Cardiff Metropolitan University | Prifysgol Metropolitan Caerdydd

Correlation & Regression Workshop 2: CIS5026

Module leader: Dr Angesh Anupam
Email: AAnupam@cardiffmet.ac.uk

Session overview

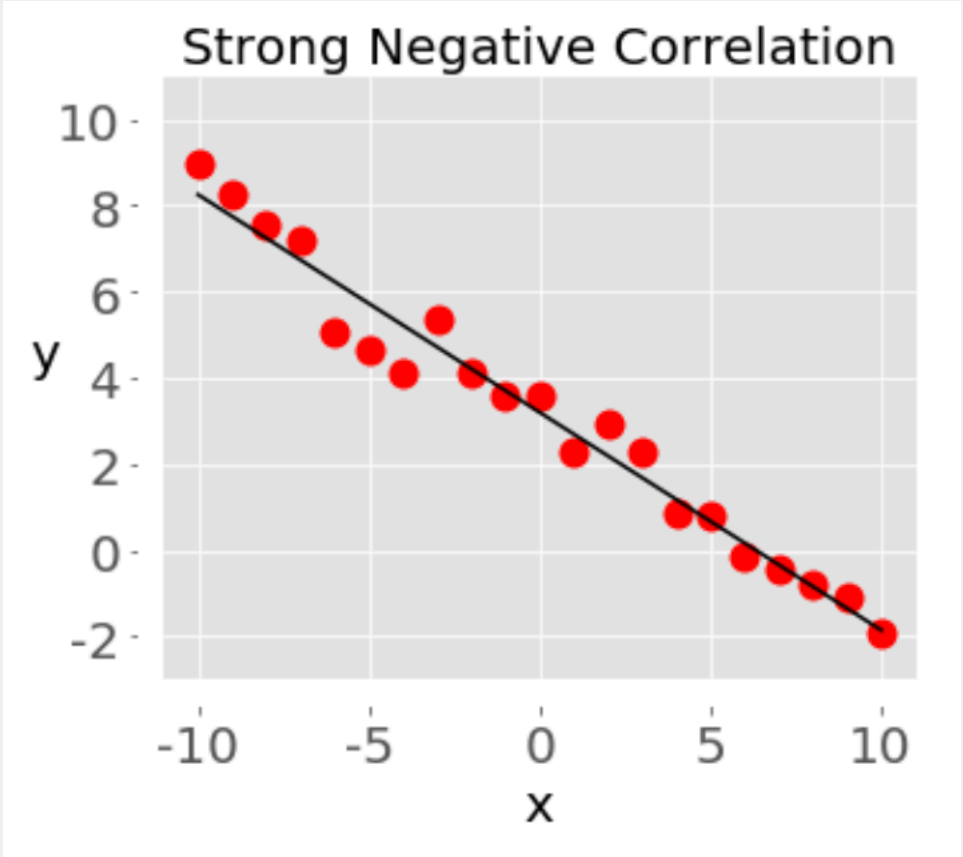
- Correlation
- Simple regression
- Advanced regression
- Assignment

