LINEAR REGRESSION INTUITION

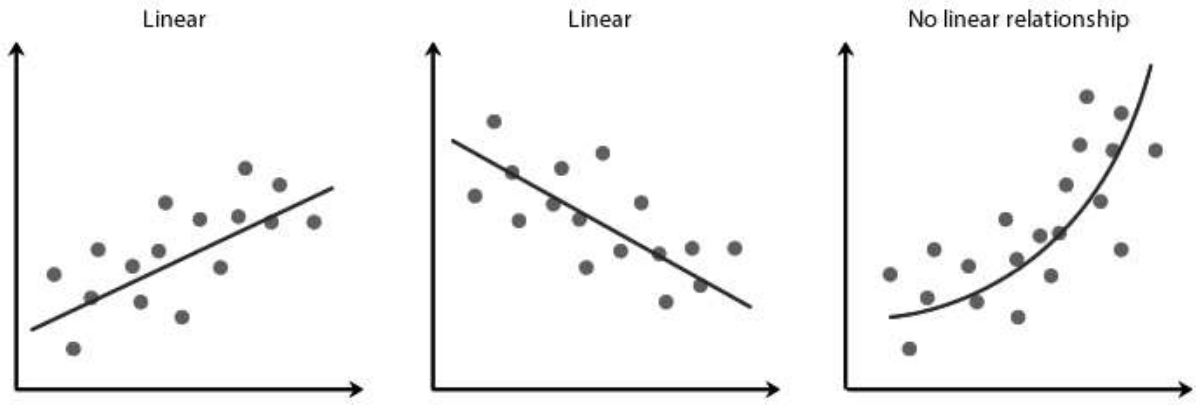
DEF^{N}:

Linear regression is the supervised machine learning model in which model finds best fit line between independent and dependent variable.

OR

DEF^{N}:

Linear regression attempts to model the relationship between two variables by fitting a linear equation (a straight line to the observed data).



There are two types linear regression (excluding regularized regressions),

1. Simple Linear Regression.
2. Multi-Linear Regression.

SIMPLE LINEAR REGRESSION

A statistical method in which only one independent variable is present and the model has to find the linear relationship of it with the dependent variable.

- One variable, denoted x, is regarded as the explanatory or independent variable.
- One variable, denoted y, is regarded as the response, predictor, outcome or dependent variable.

General EQ^N,

$$\boldsymbol{y = \beta_0 + \beta_1 x_1 + \varepsilon}$$

Consider y as your income and $x_1$as your education, the more education you have the higher income you will get.

β_1 works as multiplier and quantifies the income, while $\beta_0$is constant value (consider it as minimum wage), so if you have 0 education ($x_1 = 0$) you will get a minimum wage.

ε is the error (on average error is 0).

Linear Regression EQ^N,

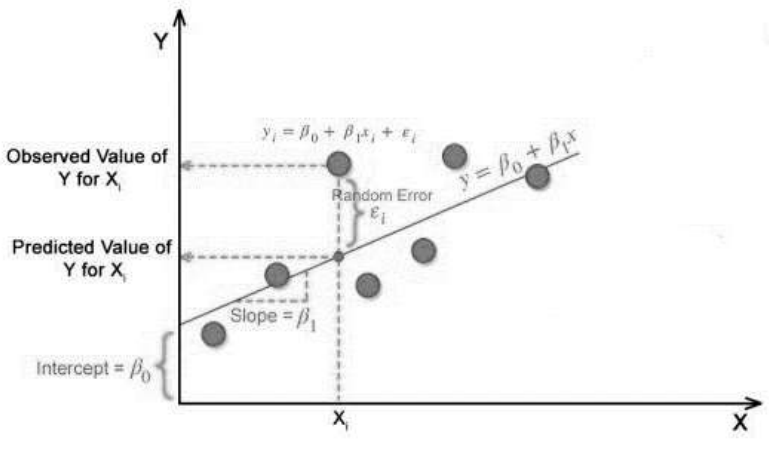
$$\boldsymbol{\hat{y} = b_0 + b_1 x_1}$$

$\boldsymbol{\hat{y}}$ **:** Estimated predicted value.

