به نام خدا

# کارگاه علم داده با پایتون پیشرفته

جلسه سوم: رگرسیون خطی ساده و رگرسیون چندگانه


مدرس :

مهرناز جلیلی

دانشجو کارشناسی ارشد علم داده ها

دانشگاه شهید بهشتی


# Regression Intro

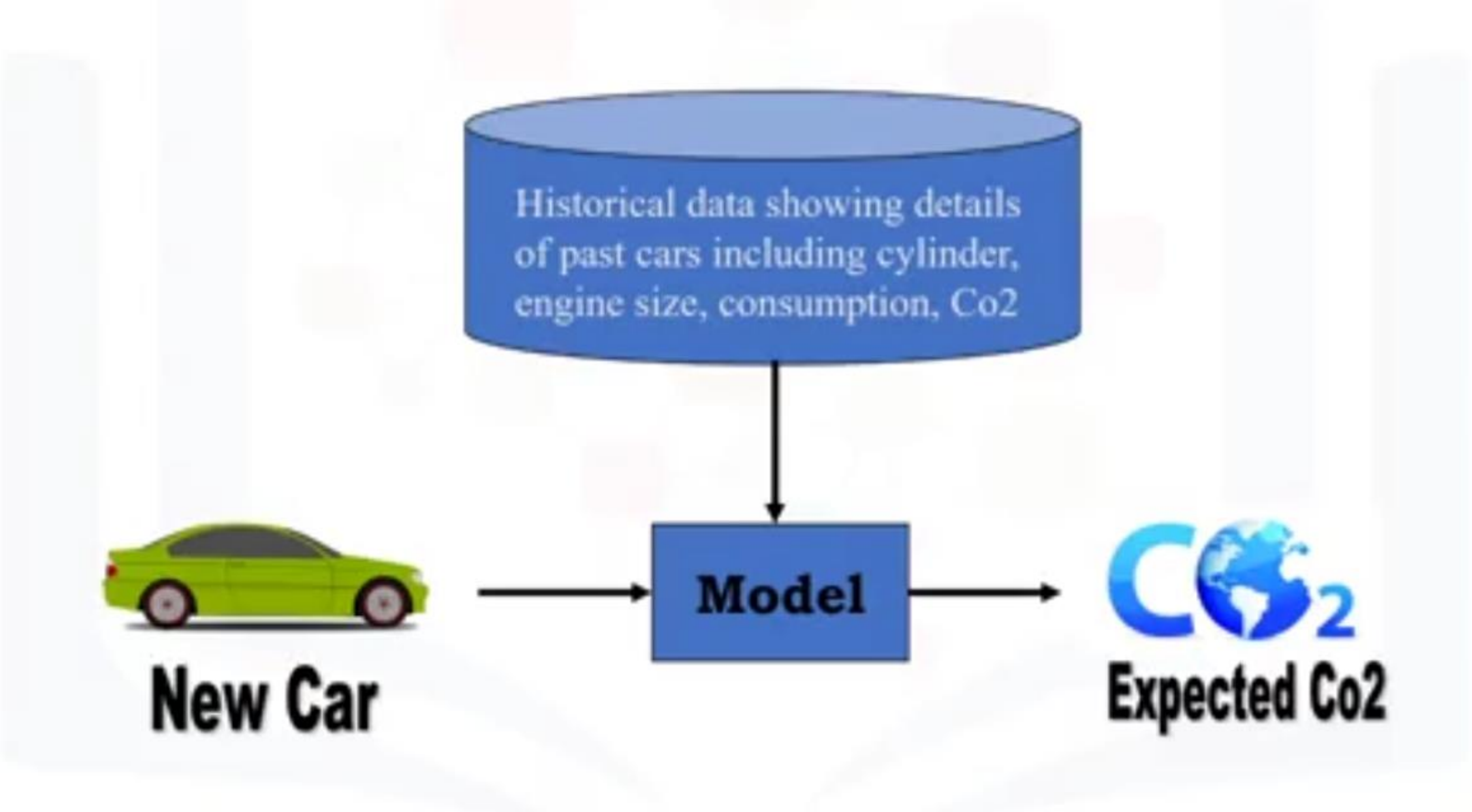
- regression is the process of predicting a continuous value
- Independent (x, desc, ...) vs Dependent (y, goal, prediction, ...) variables
- y is continuous

