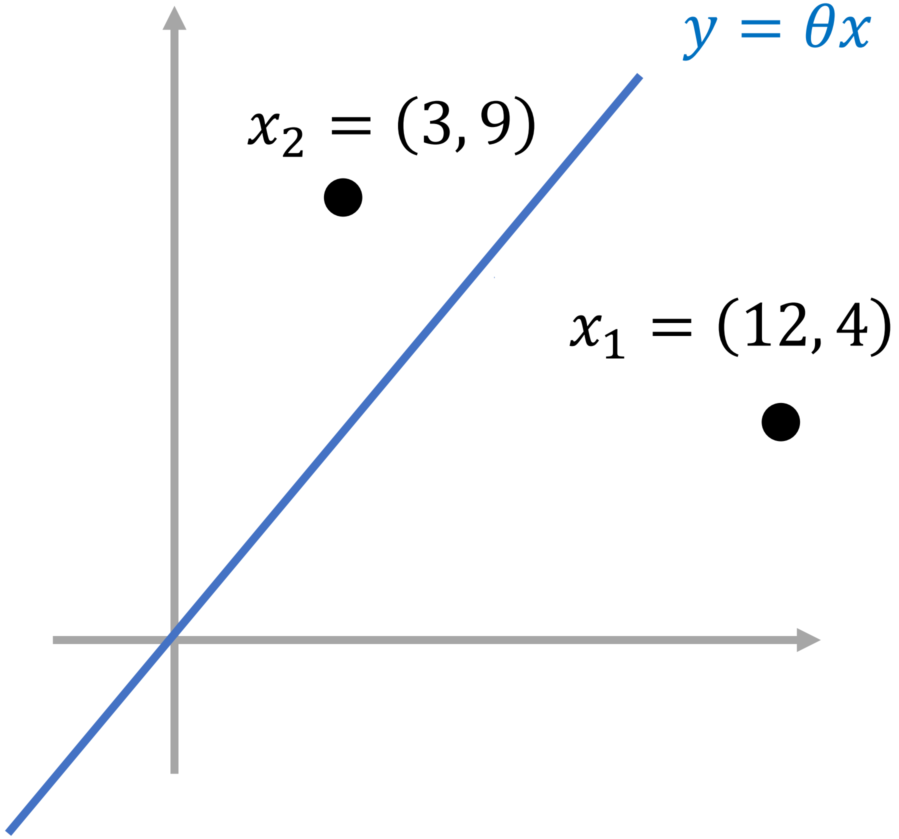




$$\varepsilon = (\theta x_1 - y_1)^2 + (\theta x_2 - y_2)^2 = \left(2\sum_{i=1}^{m} x_i^2\right)\theta^2 - \left(2\sum_{i=1}^{m} x_i y_i\right)\theta$$

```
x=np.array([12,3])
y=np.array([4,9])
theta=np.linspace(-1, 1)
error=2*np.sum(x**2)*theta**2-2*np.sum(x*y)*theta
plt.plot(theta,cost,'b-')
plt.show()
```

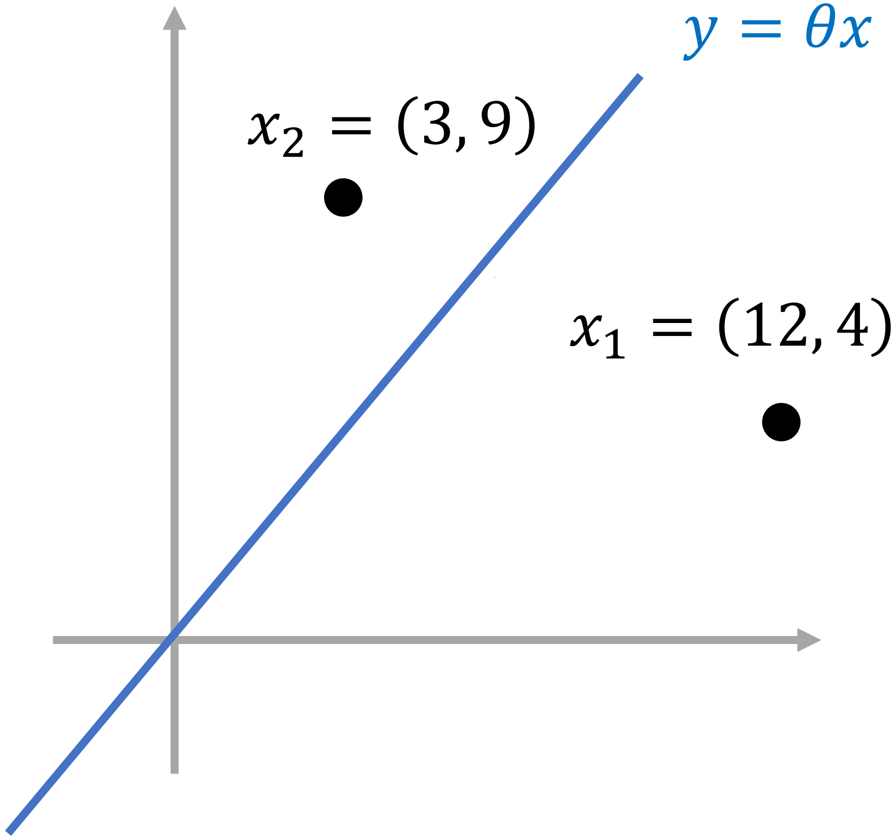
$$y = \theta x \qquad m = 2$$



$$\varepsilon = (\theta x_1 - y_1)^2 + (\theta x_2 - y_2)^2$$

$$\frac{d\varepsilon}{d\theta} = 0 \quad \Rightarrow \quad \theta = \frac{x_1 y_1 + x_2 y_2}{x_1^2 + x_2^2}$$