$\boldsymbol{b_0}$ **:** Intercept / constant.

$\boldsymbol{b_1}$ **:** Slope / quantifier.

$\boldsymbol{x_1}$ **:** Sample data for independent variable.



MULTI-LINEAR REGRESSION

A statistical method can be used to analyze the relationship between single dependent variable and multiple independent variable.

- Multiple variable, denoted $x_1, x_2, x_3, \ldots\ldots, x_n$ is regarded as the explanatory or independent variables.
- One variable, denoted y, is regarded as the response, predictor, outcome or dependent variable.

Multi-Linear Regression EQ^N,

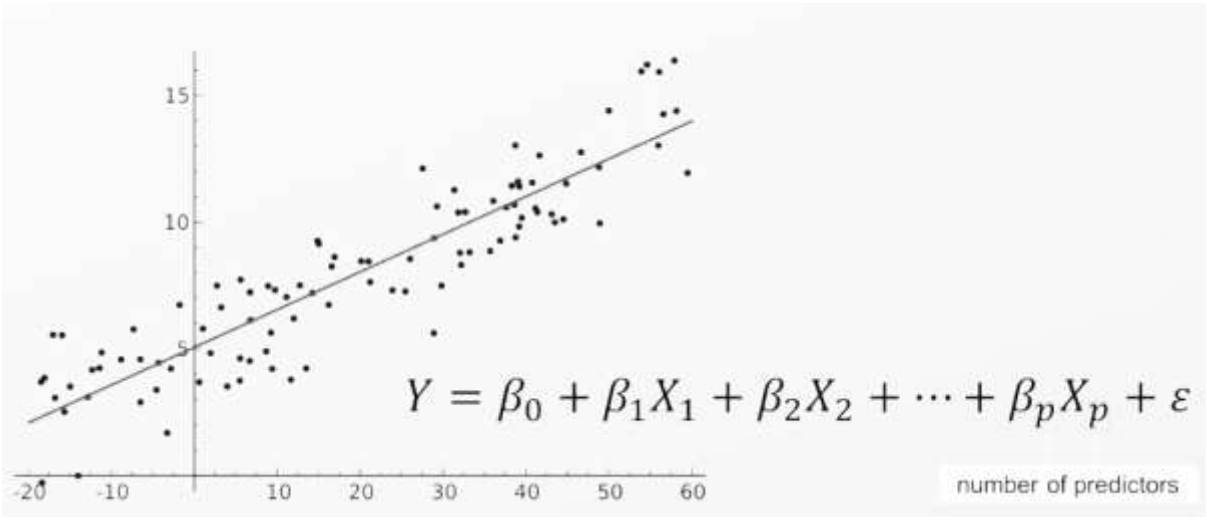
$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots\ldots + b_n x_n + \varepsilon$$

$\hat{y}$ **:** Estimated predicted value.

b_0 **:** Intercept / constant.

$b_1, b_2, b_3, \ldots\ldots, b_n$ **:** Slope / quantifier.

$x_1, x_2, x_3, \ldots\ldots, x_n$ **:** Sample data for independent variable.



ASSUMPTIONS OF LINEAR REGRESSION:

Regression is a parametric approach, which means that it makes assumptions about the data for the purpose of analysis. For

successful regression analysis, it's essential to validate the following assumptions.

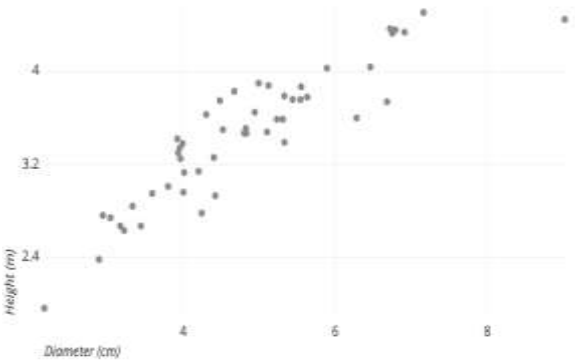
There are following five assumptions of linear regression;

1. Linearity.
2. No Endogeneity.
3. Normality and Homoscedasticity.
4. No Auto-Correlation of Residuals.
5. No Multi-Collinearity.