[5]:

| | ENGINESIZE | CYLINDERS | FUELCONSUMPTION_COMB | CO2EMISSIONS |
|---|---|---|---|---|
| **0** | 2.0 | 4 | 8.5 | 196 |
| **1** | 2.4 | 4 | 9.6 | 221 |
| **2** | 1.5 | 4 | 5.9 | 136 |
| **3** | 3.5 | 6 | 11.1 | 255 |
| **4** | 3.5 | 6 | 10.6 | 244 |
| **5** | 3.5 | 6 | 10.0 | 230 |
| **6** | 3.5 | 6 | 10.1 | 232 |
| **7** | 3.7 | 6 | 11.1 | 255 |
| **8** | 3.7 | 6 | 11.6 | 267 |

# Regression Intro

## Model



# Regression Intro

## Types

- Simple (only one independent)
  - Linear
  - Non-Linear
- Multiple (multiple independent)
  - Linear
  - Non-Linear

# Regression Intro

## Samples

- Household Price
- Customer Satisfaction
- Sales Forecast
- Employment Income

# Regression Intro

## Algorithms

---

- Ordinal
- Poisson
- Fast Forest quantile
- Linear, Polynominal, Lasso, Stepwise, Ridge
- Bayesian Linear
- Nerural Network
- Decision Forest
- Boosted decision tree
- K-nearest neighbors

# Simple Linear Regression

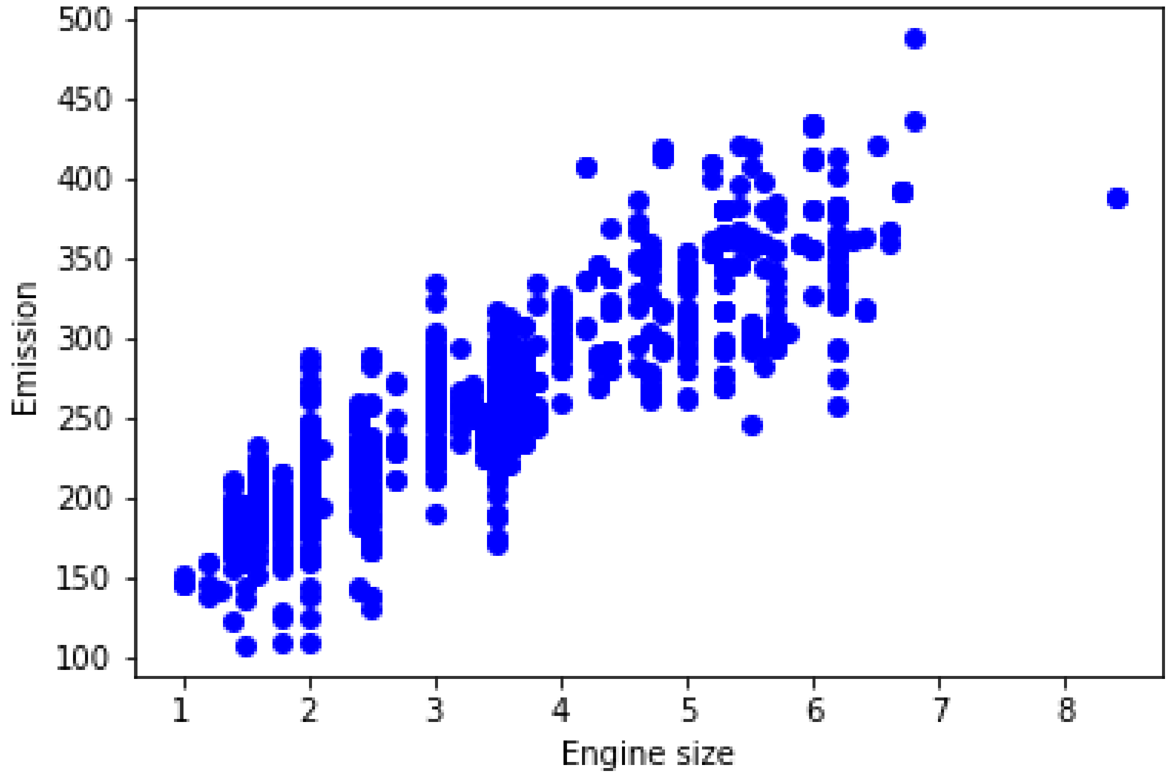
- Can we predict co2 emission from one of the independents (this is why we call it Simple)
- Lets try engine size...