Correlation

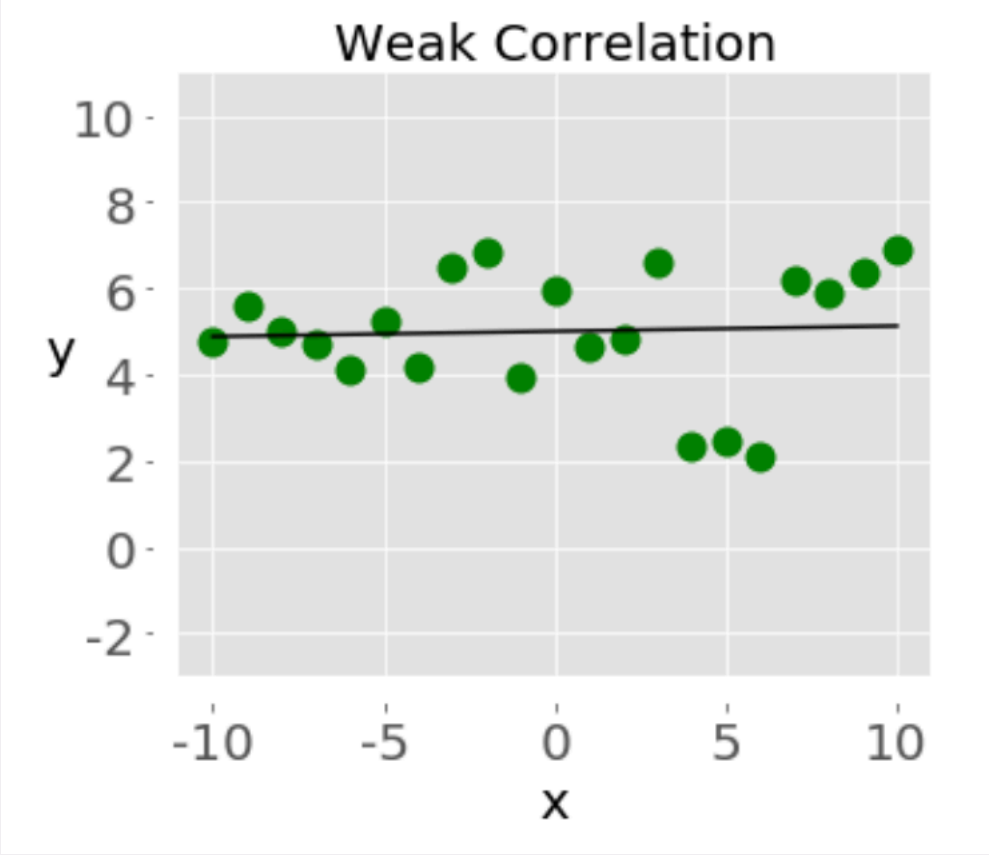
Introduction

- In data science we are often concerned about the relationships between two or more variables (or features) of a dataset.
- Each data point in the dataset is an observation, and the features are the properties or attributes of those observations.
- Example question – What mathematical function exists between population density and the GDP per capita of different countries?