$$y = \theta x \qquad m = 2$$



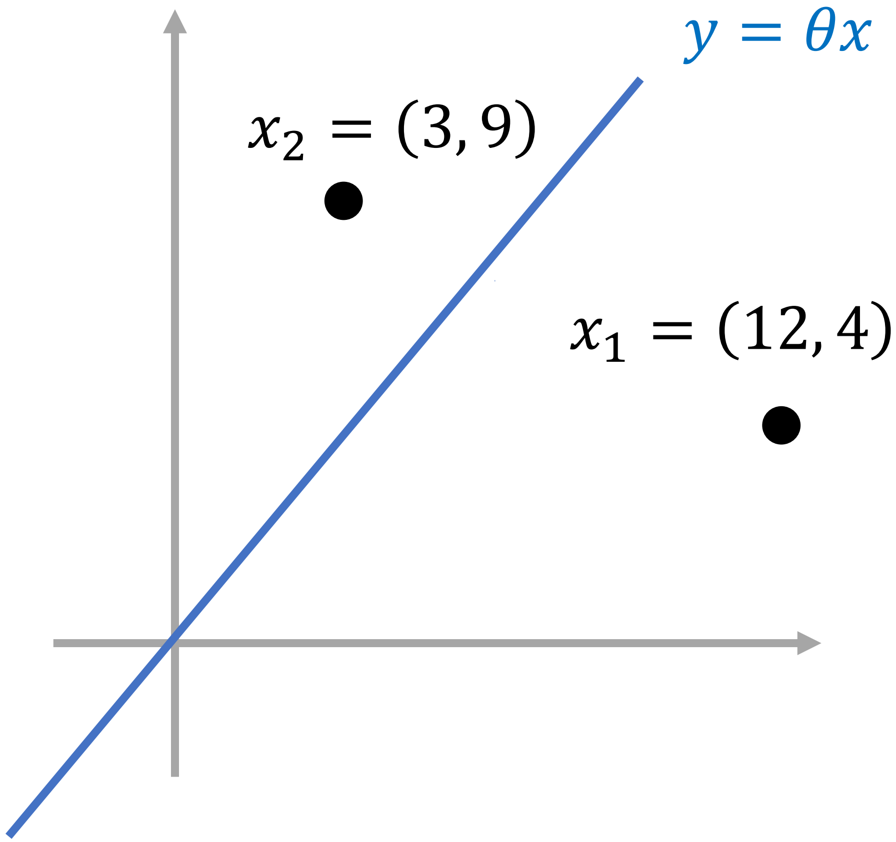
$$\varepsilon = (\theta x_1 - y_1)^2 + (\theta x_2 - y_2)^2$$

$$\frac{d\varepsilon}{d\theta} = 0 \quad \Rightarrow \quad \theta = \frac{x_1 y_1 + x_2 y_2}{x_1^2 + x_2^2}$$

Matrix form for fast vectorized computations

$$\begin{pmatrix} \square & \square \\ \square & \square \end{pmatrix}$$

$$y = \theta x \qquad m = 2$$



$$\varepsilon = (\theta x_1 - y_1)^2 + (\theta x_2 - y_2)^2$$

$$\frac{d\varepsilon}{d\theta} = 0 \quad \Rightarrow \quad \theta = \frac{x_1 y_1 + x_2 y_2}{x_1^2 + x_2^2}$$

Matrix form for fast vectorized computations

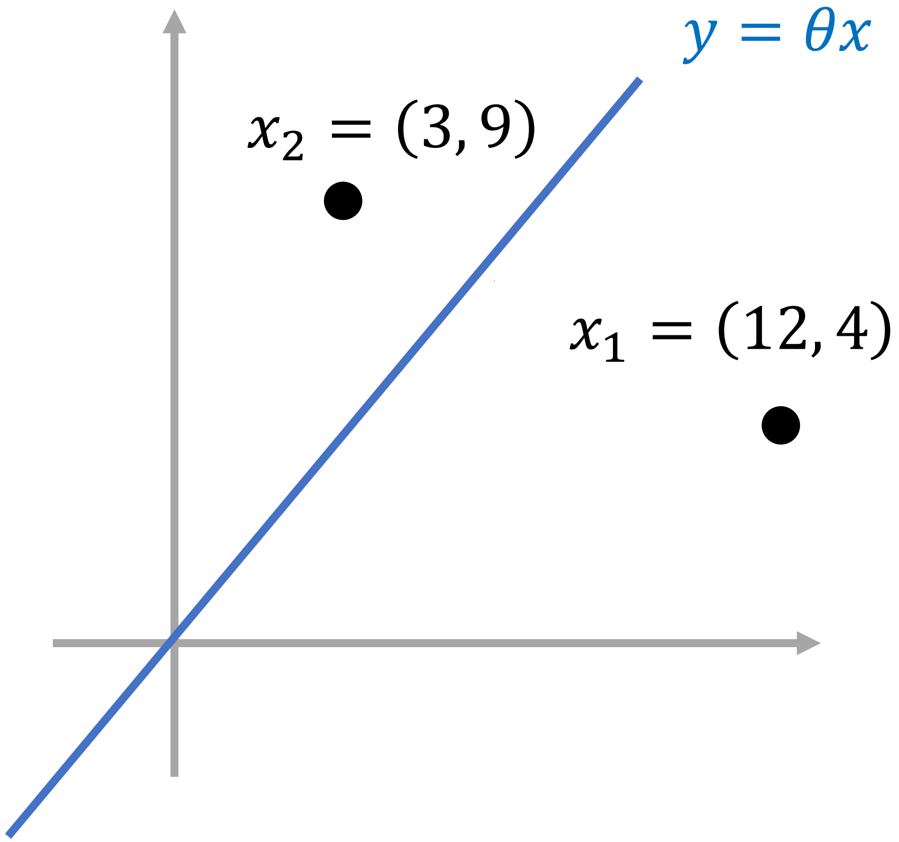
$$\boldsymbol{x} := \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \qquad \boldsymbol{y} := \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