LINEARITY:

Linear regression needs the relationship between the independent and dependent variables to be linear.

The linearity assumption can best be tested with scatter plots, the following two examples depict two cases, where no linearity and linearity are present.



Linear Relationship

Non-Linear Relationship

NO ENDOGENEITY

Endogeneity refers to situations in which a predictor (e.g. $\hat{y}$) in a linear regression model is correlated to the error term.
Endogeneity is caused by omitted variable bias.

Omitted variable bias occurs when a statistical model fails to include one or more relevant variable.

Example:

Say, you want to predict the effect of **education** on people's **salaries.**

We know that only education is not enough for predicting salaries, ability(skill) is also an important factor.

EQ^{N},

$$\boldsymbol{Salary = \beta_0 + \beta_1 * Education + \varepsilon}$$

Not including the ability this will cause the omitted variable bias.

Salary is also likely to be related to ability, which we previously decided to exclude. In turn, ability is also likely related to the level of education a person attains, as those with greater ability are likely to pursue higher education.

The omitted variable (ability) affects your analysis of both education (the independent variable) and earnings (the dependent variable).

Ability is in the error term due to endogeneity. Since ability is not in the regression model, our estimate of β_1 will absorb some of the effect of ability (as it is correlated with education).

The estimate is now biased, so we can no longer make a causal claim about education.

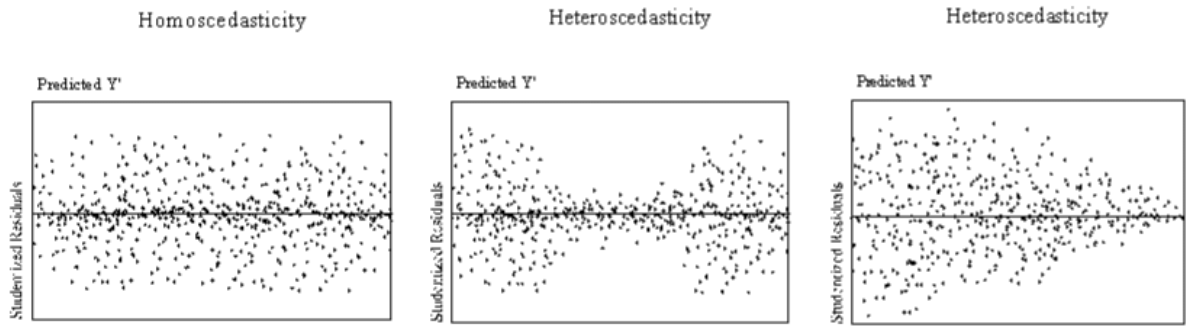
NORMALITY AND HOMOSCEDASTICITY

Normality means your variables must follow gaussian distribution, you can check normality using histograms or Q-Q plot.

If normality is not observed, transform the variables (using log transformation, exponential transformation, reciprocal transformation, square root transformation, boxcox transformation etc.).

Homo means same **Scedasticity** means scatter, homoscedasticity means having same scatter or generally having same variance.

We can check homoscedasticity by plotting predicted values against residuals, if the spread is equal then homoscedasticity condition is fulfilled.