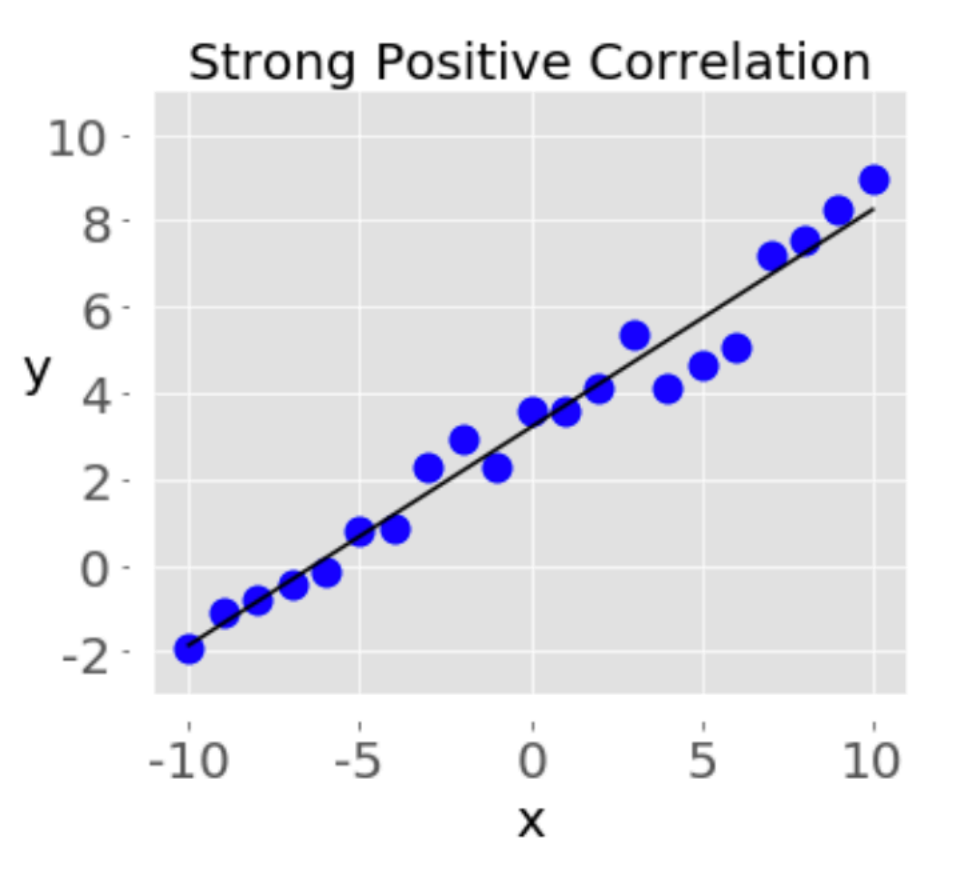
Strong negative correlation



Weak correlation



Strong positive correlation



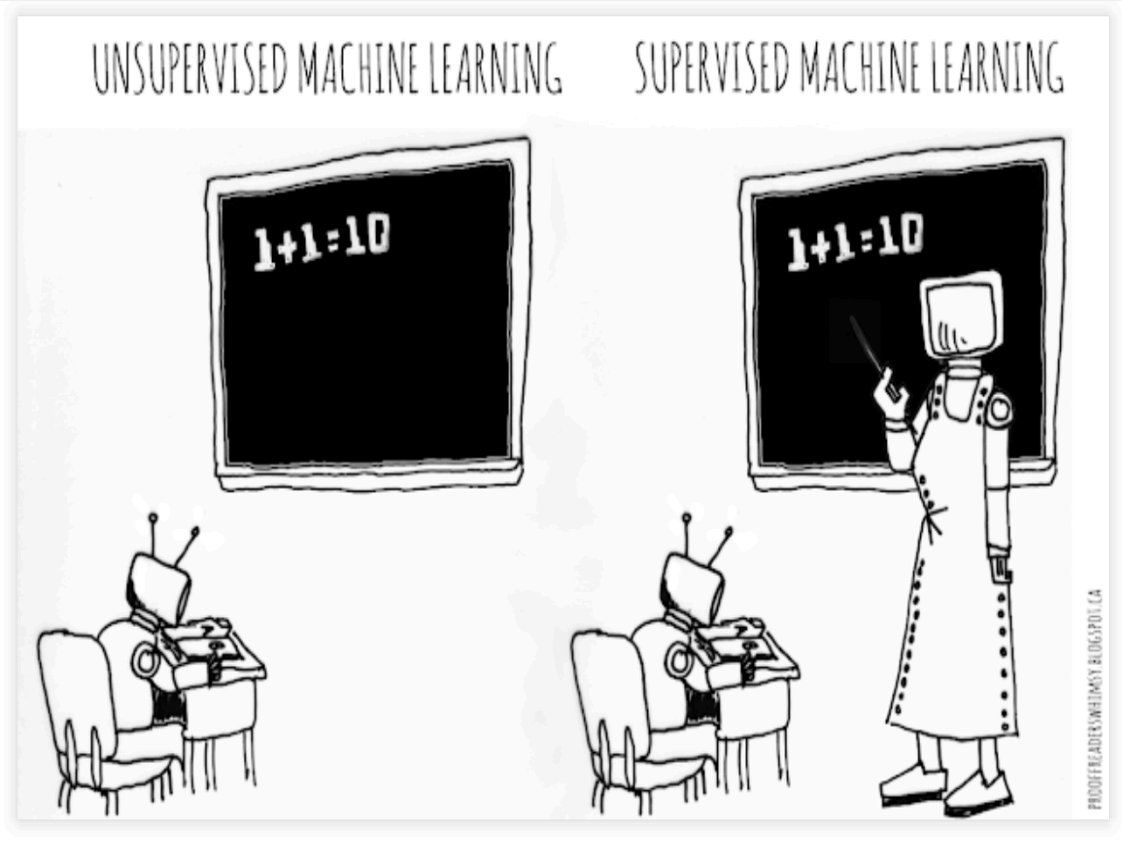
Pearson correlation coefficient

- Pearson correlation coefficient (PCC) also referred to as Pearson's r, or the bivariate correlation is a measure of linear correlation between two sets of data
- It is the covariance of two variables, divided by the product of their standard deviation
- It is essentially a normalised measurement of the covariance, such that the result always has a value between -1 and 1
- The measure can only reflect a linear correlation of variables, and ignores many other types of relationship

Demonstration

Supervised Vs unsupervised learning

- Supervised learning – teach the model like a teacher. As simple as that! Teaching is done through training the model with a labeled dataset. Examples – classification, regression.
- Unsupervised learning – as the name suggests, the model works on its own to discover information and get trained. Here, unlabeled dataset is used for training. Examples – clustering, dimension reduction, etc.



Supervised Vs unsupervised learning

Supervised learning

- Example – Classification, Regression
- Wide range of evaluation methods
- Controlled environment

Unsupervised learning

- Example – Clustering
- Compared to supervised learning, fewer evaluation methods exist.
- Less controlled environment

Labeled dataset

ID	Clump	UnifSize	UnifShap	MargAd	SingEpiSi	BarNuc	NormNu	M4it	Class
02	2	3	4	1	6	0	2	1	B
06	5	5	7	8	2	2	8	7	M
07	7	8	1	9	8	7	9	2	B
13	1	1	9	3	0	1	1	0	M
67	9	3	0	5	1	6	2	4	M
98	0	0	3	2	2	9	0	8	B
64	4	2	6	0	7	3	3	2	M

scikit-learn

- It’s free to use.
- Covers basic classification, clustering, and regression algorithms.
- Works very well with NumPy and SciPy
- Very easy to implement