$$\Rightarrow \quad \theta = \frac{(x_1, x_2)^T \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}}{(x_1, x_2)^T \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}} = \boldsymbol{x}^T \boldsymbol{y} / \boldsymbol{x}^T \boldsymbol{x}$$

>_ Code

$$y = \theta x \qquad m = 2$$

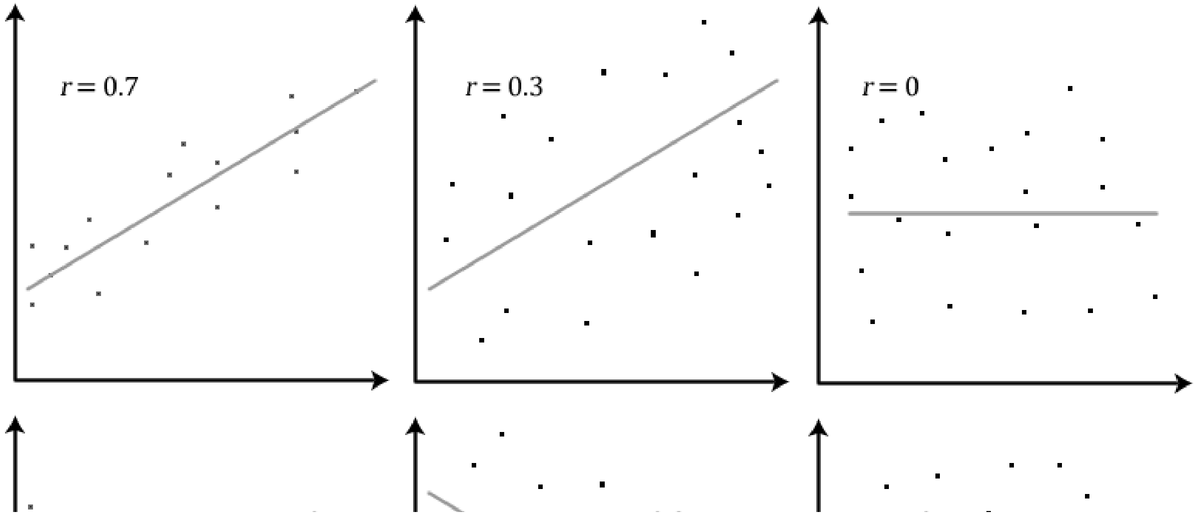
$$\theta = \frac{(x_1, x_2)^T \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}}{(x_1, x_2)^T \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}} = \boldsymbol{x}^T \boldsymbol{y} / \boldsymbol{x}^T \boldsymbol{x}$$



>_ Code

```python
import numpy as np
import matplotlib.pyplot as plt
xi=np.array([3, 12]) # data points o o o
yi=np.array([9, 4])
theta=sum(xi*yi)/sum(xi**2)
plt.scatter(xi,yi)
# model
x=np.linspace(min(xi),max(xi),num=5)
y=theta*x
plt.plot(x, y)
plt.axis('equal')
plt.show()
```

Model Evaluation. Pearson correlation coefficient



Pearson Correlation Coefficient

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

[redacted]

Pearson

We need to calc

Age

Height (cm)

$\Sigma(x_i^2) =$

$\Sigma(y_i^2) =$

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \quad \text{(Eq.1)}$$

where:

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

[redacted]

The formula for ρ can be expressed in terms of mean and expectation. Since

$$\operatorname{cov}(X,Y) = \operatorname{E}[(X-\mu_X)(Y-\mu_Y)],\text{[7]}$$

the formula for ρ can also be written as

$$\rho_{X,Y} = \frac{\operatorname{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \quad \text{(Eq.2)}$$

where:

- σ_Y and σ_X are defined as above
- μ_X is the mean of X

[redacted]

