Linear Regression

Overview / Contents

- Regression Concepts
    - Types of Regression
- In-depth intuition of OLS
- Lose Functions
- Cost Function
- R Squared Values
- Coding with Python:
    - Implementing Linear Regression
    - Simple ML Project
    - Assignment

- **Regression in Machine Learning:**

Regression is a technique used to predict numerical values based on input features. It models the relationship between a dependent variable (what you want to predict) and independent variables (features).

- **Example: Predicting House Prices:**

Imagine you're predicting house prices based on square footage. The regression model finds a line that best fits the data: Price = 100 * SquareFootage + 50000. Here, 100 is the increase in price for each square foot increase, and $50,000 is the starting price estimate. This model helps estimate prices for different house sizes.

Regression Concepts

Examples

1. Economics: GDP Prediction:
Using historical data, economists can predict a country's future GDP based on factors like inflation rate, unemployment rate, and consumer spending.

2. Healthcare: Patient Outcome:
Doctors can predict a patient's recovery time after surgery based on variables like age, pre-existing conditions, and the complexity of the procedure.

3. Retail: Sales Forecasting:
Retailers can use regression to forecast sales based on parameters like advertising spend, holiday season, and previous sales data.

4. Finance: Stock Price Prediction:
Traders and investors can predict stock prices by analyzing factors like trading volume, historical prices, and economic indicators.

5. Agriculture: Crop Yield Estimation:
Regression helps farmers predict crop yields based on factors like weather conditions, soil quality, and type of crop.

Regression Concepts

Examples

6. Marketing: Customer Lifetime Value:
Marketers use regression to estimate a customer's lifetime value based on purchase history, engagement, and demographic information.

7. Education: Student Performance:
Educators can predict student performance on standardized tests using factors like attendance, study time, and past test scores.

8. Energy: Energy Consumption:
Energy companies can predict household energy consumption based on variables like weather, household size, and appliance usage.

9. Transportation: Fuel Efficiency:
Manufacturers predict a vehicle's fuel efficiency based on engine specifications, weight, and aerodynamics.

10. Real Estate: Property Valuation:
Regression helps in estimating property values based on features like location, square footage, and nearby amenities.

Types of Regression?

There are several types of regression techniques, each designed to handle different types of data and relationships between variables. Here are some common types of regression:

1. Linear Regression:
   - Simple Linear Regression: Predicting a continuous dependent variable using a single independent variable.
   - Multiple Linear Regression: Predicting a dependent variable using multiple independent variables.

2. Polynomial Regression:
   - Modeling nonlinear relationships by adding polynomial terms to the regression equation.

3. Ridge Regression:
   - Adding a penalty term to the coefficients to prevent overfitting.

4. Lasso Regression:
   - Similar to ridge regression, but with a penalty that encourages some coefficients to become exactly zero, leading to feature selection.

5. Elastic Net Regression:
   - A combination of ridge and lasso regression, providing a balance between their strengths.

Types of Regression

Common types

6. Logistic Regression:
   - Used for binary or multinomial classification tasks, predicting the probability of an event occurring.

7. Poisson Regression:
   - Modeling count data, often used in situations where the dependent variable represents counts.

8. Time Series Regression:
   - Modeling time-dependent data, considering temporal patterns and autocorrelation.

9. Nonlinear Regression:
   - Fitting a nonlinear function to the data to capture complex relationships.

10. Quantile Regression:
    - Modeling different quantiles of the dependent variable, useful for understanding conditional distributions.

11. Support Vector Regression (SVR):
    - Utilizes support vector machines for regression tasks, particularly suited for high-dimensional spaces.

Types of Regression

Common types

12. Bayesian Regression:
    - Incorporates Bayesian statistics to estimate parameters and uncertainties in regression models.

13. Kernel Regression:
    - Uses kernel functions to capture complex patterns in the data.

14. Generalized Linear Models (GLM):
    - Generalization of linear regression for various types of dependent variables, including binary and count data.