10 minutes break

Introduction to regression

Introduction to regression

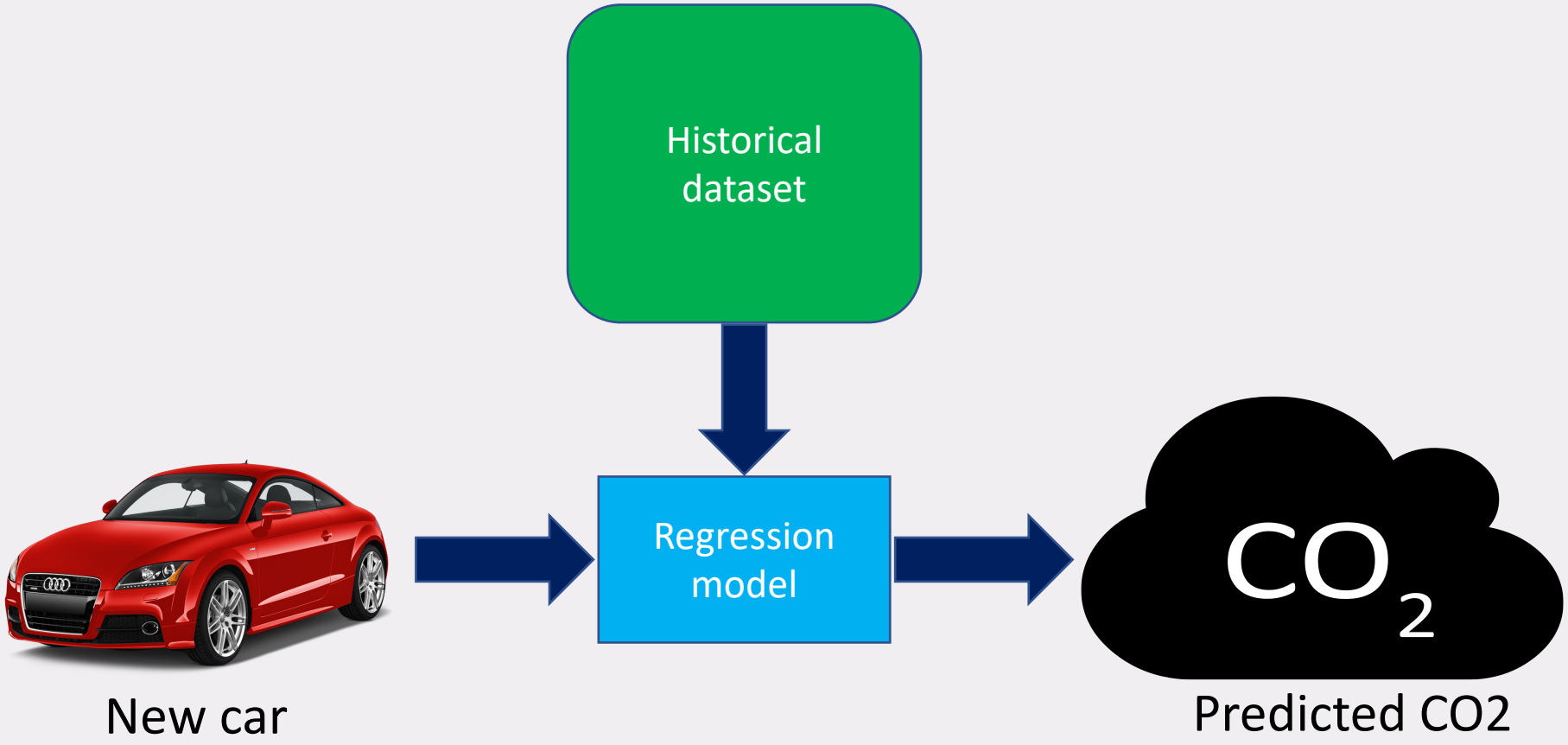
Serial No	Engine size	Cylinders	Fuel consumption	C02 Emission
1	3	6	8.3	239
2	4	12	12.6	359
3	3	12	12.6	359
4	3	8	11.3	338
5	2	8	12.2	354
6	2	8	11.3	338
7	3	8	12.2	354
8	3	12	12.6	?

A regression method can be used to predict a continuous variable, for example, CO2 Emission.

Introduction to regression

- Dependent variable such as CO2 emission is usually denoted by ’y’
- Independent variables such as, engine size, cylinder, fuel consumption etc. can be denoted as a matrix ‘X’.
- Relation between X & Y can mathematically be denoted as, Y= f(X). Which means, Y is a function of X. This function can be linear or nonlinear depending upon the state of the underlying system.

Introduction to regression



Regression models

Simple regression

- Simple linear regression
- Simple nonlinear regression
- Single input single input (SISO)

Multiple regression

- Multiple linear regression
- Multiple nonlinear regression
- Multiple inputs, single output (MISO) and multiple inputs multiple outputs (MIMO)