Model Evaluation. R^2 Statistic

Pearson Correlation Coefficient

r =

To calculate R^2, we use the formula

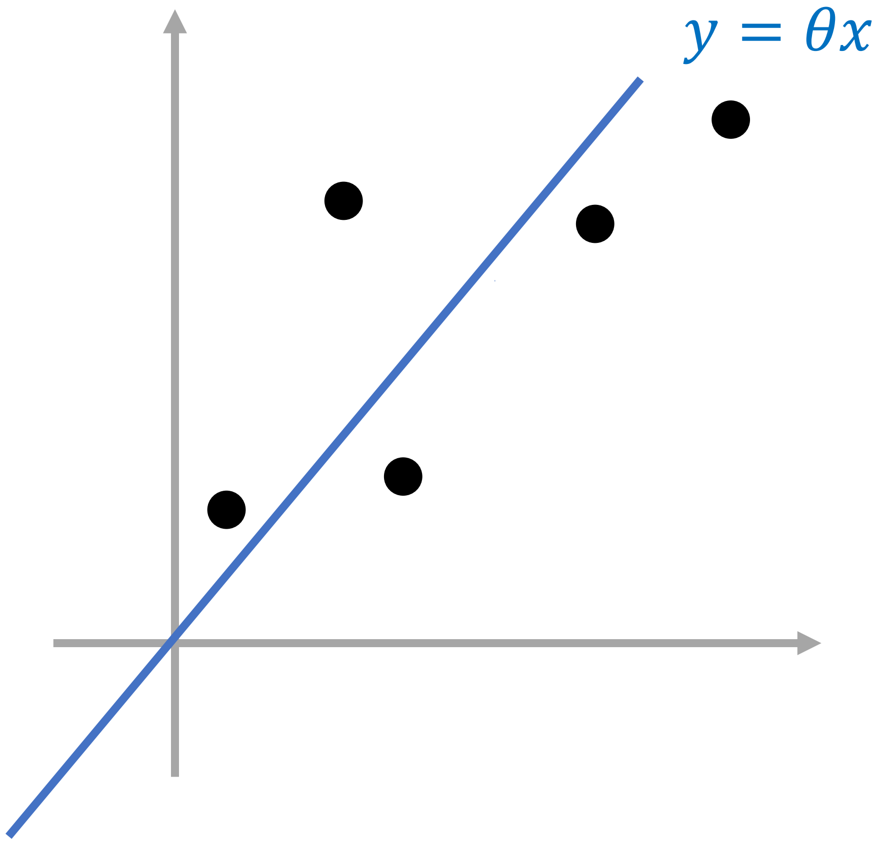
$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} \tag{3.17}$$

where $\text{TSS} = \sum(y_i - \bar{y})^2$ is the *total sum of squares*, and RSS is defined in (3.16). TSS measures the total variance in the response Y, and can be thought of as the amount of variability inherent in the response before the

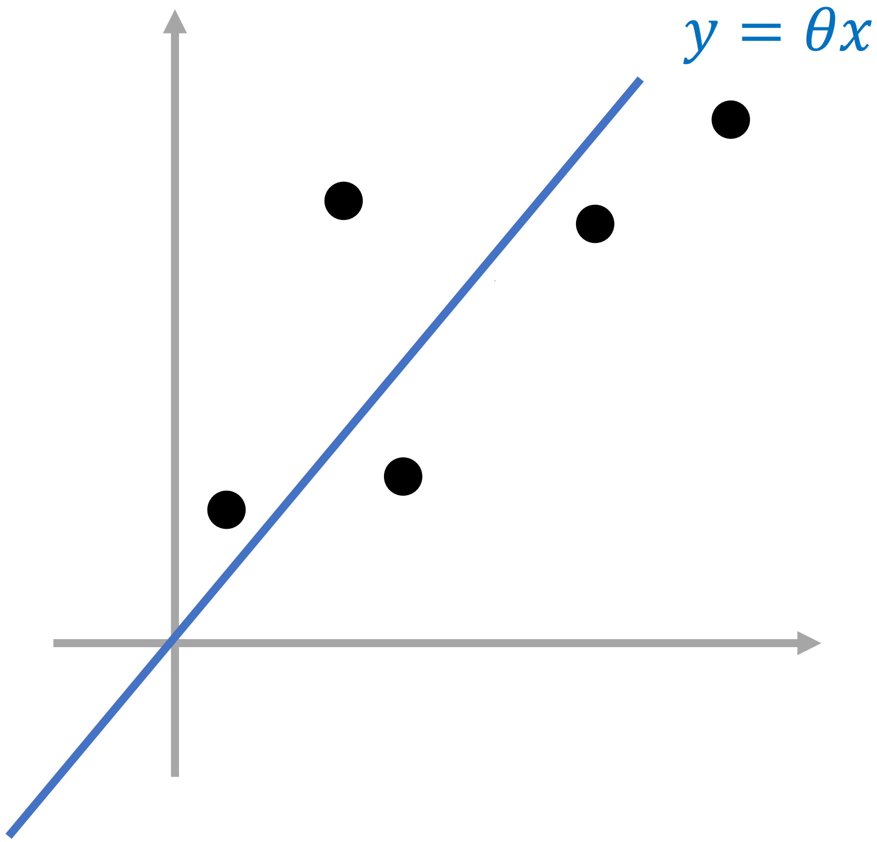
total sum of squares

Pages 69,70,71. Gareth & Hastie 2014.