15. Stepwise Regression:
    - An automated method for selecting a subset of important features.

16. Piecewise Regression:
    - Fits different regression models to different segments of the data, useful for data with changing trends.

17. Principal Component Regression (PCR):
    - Combines principal component analysis (PCA) and linear regression.

All you need to know about Linear Regression!

- Linear regression is a fundamental supervised machine learning algorithm used for predicting a continuous numerical value (also known as the dependent variable) based on one or more input features (independent variables).
- It models the relationship between the dependent variable and the independent variables as a linear equation.
- The goal is to find the best-fitting line (or hyperplane in higher dimensions) that minimizes the difference between the observed and predicted values.
- This best-fitting line represents the linear relationship between the input features and the target variable.

General Equation for Linear Regression:

Y = M*X + C

Y = M1*X1 + M2*X2 + ………..Mn*Xn + C

Here,

M = Coefficient of the input feature X

C = Intercept

X = Features

Y = Predicted Output / Label

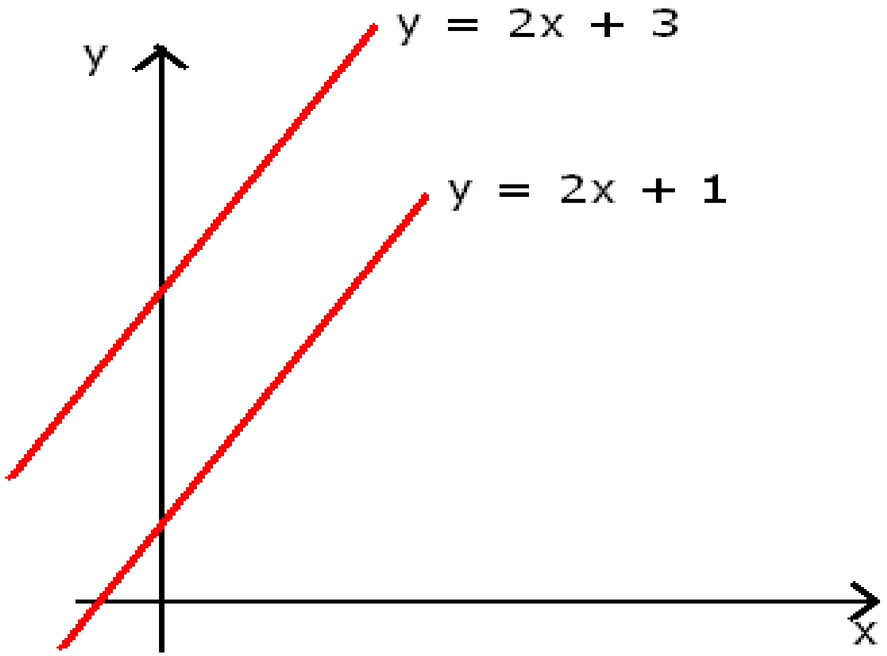


Fig: Straight Line

X = 10, 30, 50
Y = 2*10 + 3 =23
Y = 2*30 + 3 = 63
Y = 2*50 + 3 = 103

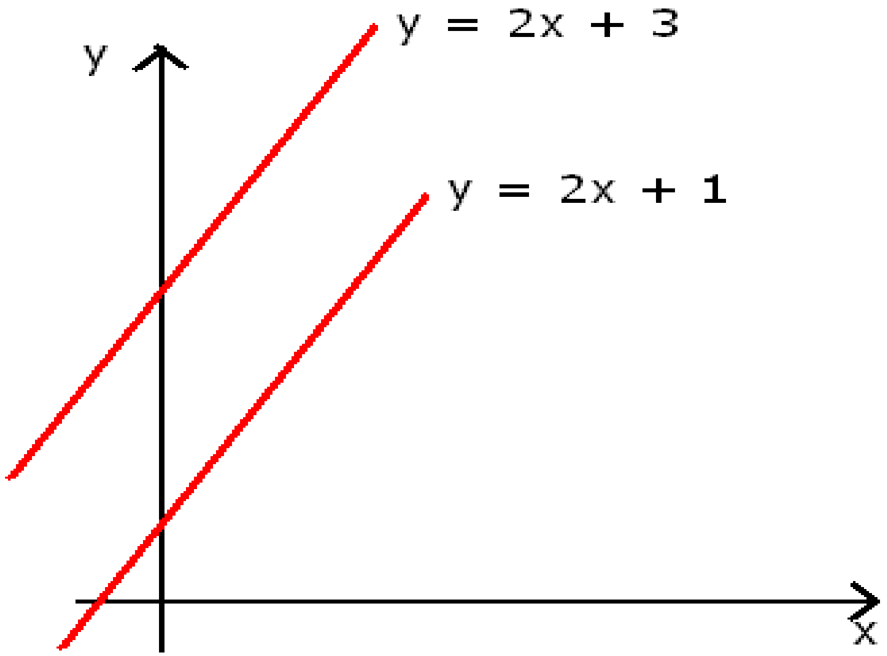


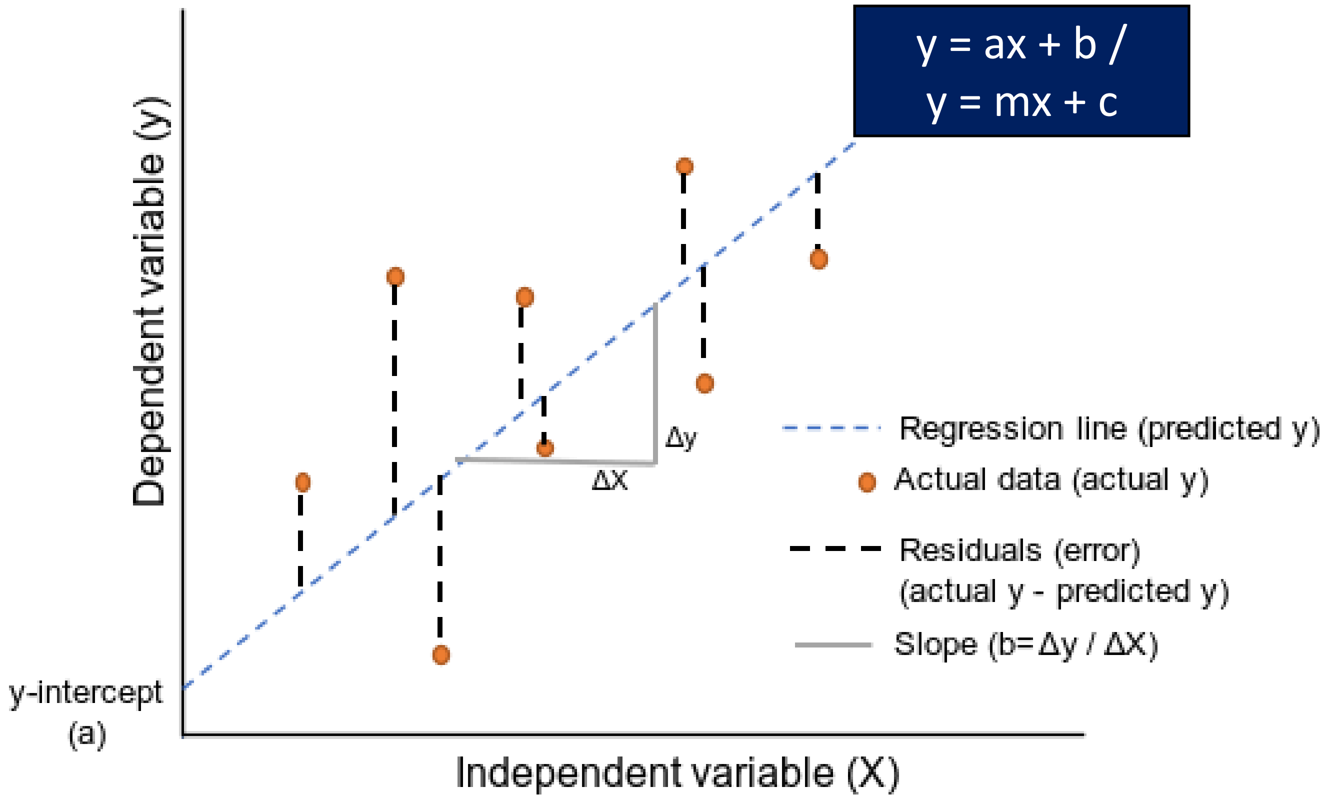
Fig: Straight Line

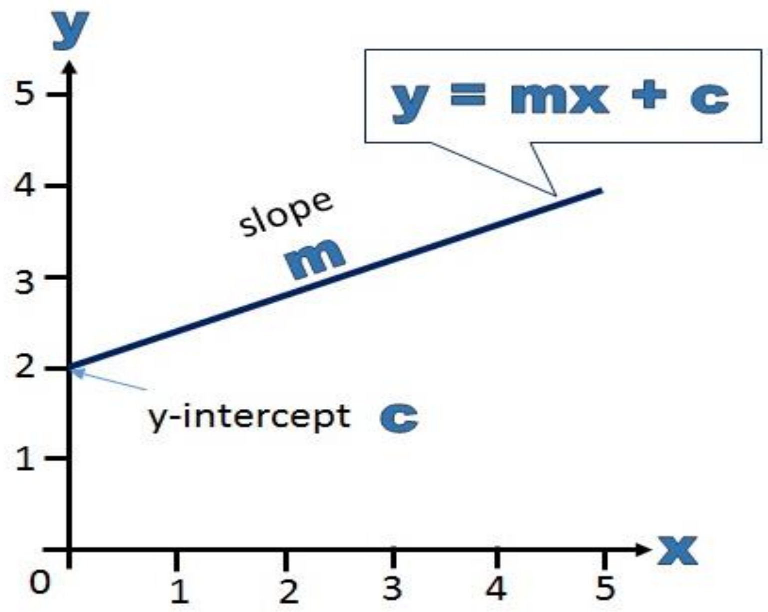
X = 10, 30, 50
Y = 2*10 + 3 =23
Y = 2*30 + 3 = 63
Y = 2*50 + 3 = 103