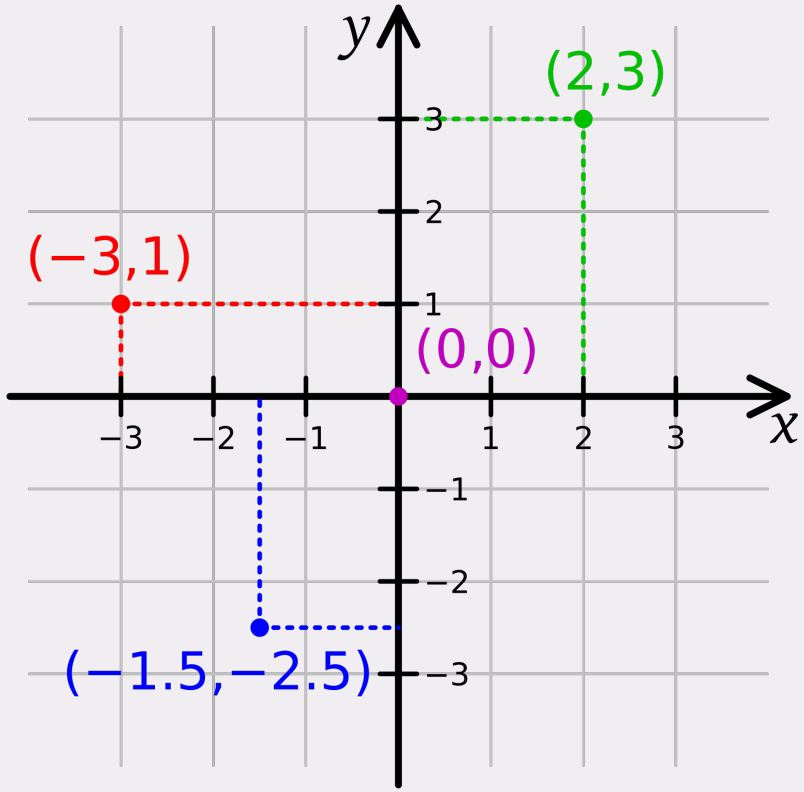
Applications of regression

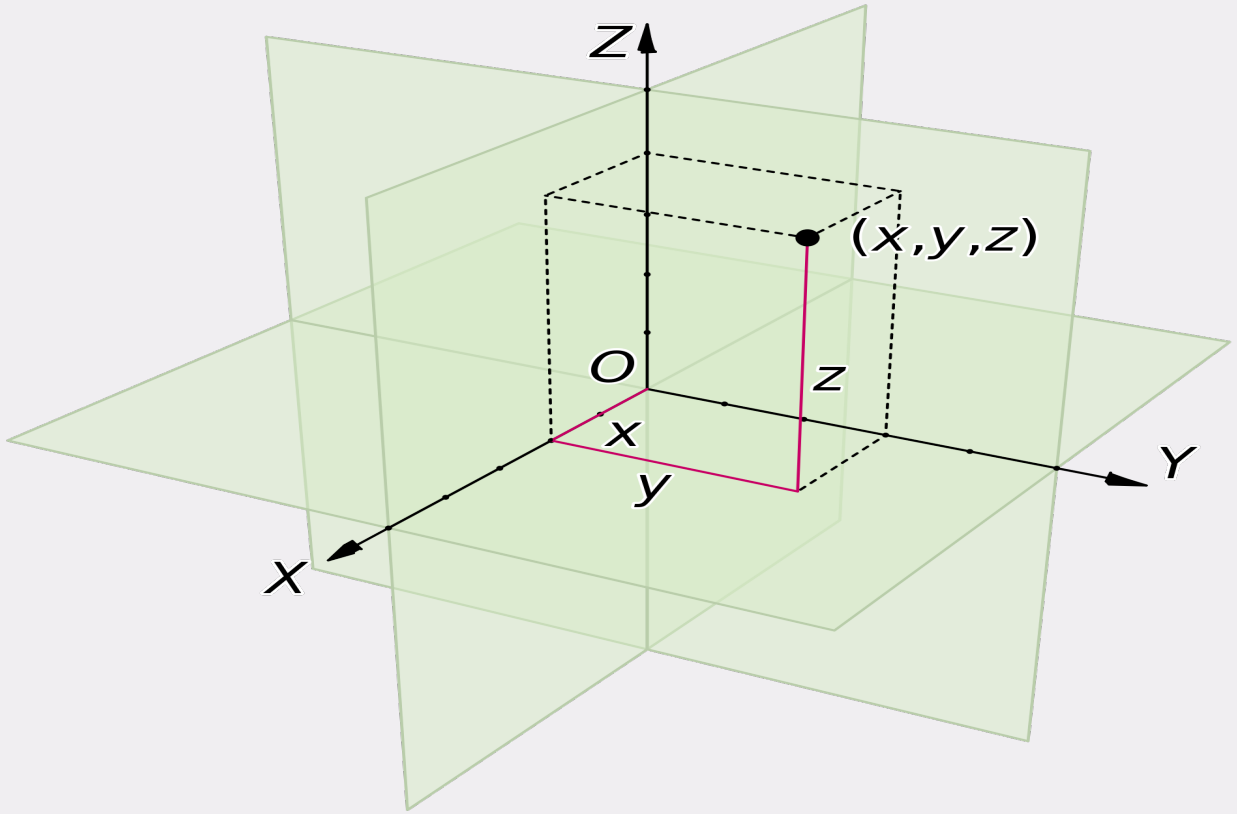
Source - https://people.com/