$y = \theta x \qquad m$

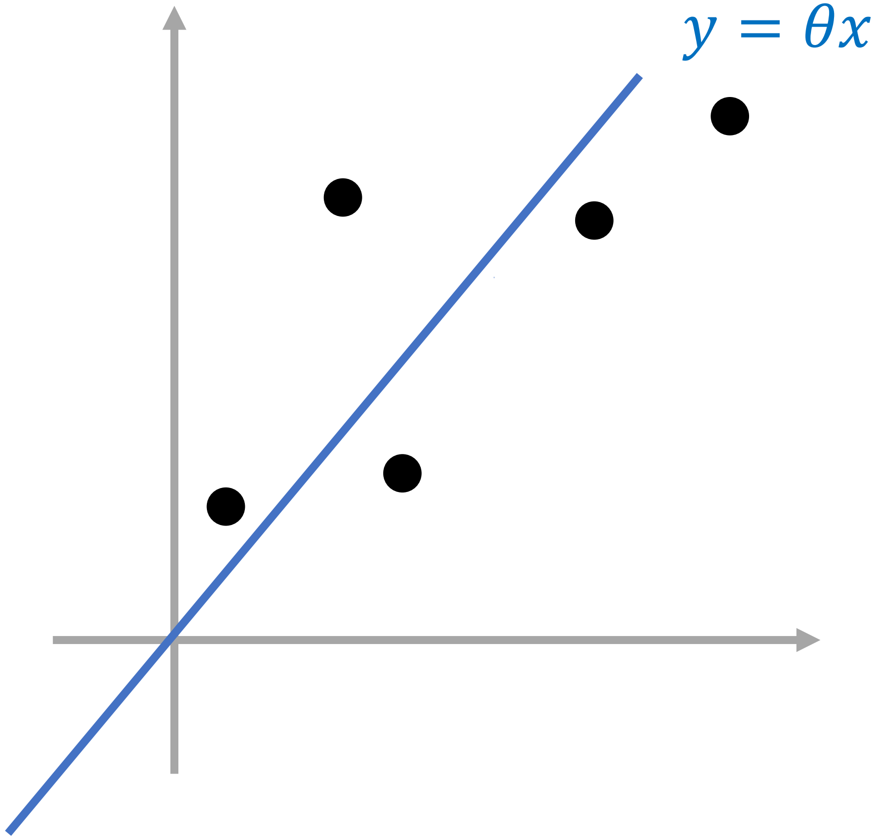


$y = \theta x \qquad m$



$$\begin{aligned} \varepsilon = \ & (\theta x_1 - y_1)^2 \\ & + (\theta x_2 - y_2)^2 \\ & \quad \vdots \\ & + (\theta x_m - y_m)^2 \end{aligned}$$

$y = \theta x \qquad m$



$$\varepsilon = (\theta x_1 - y_1)^2 + (\theta x_2 - y_2)^2 + \cdots + (\theta x_m - y_m)^2$$

$$\frac{d\varepsilon}{d\theta} = 0 \quad \Rightarrow \quad \theta = \frac{\sum_{i=1}^{m} x_i y_i}{\sum_{i=1}^{m} x_i^2}$$

Matrix form for fast vectorized computations

$$\boldsymbol{x} := \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} \qquad \boldsymbol{y} := \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix}$$

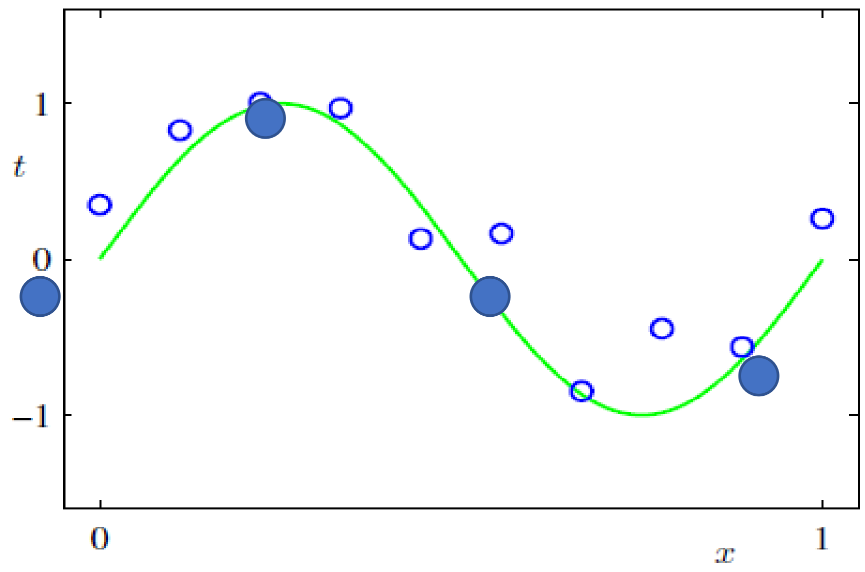
$$\Rightarrow \quad \theta = \boldsymbol{x}^T \boldsymbol{y} / \boldsymbol{x}^T \boldsymbol{x}$$