[5]:

| | ENGINESIZE | CYLINDERS | FUELCONSUMPTION_COMB | CO2EMISSIONS |
|---|---|---|---|---|
| **0** | 2.0 | 4 | 8.5 | 196 |
| **1** | 2.4 | 4 | 9.6 | 221 |
| **2** | 1.5 | 4 | 5.9 | 136 |
| **3** | 3.5 | 6 | 11.1 | 255 |
| **4** | 3.5 | 6 | 10.6 | 244 |
| **5** | 3.5 | 6 | 10.0 | 230 |
| **6** | 3.5 | 6 | 10.1 | 232 |
| **7** | 3.7 | 6 | 11.1 | 255 |
| **8** | 3.7 | 6 | 11.6 | 267 |

# Simple Linear Regression

- Relationship is obvious $\hat{y} = \theta_0 + \theta_1 x_1$
- There is a line, we assume a straight line
- we can predict an emission for say, a car with 2.4
- y hat is the dependent variable of the predicted value.
- x1 is the independent variable.
- Theta 0 and theta 1 are the parameters of the line
- Theta 1 is known as the slope or gradient of the fitting line and theta 0 is known as the intercept.
- Theta 0 and theta 1 are also called the coefficients of the linear equation.

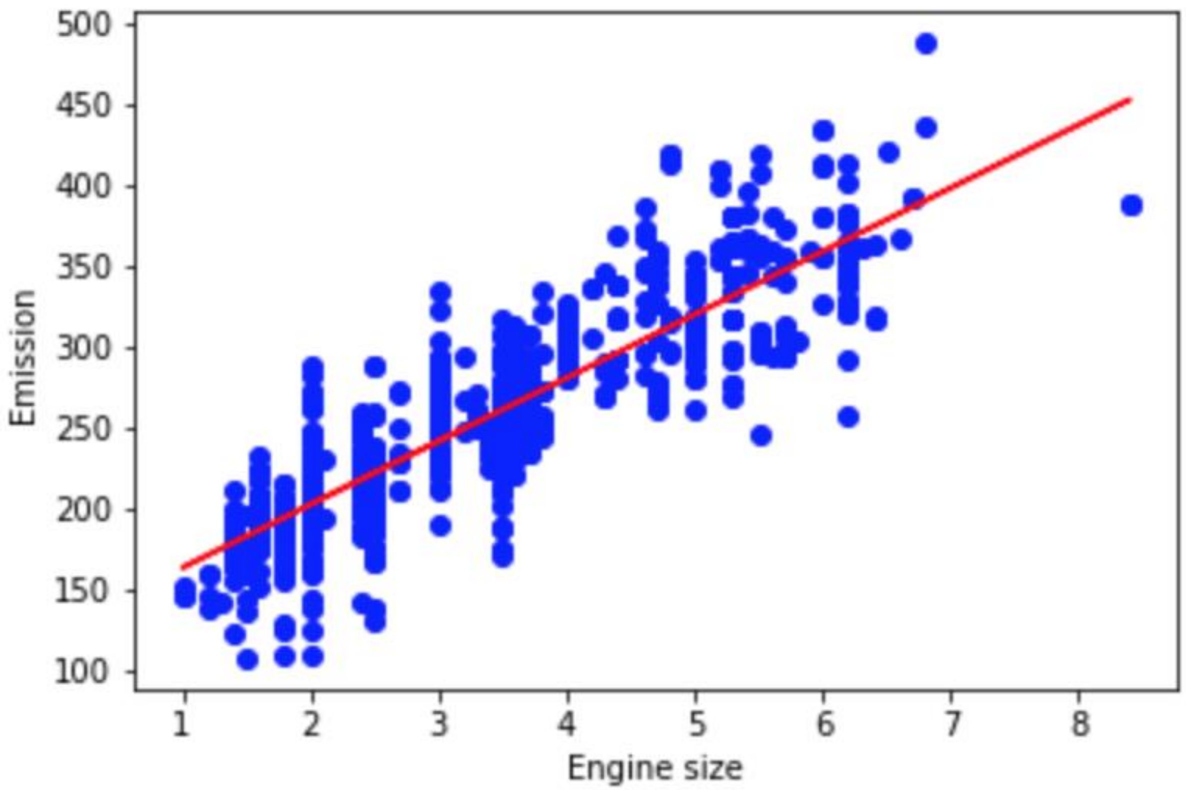


# Simple Linear Regression

## MSE

- Residual error for each point is the distance of the prediction from the actual point. So Mean Square Error (MSE should be minimized)