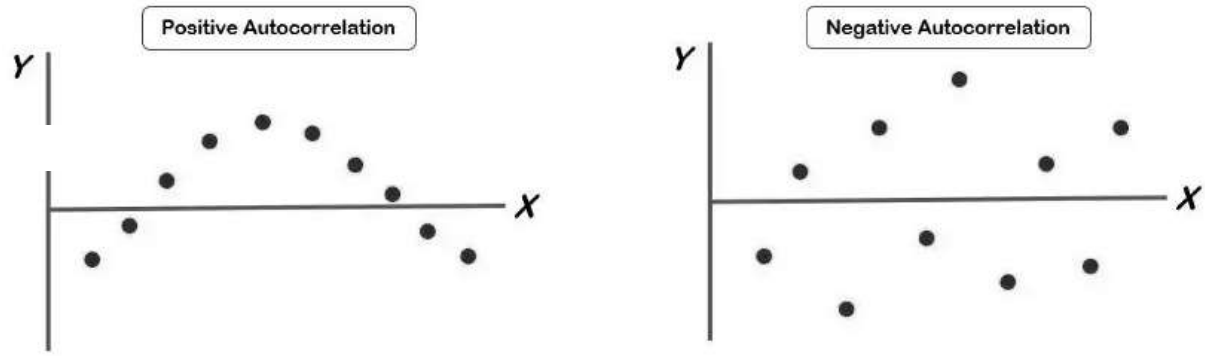


The advantage of homoscedasticity is that the best fit line can fit anywhere in the plot due to the spread of residuals avoiding overfitting and underfitting scenarios.

NO AUTO-CORRELATION OF RESIDUALS:

No Auto-correlation of residuals means there should be no correlation between the residuals, meaning when you plot the residuals they must not follow a specific pattern (+ve autocorrelation), they must be randomly dispersed (-ve autocorrelation).

This suggests that the model is correctly capturing the underlying relationships in the data and that the predictions made by the model are reliable.



NO MULTI-COLLINEARITY

Multicollinearity is observed when two or more independent variables are correlated to one another.

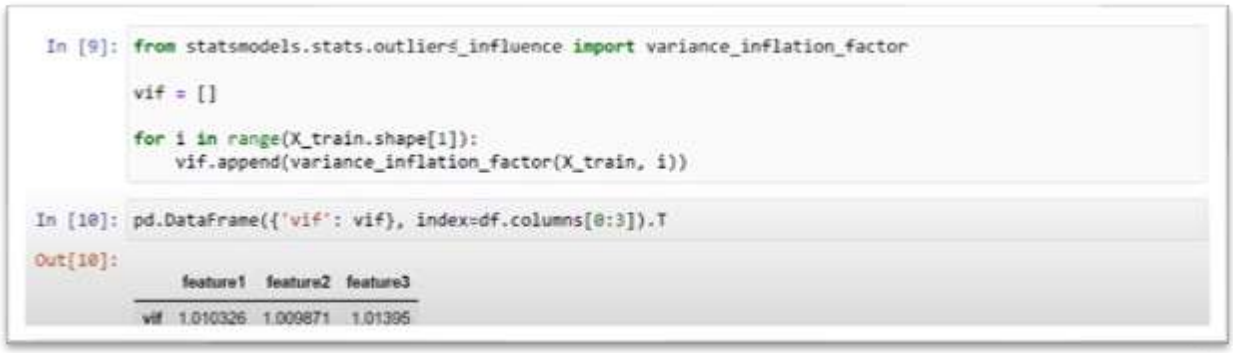
If we found any variable causing collinearity problem we must drop it.

Example:

Consider a kid who loves watching tv and eating snacks. Now, the kid is watching tv while eating snacks, can you tell what activity (either watching tv or eating snacks) does the kid love more? No, right? because he loves both therefore we can't decide as watching

tv and eating snacks are correlated, in order to decide we have to drop one variable so that we can decide.

The best method to test for collinearity assumption is the **Variance Inflation Factor** method.



```
In [9]: from statsmodels.stats.outliers_influence import variance_inflation_factor

        vif = []

        for i in range(X_train.shape[1]):
            vif.append(variance_inflation_factor(X_train, i))

In [10]: pd.DataFrame({'vif': vif}, index=df.columns[0:3]).T

Out[10]:
```

	feature1	feature2	feature3
vif	1.010326	1.009871	1.01395

COST FUNCTION

Cost function is the calculation of the residuals between predicted values and actual values.

It tells us how badly the model is performing.