X	Actual	Predicted
10	25	23
30	60	63
50	100	103

Linear Regression

Basic Concepts

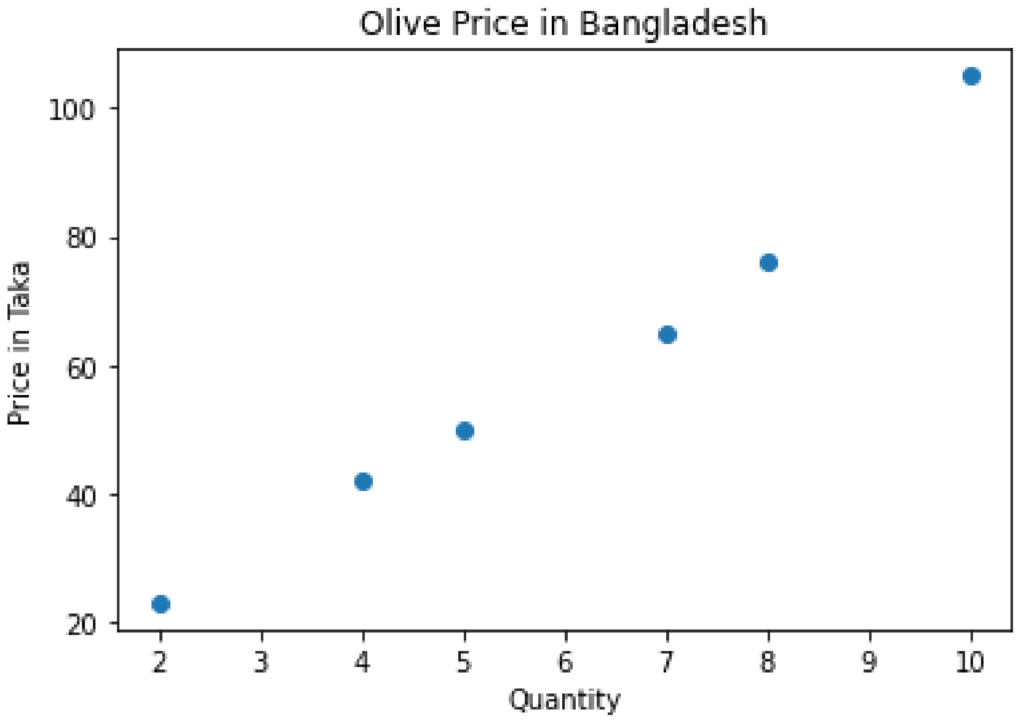




- **Residual** = Observed Value - Predicted Value

Observed	Predicted	Residual
25	23	2
60	63	3
100	103	3

	x	y
0	5	50
1	7	65
2	4	42
3	8	76
4	2	23
5	10	105



	x	y
0	5	50
1	7	65
2	4	42
3	8	76
4	2	23
5	10	105

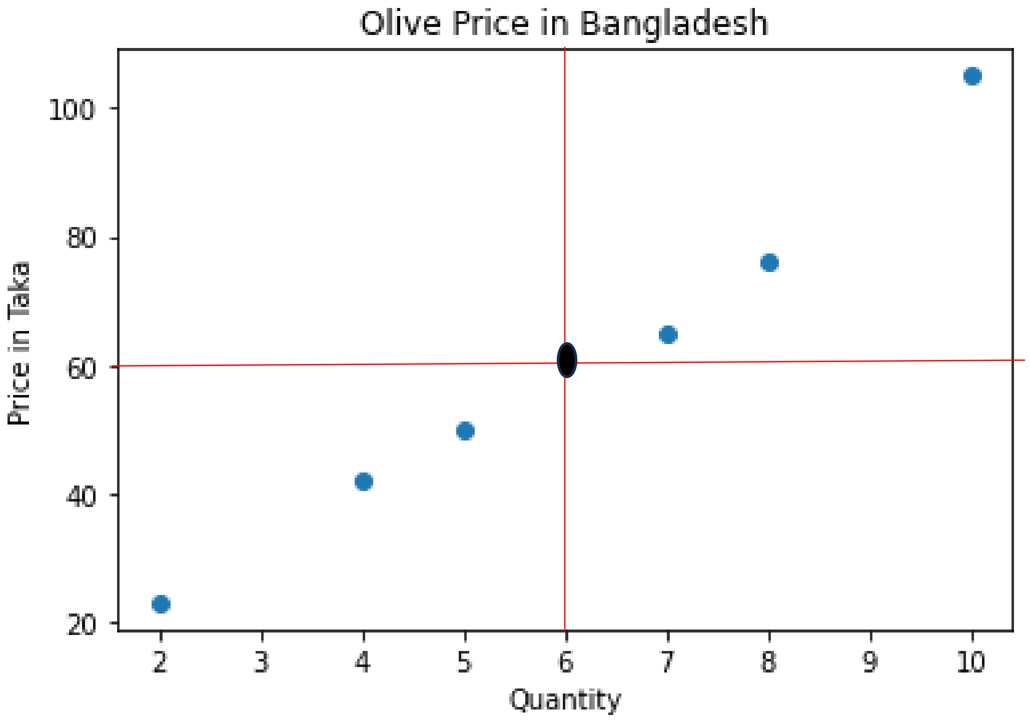
Mean Values

```
df.x.mean()
```

```
6.0
```

```
df.y.mean()
```

```
60.166666666666664
```



Formula of Linear Regression

$$Y = MX + C$$

$$C = \bar{Y} - M\bar{X}$$

$$M = \frac{\bar{X}\cdot\bar{Y} - \overline{XY}}{(\bar{X})^2 - \overline{X^2}}$$

$\bar{X}$ = Mean X

$\bar{Y}$ = Mean Y

Now Solve it

Data Set

	A	B
1	x	y
2	5	50
3	7	65
4	4	42
5	8	76
6	2	23
7	10	105
8	7	?