- Minimum MSE can be achieved with two methods: Math or Optimization




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

# Simple Linear Regression

## MSE (Math)

| | ENGINESIZE | CYLINDERS | FUELCONSUMPTION_COMB | CO2EMISSIONS |
|---|---|---|---|---|
| 0 | 2.0 | 4 | 8.5 | 196 |
| 1 | 2.4 | 4 | 9.6 | 221 |
| 2 | 1.5 | 4 | 5.9 | 136 |
| 3 | 3.5 | 6 | 11.1 | 255 |
| 4 | 3.5 | 6 | 10.6 | 244 |
| 5 | 3.5 | 6 | 10.0 | 230 |
| 6 | 3.5 | 6 | 10.1 | 232 |
| 7 | 3.7 | 6 | 11.1 | 255 |
| 8 | 3.7 | 6 | 11.6 | 267 |

$X_1$

$y$

$$\hat{y} = \theta_0 + \theta_1 x_1$$

$$\theta_1 = \frac{\sum_{i=1}^{s}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{s}(x_i - \bar{x})^2}$$

$$\bar{x} = (2.0 + 2.4 + 1.5 + \dots)/9 = 3.03$$

$$\bar{y} = (196 + 221 + 136 + \dots)/9 = 226.22$$

$$\theta_1 = \frac{(2.0 - 3.03)(196 - 226.22) + (2.4 - 3.03)(221 - 226.22) + \dots}{(2.0 - 3.03)^2 + (2.4 - 3.03)^2 + \dots}$$