It helps find the optimal values of $\boldsymbol{\beta_0}$ & $\boldsymbol{\beta_1}$, such a way that we get as least value of cost function as possible.

$$J_{(\beta_0,\beta_1)} = \frac{1}{2m}\sum_{i=1}^{m}[\, f_{\beta_0,\beta_1}(x^i) - (y^i)]^2$$

$$J_{(\beta_0,\beta_1)} = \frac{1}{2m}\sum_{i=1}^{m}[\, f_{\beta_0,\beta_1}(\beta_1 x^i + \beta_0) - (y^i)]^2$$

Example:

Let's consider an example of cost function using simple linear regression.

As we remember the linear regression equation,

$$\hat{y} = b_0 + b_1 x_1$$

For simple linear regression,

we substitute $b_0 = 0$

$$\hat{y} = b_1 x_1$$

Substituting $b_0 = 0$ does nothing more than passing our best fit line from origin (you can try putting some random values of x_1 and b_1 considering $b_0 = 0, 1, 2$ or any value of your choice and you will get the idea of how b_0 is affecting the eqn).

Substituting $b_0 = 0$ will transform cost function as,

$$J_{(\beta_1)} = \frac{1}{2m} \sum_{i=1}^{m} [\, f_{\beta_1}(\beta_1 x^i) - (y^i)]^2$$

where,

- β_1 is fixed
- x^i is independent variable
- y^i is observed value
- m is total number of observations

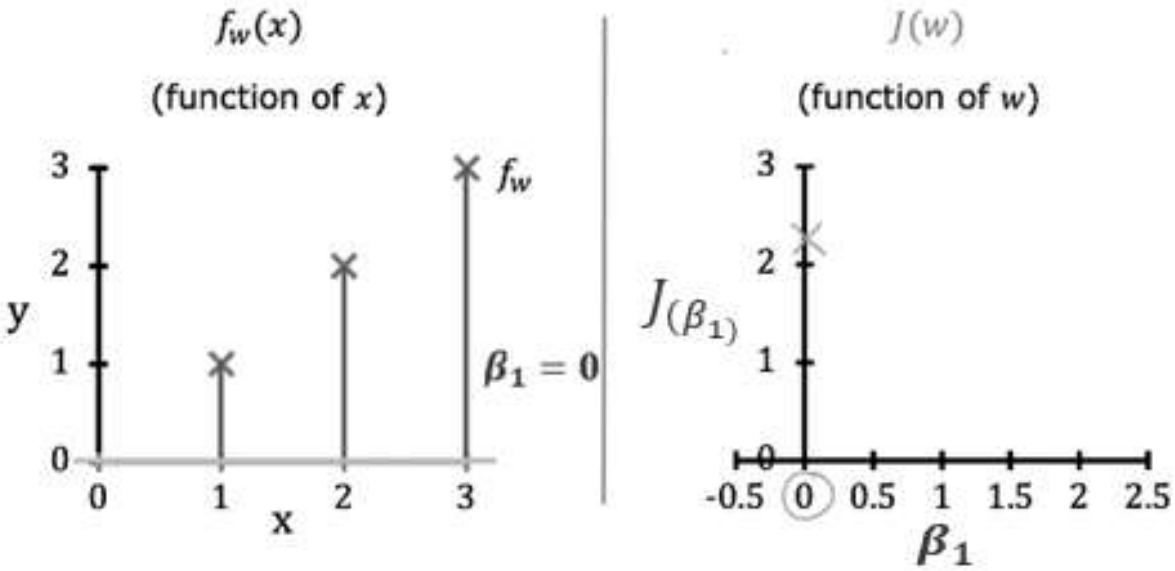
Considering following values of x & y.

x	y
1	1
2	2
3	3
.	.
.	.
n	n

Let's fix the value of $\boldsymbol{\beta_1 = 0}$,

$$J_{(\beta_1)} = \frac{1}{2*3}\sum_{i=1}^{3}[(0*1-1)^2 + (0*2-2)^2 + (0*3-3)^2]$$