>_ Code

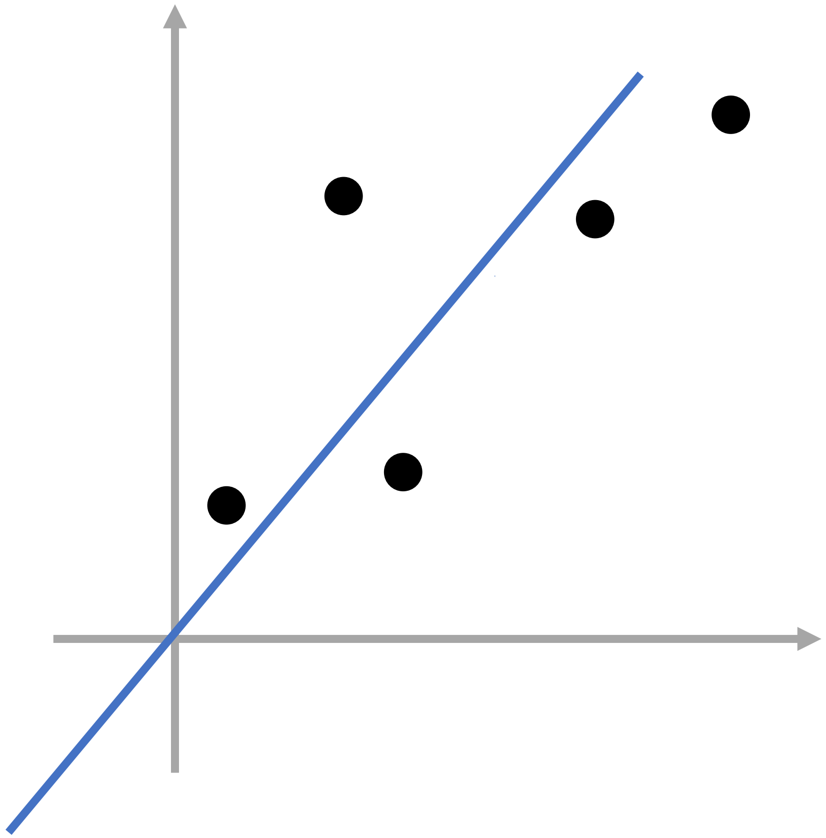
Your turn

1. $y = \theta x + 5$

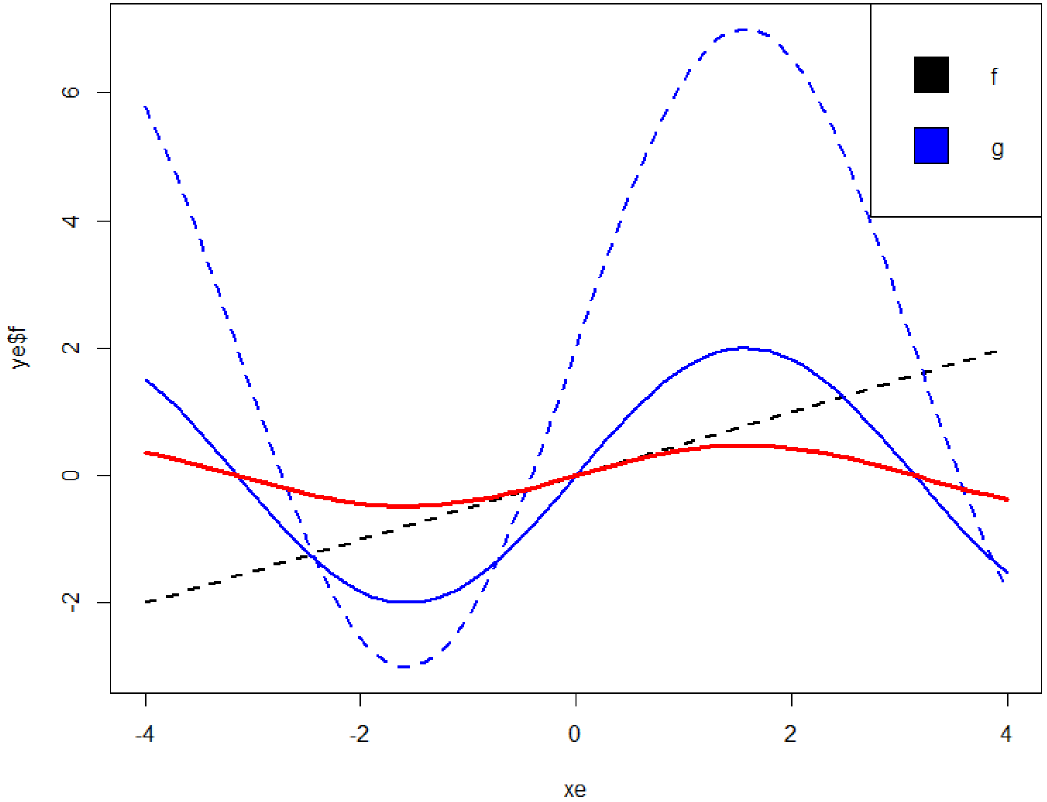
2. $y = A\sin(x)$



$$y = \theta x + b$$



Aplicacion (reescalamiento de datos). Las dos curvas punteadas son los datos originales (regularmente muestreados). La curva azul punteada (g) es la curva a reescalar verticalmente (dilatación/contraccion) y a trasladar de tal manera que se parezca lo más posible a la curva punteada negra (g). La solución mediante regresión es la curva roja continua. Encuentren la formula matematica para obtener la curva continua. Esto fue aplicado en el caso de well logs para hacer el match y en recibos de CFE con datos de sensors.