$$\theta_1 = 39$$

$$\theta_0 = \bar{y} - \theta_1 \bar{x}$$

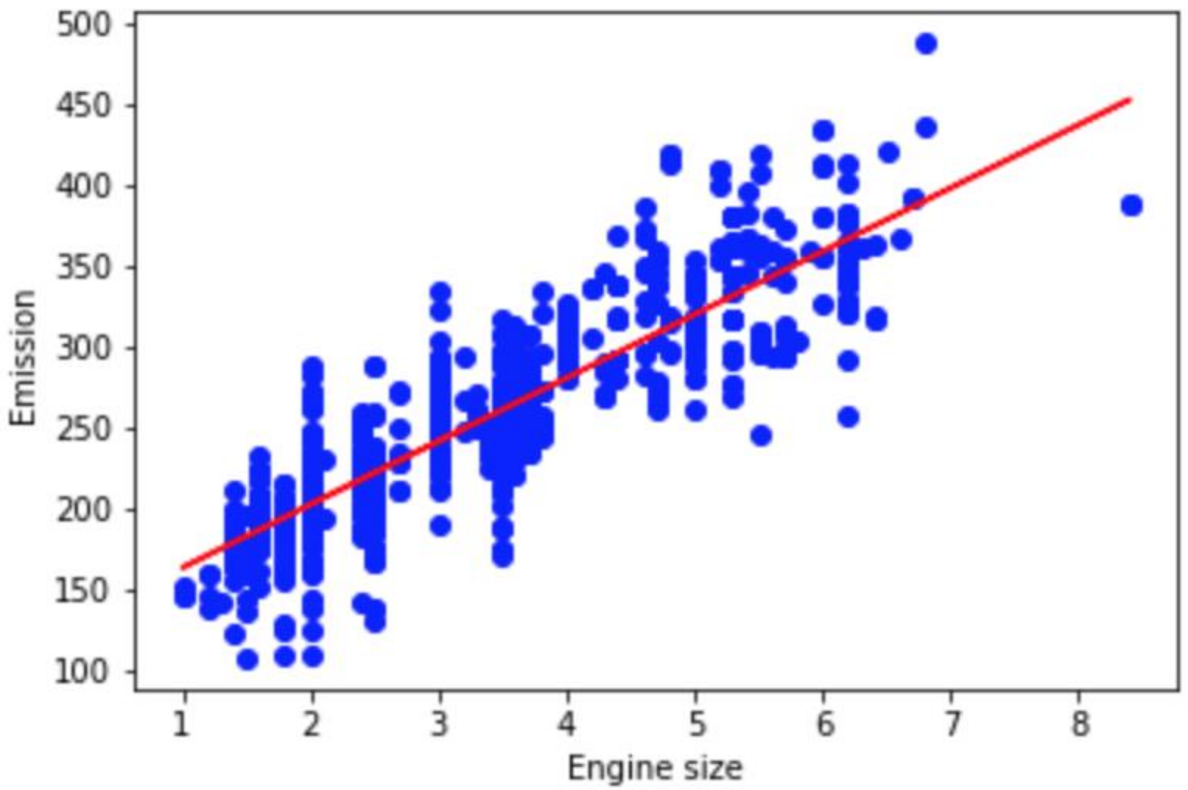
$$\theta_0 = 226.22 - 39 * 3.03$$

$$\theta_0 = 125.74$$

# Simple Linear Regression

## Pros

- Very Fast
- Easy to understand and interpret
- No need for parameter tuning
  (say like in KNN)



# Model Evaluation

- goal is to build a model to accurately predict an unknown case.
- You need to evaluate to see how much you can trust your model/prediction
- Two main methods:
  - Train and Test on Same data
  - Train / Test split
- Regression Evaluation Metrics

# Model Evaluation

## Train and Test on Same data

- High "training accuracy"
  - not always good
  - overfitting the data (say capture noise and produce non generalized model)
- Low "out of sample accuracy"
  - Important to have

| | ENGINESIZE | CYLINDERS | FUELCONSUMPTION_COMB | CO2EMISSIONS |
|---|---|---|---|---|
| 0 | 2.0 | 4 | 8.5 | 196 |
| 1 | 2.4 | 4 | 9.6 | 221 |
| 2 | 1.5 | 4 | 5.9 | 136 |
| 3 | 3.5 | 6 | 11.1 | 255 |
| 4 | 3.5 | 6 | 10.6 | 244 |
| 5 | 3.5 | 6 | 10.0 | 230 |
| 6 | 3.5 | 6 | 10.1 | 232 |
| 7 | 3.7 | 6 | 11.1 | 255 |
| 8 | 3.7 | 6 | 11.6 | 267 |
| 9 | 2.4 | 4 | 9.2 | 212 |