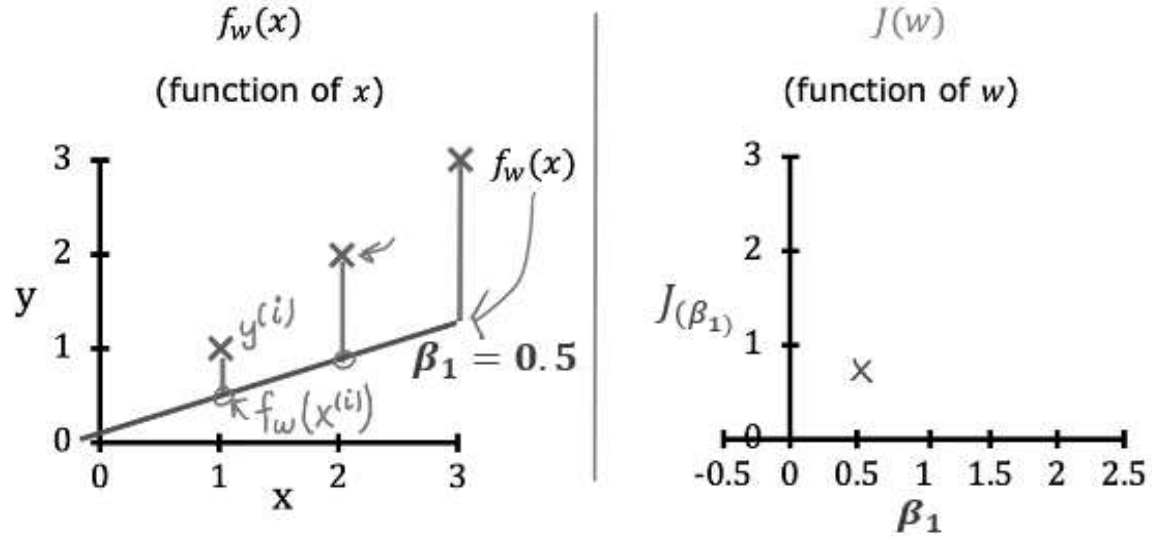
$$J_{(\beta_1)} \approx 2.3$$



Let's fix the value of $\boldsymbol{\beta_1 = 0.5}$,

$$J_{(\beta_1)} = \frac{1}{2*3}\sum_{i=1}^{3}[(0.5*1-1)^2 + (0.5*2-2)^2 + (0.5*3-3)^2]$$