>_ Code

Evaluation metrics for regression

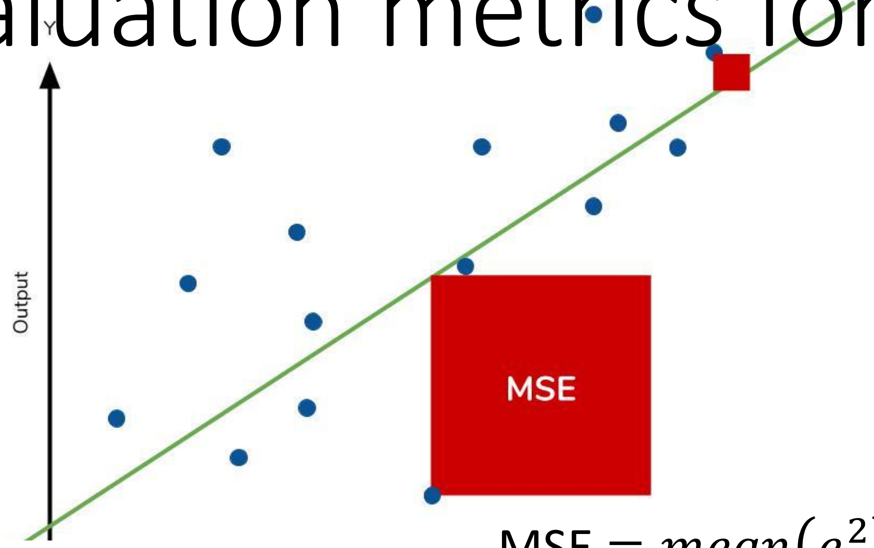
metrics.explained_variance_score(y_true, …)	Explained variance regression score function
metrics.max_error(y_true, y_pred)	max_error metric calculates the maximum residual error.
metrics.mean_absolute_error(y_true, y_pred, *)	Mean absolute error regression loss
metrics.mean_squared_error(y_true, y_pred, *)	Mean squared error regression loss
metrics.mean_squared_log_error(y_true, …)	Mean squared logarithmic error regression loss
metrics.median_absolute_error(y_true, y_pred, *)	Median absolute error regression loss
metrics.r2_score(y_true, y_pred, *[, …])	R^2 (coefficient of determination) regression score function.
metrics.mean_poisson_deviance(y_true, y_pred, *)	Mean Poisson deviance regression loss.
metrics.mean_gamma_deviance(y_true, y_pred, *)	Mean Gamma deviance regression loss.
metrics.mean_tweedie_deviance(y_true, y_pred, *)	Mean Tweedie deviance regression loss.

https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics

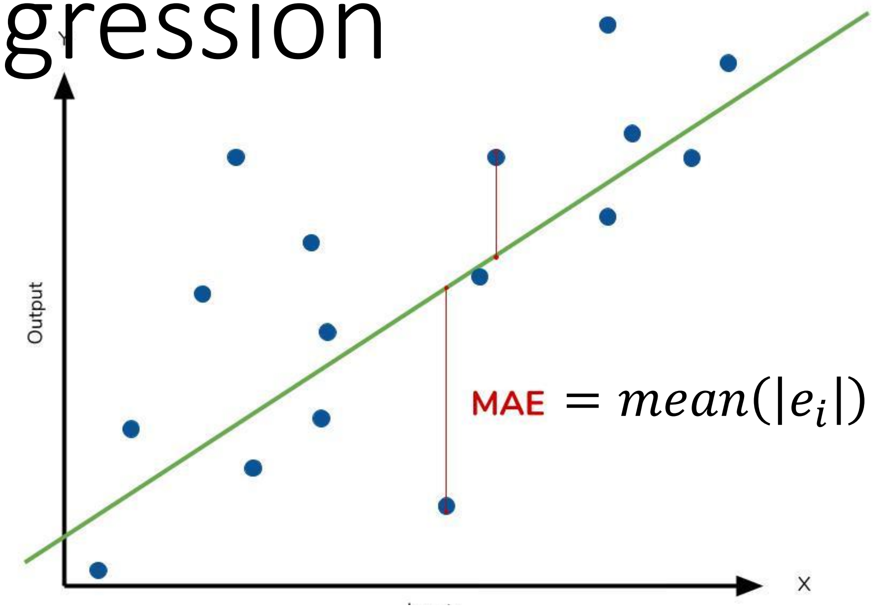
https://www.dataquest.io/blog/understanding-regression-error-metrics/