Test

$y$

Actual values

$$Error = \frac{(232-234)+(255-256)+\ \dots}{4}$$

$$Error = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

$\hat{y}$

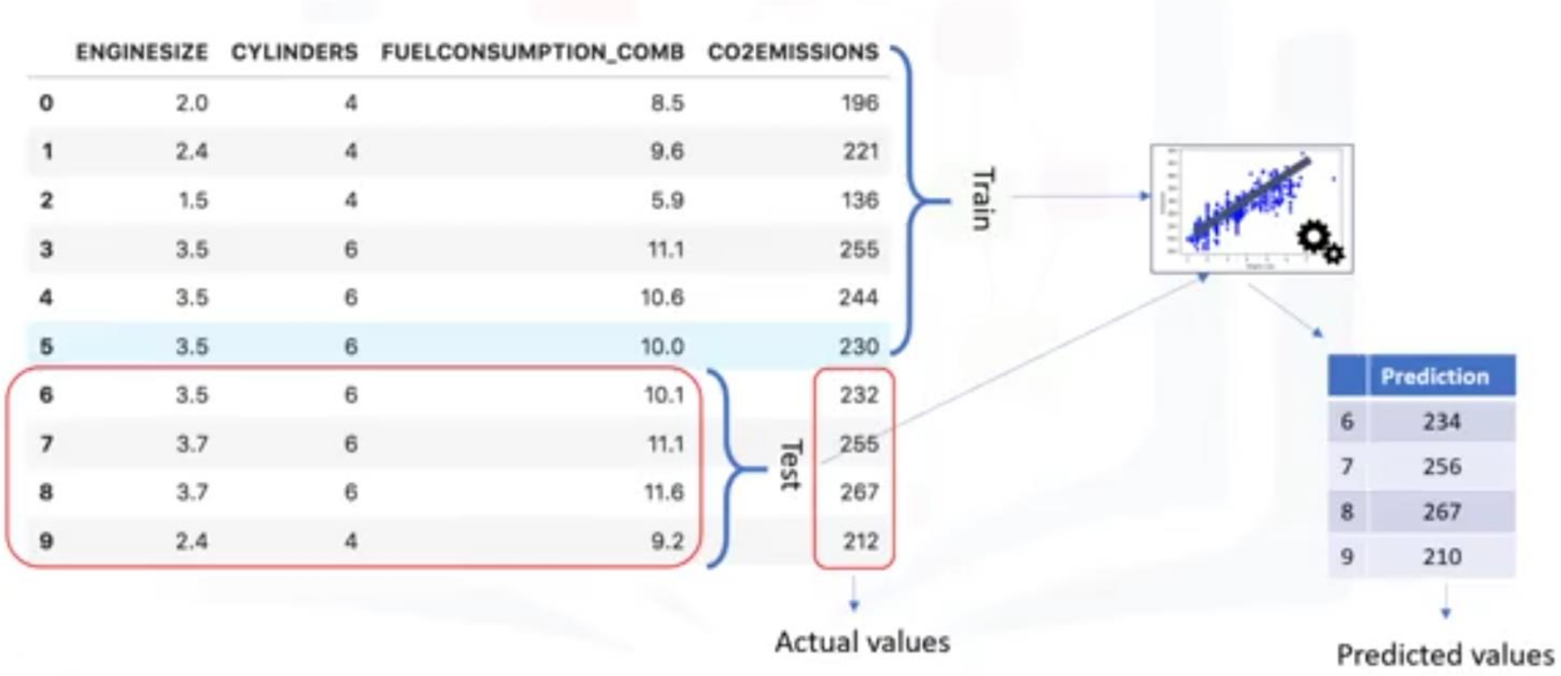
| | Prediction |
|---|---|
| 6 | 234 |
| 7 | 256 |
| 8 | 267 |
| 9 | 210 |

Predicted values

# Model Evaluation

## Train/Test split

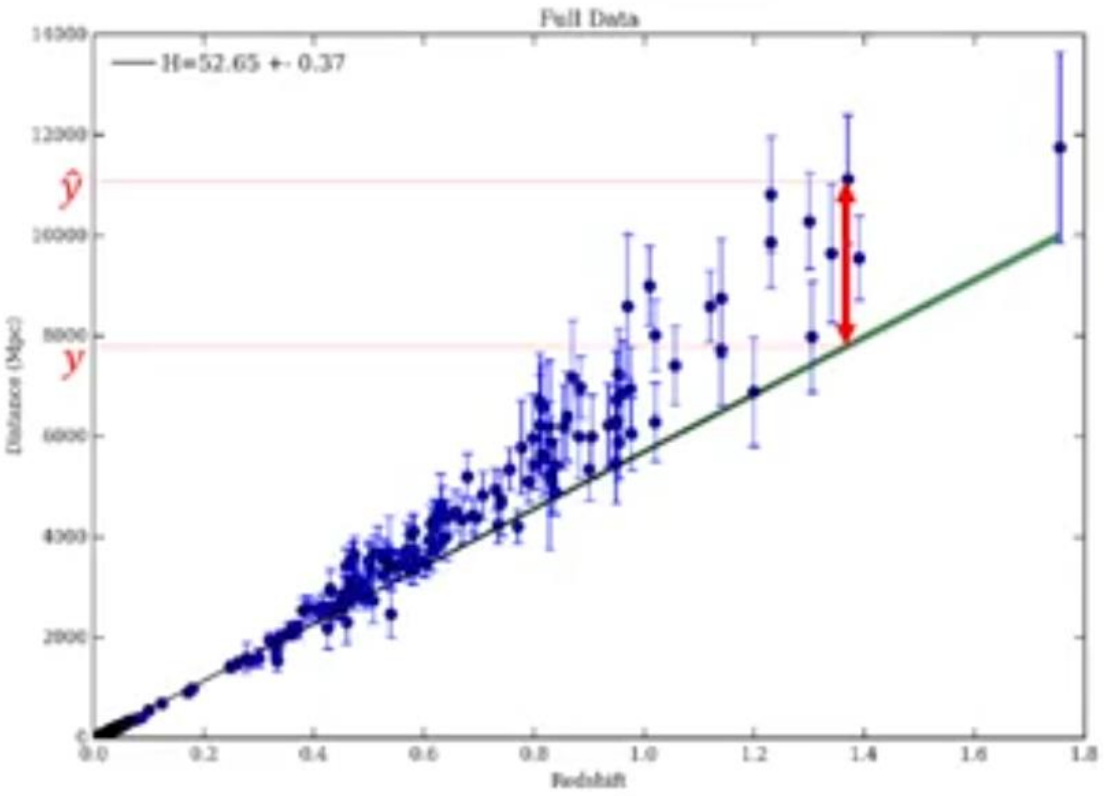
- Mutually exclusive split
- More accurate on out-of-sample
- ensure that you train your model with the testing set afterwards, as you don't want to lose potentially valuable data.
- Dependent on which datasets the data is trained and tested