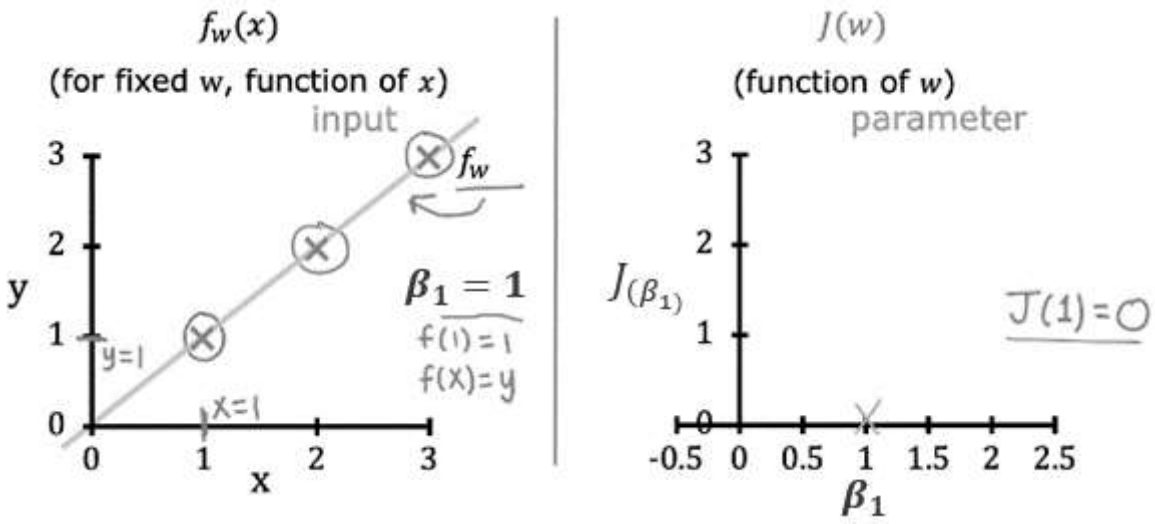
$$J_{(\beta_1)} \approx 0.58$$



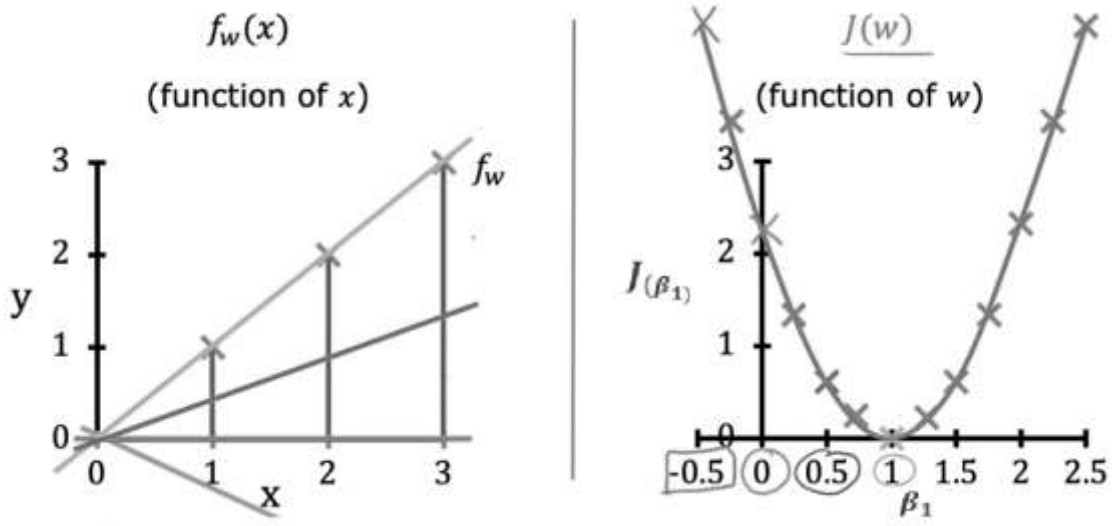
Let's fix the value of $\boldsymbol{\beta_1 = 1}$,

$$J_{(\beta_1)} = \frac{1}{2*3}\sum_{i=1}^{3}[(1*1-1)^2 + (1*2-2)^2 + (1*3-3)^2]$$

$$\boldsymbol{J_{(\beta_1)} = 0} \text{ (optimal)}$$



Here we plotted multiple different values of **x** using cost function in order to minimize loss.



We can observe that as the value of $\boldsymbol{\beta_1}$ is increasing or decreasing the cost function is increasing except when $\boldsymbol{\beta_1 = 1}$ at which cost is **0**.

GRADIENT DESCENT

Gradient descent is an optimization algorithm used to find the values of $\boldsymbol{\beta_0}$ & $\boldsymbol{\beta_1}$.

This algorithm is best used when $\boldsymbol{\beta_0}$ & $\boldsymbol{\beta_1}$ cannot be calculated manually (using cost function).

We can use gradient descent to minimize cost function.

Gradient descent works based on the value of **alpha** (a small value) that we assign, gradient descent takes baby steps according to alpha and tries to reach global minima.

Generalized formula,

$$\boldsymbol{\theta_{j:=}\ \theta_{j-}\ \ \frac{\alpha}{m}\sum_{i=1}^{m}[\,h_\theta(x^i) - y^i]x^i}$$

Formula with respect to $\boldsymbol{\beta_0}$ & $\boldsymbol{\beta_1}$,

$$\boldsymbol{\beta_1 = \beta_1 - \frac{\alpha}{m}\sum_{i=1}^{m}[\,f_{\beta_0,\beta_1}(x^i) - y^i]x^i}$$

$$\boldsymbol{\beta_0 = \beta_0 - \frac{\alpha}{m}\sum_{i=1}^{m}[\,f_{\beta_0,\beta_1}(x^i) - y^i]}$$