Calculation Table for Single Variable Linear Regression

	A	B	C	D	E	F	G	H	I
1	x	y	xy	x2	x̄	ȳ	(xy) bar	(x̄)2	(x2) bar
2	5	50	250	25					
3	7	65	455	49	Sum=36	Sum=361	Sum=2577		Sum=258
4	4	42	168	16	36/6	361/6	2577/6		258/6
5	8	76	608	64					
6	2	23	46	4	Avg=6	Avg=60.17	Avg=429.5	36	Avg=43
7	10	105	1050	100	Average	Average	Average		Average

Linear Regression

Raw Calculation

Formula of Linear Regression

$$Y = MX + C$$

$$C = \bar{Y} - M\bar{X}$$

$$M = \frac{\bar{X}\cdot\bar{Y} - \overline{XY}}{(\bar{X})^2 - \overline{X^2}}$$

$\bar{X}$ = Mean X

$\bar{Y}$ = Mean Y

Final Calculations
M = ((6*60.17)-429.5) / (36-43)
M = 9.782
C = 60.17-(9.782*6)
C = 1.48
Y = (9.782 * X) + 1.48
Predict, y = (9.782*7)+1.48
Ans = 69.95

	A	B	C	D	E	F	G	H	I	J
1	x	y	xy	x2	x̄	ȳ	(xy) bar	(x̄)2	(x2) bar	Final Calculations
2	5	50	250	25						M = ((6*60.17)-429.5) / (36-43)
3	7	65	455	49	Sum=36	Sum=361	Sum=2577		Sum=258	M = 9.782
4	4	42	168	16	36/6	361/6	2577/6		258/6	C = 60.17-(9.782*6)
5	8	76	608	64						C = 1.48
6	2	23	46	4	Avg=6	Avg=60.17	Avg=429.5	36	Avg=43	Y = (9.782 * X) + 1.48
7	10	105	1050	100	Average	Average	Average		Average	Predict, y = (9.782*7)+1.48
8	7	69.95		49						Ans = 69.95

Slope, $m = \Sigma((x - \bar{x}) * (y - \bar{y})) / \Sigma((x - \bar{x})^2)$

Intercept, $c = \bar{y} - m * \bar{x}$

Where:

x is a data point on the independent variable (x-axis).

y is the corresponding dependent variable (y-axis).

$\bar{x}$ is the mean of the independent variable.

$\bar{y}$ is the mean of the dependent variable.