# Model Evaluation

## Evaluation Matrix

- used to explain the performance of a model
- say comparing actual with predicted
- error of the model is the difference between the data points and the trend line generated by the algorithm
- There different metrics (next slide) but the choice is based on the model, data type, domain, ...



# Model Evaluation

## Errors

- mean absolute error (MAE)
- mean squared error (MSE)
- root mean squared error (RMSE); interpretable in the same units as the response vector or y units
- Relative absolute error, also known as residual sum of square (RAE)
- Relative squared error  (RSE)
- R2; Popular metric for the accuracy of your model. represents how close the data values are to the fitted regression line. The higher the better

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

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

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

$$RAE = \frac{\sum_{j=1}^{n}|y_j - \hat{y}_j|}{\sum_{j=1}^{n}|y_j - \bar{y}|}$$

$$RSE = \frac{\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{\sum_{j=1}^{n}(y_j - \bar{y})^2}$$

$$\boxed{R^2 = 1 - RSE}$$

# Lab: Simple Linear Regression

- L03-MehrnazJalili.ipynb

# Multiple Linear Regression

- Simple / Multiple
- kind of same as simple
- usages:
  - find the strength of each independent variable
  - predict the impact of the change on one of the independent variables

# Multiple Linear Regression

## Formula

$$Co2\ Em = \theta_0 + \theta_1 Engine\ size + \theta_2 Cylinders + \ \dots$$

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ \dots + \theta_n x_n$$

$$\hat{y} = \theta^T X$$

$$\theta^T = [\theta_0, \theta_1, \theta_2, \dots] \qquad X = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \dots \end{bmatrix}$$

| | $\overbrace{\text{ENGINESIZE}}$ | $\overbrace{\text{CYLINDERS}}^{\text{X: Independent variable}}$ | $\overbrace{\text{FUELCONSUMPTION\_COMB}}$ | $\overbrace{\text{CO2EMISSIONS}}^{\text{Y: Dependent variable}}$ |
|---|---|---|---|---|
| 0 | 2.0 | 4 | 8.5 | 196 |
| 1 | 2.4 | 4 | 9.6 | 221 |
| 2 | 1.5 | 4 | 5.9 | 136 |
| 3 | 3.5 | 6 | 11.1 | 255 |
| 4 | 3.5 | 6 | 10.6 | 244 |
| 5 | 3.5 | 6 | 10.0 | 230 |
| 6 | 3.5 | 6 | 10.1 | 232 |
| 7 | 3.7 | 6 | 11.1 | 255 |
| 8 | 3.7 | 6 | 11.6 | 267 |
| 9 | 2.4 | 4 | 9.2 | ? |

# Multiple Linear Regression

## Finding parameters

- Again we can find the MSE
- the best model is the one with he minimized MSE
- The method is called Ordinary Least square
  - linear algebra
  - slow! for less than 10K samples
- Optimization Algorithms
  - Gradient Descent (Starts with random, then changes in multiple iterations)

# Multiple Linear Regression

## Some notes

- Try to have theoretical defense when choosing the independent variables. too many Xs might result in over fitting
- Xs do not need to be continues. If they are not try to assign values (like 1 and 2) to categories
- there needs to be a linear relationship. Test your Xs with scatter plots or use your logic. If the relationship displayed in your scatter plot is not linear, then you need to use non-linear regression.

# Lab: Multiple Linear Regression

- L03-MehrnazJalili.ipynb