Evaluation metrics for regression

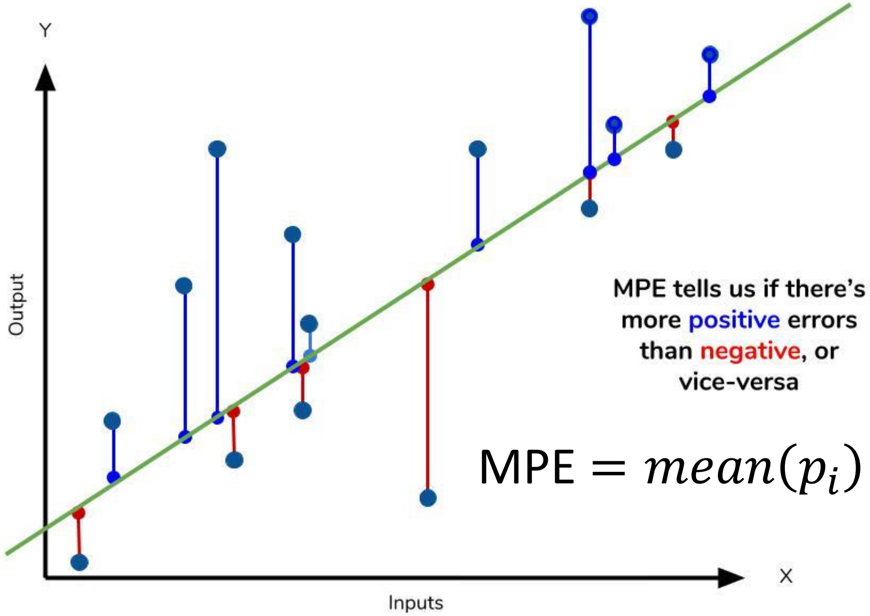
$$e_i = y_i - \hat{y}_i$$



$$\text{MSE} = mean\left(e_i^2\right)$$



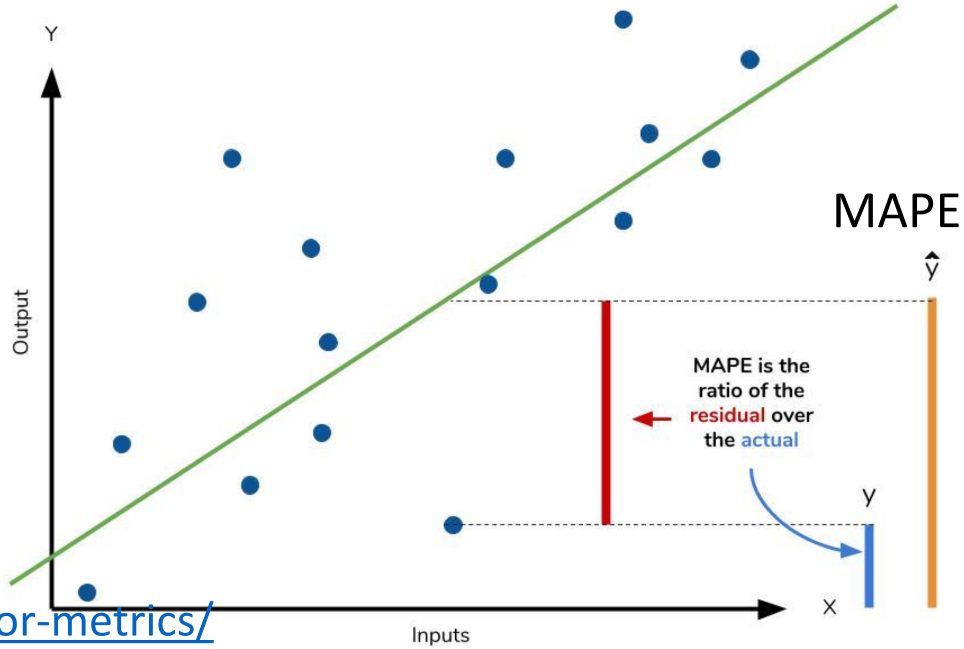
$$\text{MAE} = mean(|e_i|)$$



$$\text{MPE} = mean(p_i)$$

$$p_i = 100\,\frac{e_i}{y_i}$$

$$\text{MAPE} = mean(|p_i|)$$



https://www.dataquest.io/blog/understanding-regression-error-metrics/

Evaluation metrics for regression (Accuracy)

$$\text{Mean absolute error: MAE} = \text{mean}(|e_t|),$$

$$\text{Root mean squared error: RMSE} = \sqrt{\text{mean}(e_t^2)}.$$

$$\text{Mean absolute percentage error: MAPE} = \text{mean}(|p_t|).$$

$$\text{sMAPE} = \text{mean}\left(200|y_t - \hat{y}_t|/(y_t + \hat{y}_t)\right).$$

For time series:

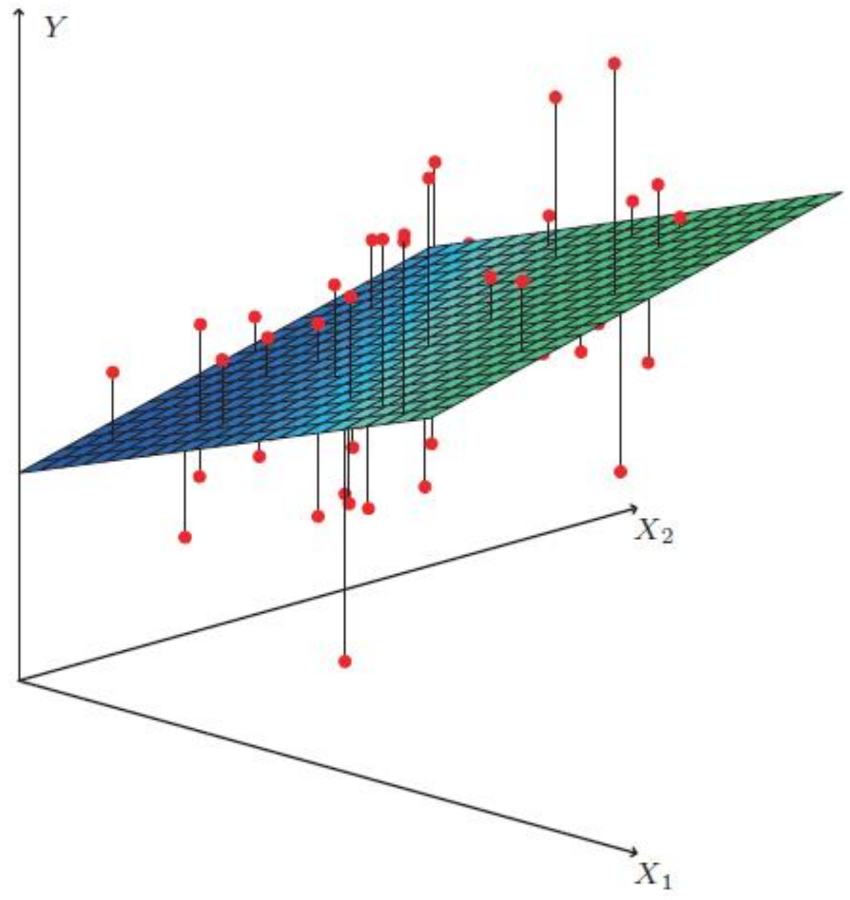
[2013 A Survey of forecast error measures]

https://otexts.com/fpp2/accuracy.html

Rob J Hyndman, Anne B Koehler (2006). Another look at measures of forecast accuracy

2009 Elena Deza, Michel Marie Deza_Encyclopedia of Distances, metric dissimilarity error

Multiple linear Regression



Page 73, Gareth & Hastie, 2014.