Data Set

	x	y
0	5	50
1	7	65
2	4	42
3	8	76
4	2	23
5	10	105

Value of M & C

```
reg.coef_
```

array([9.78571429])

```
reg.intercept_
```

1.4523809523809703

Linear Regression

Raw Calculation & Visual Prediction

Data Set

	x	y
0	5	50
1	7	65
2	4	42
3	8	76
4	2	23
5	10	105

Value of M & C

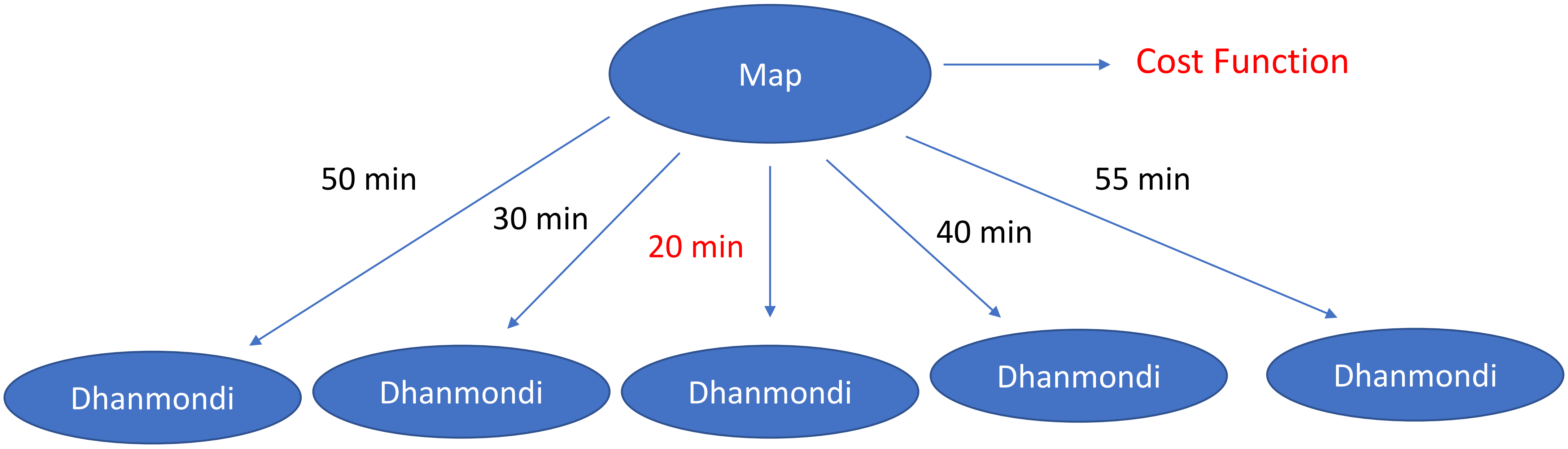
```
reg.coef_
```

```
array([9.78571429])
```

```
reg.intercept_
```

```
1.4523809523809703
```

The cost function is a function, which is associates a cost with a **decision.**



Reseduals

- Residuals are the differences between the observed values of the dependent variable and the predicted values generated by the regression model.
- They are calculated as (Yi - Ypred), where Yi is the observed value and Ypred is the predicted value.
- Residuals are used to assess the fit of a regression model and to diagnose potential issues like underfitting, overfitting, or the presence of outliers.

- **L1, L2 loss,** and **residuals** are related concepts, both involving differences between predicted and actual values in regression analysis.
- Loss is a measure of the differences, while residuals are the actual differences themselves.

- **However, loss specifically refers to a loss function used for optimization purposes, while residuals are used for model assessment and diagnosis.**

Residual = Observed Value - Predicted Value

Observed	Predicted	Residual
25	23	2
60	63	3
100	103	3

L1 Loss (Absolute Loss or Mean Absolute Error):

- L1 loss is a type of loss function used to measure the difference between predicted values and actual observed values in regression problems.
- It calculates the absolute difference between the predicted value and the actual value for each data point and then averages these absolute differences.
- Mathematically, the L1 loss for the ith data point is (|Yi - Ypred|), where Yi is the observed value and Ypred is the predicted value.
- L1 loss tends to be less sensitive to outliers compared to squared loss (L2 loss).

L2 Loss (Squared Loss or Mean Squared Error):

- - L2 loss measures the squared difference between predicted values and actual observed values in regression problems.
- - It calculates the squared difference between the predicted value and the actual value for each data point and then averages these squared differences.
- - Mathematically, the L2 loss for the ith data point is (Yi - Ypred)^2, where Yi is the observed value and Ypred is the predicted value.
- - L2 loss penalizes larger errors more heavily due to the squaring operation.

Mean Absolute Error:

MAE = $\frac{1}{n}\sum_{i=1}^{n} | y_i - \widehat{y}_i |$

Mean Squared Error :

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

1. Loss (or Error) for a Single Sample:

- When you calculate the difference between the actual value and the predicted value for a single data point, it's generally referred to as a "loss" or "error" for that specific data point.
- This term is used to describe the discrepancy between the prediction and the true value for a single instance.

2. Cost (or Loss) for the Entire Dataset:

- When you calculate the average or total of these losses/errors across the entire dataset, it's often referred to as the "cost" or "loss" for the dataset.
- The term "cost" or "loss" is used to describe the overall quality of the model's predictions for the entire dataset.

Thanks for your patience!

www.aiquest.org