Source - https://blogs.biomedcentral.com/

Source - https://climate.nasa.gov/

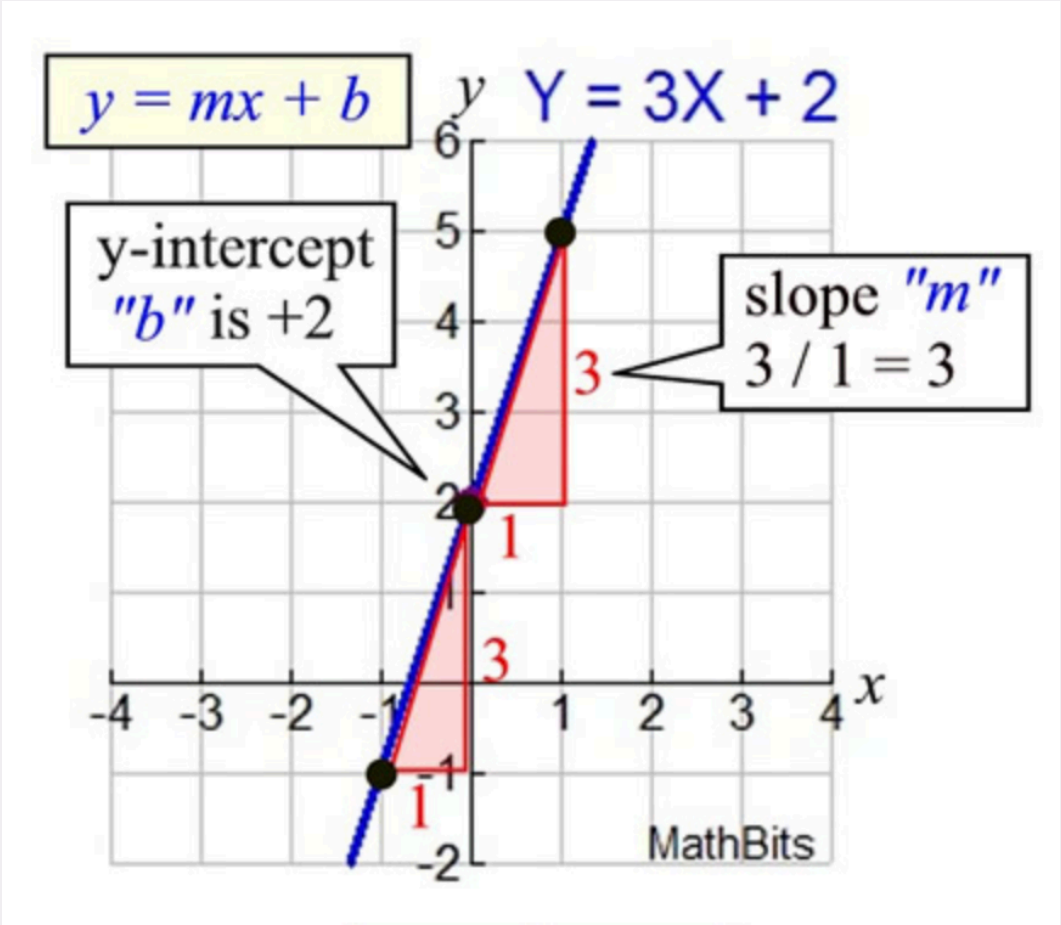
Cartesian coordinate plane

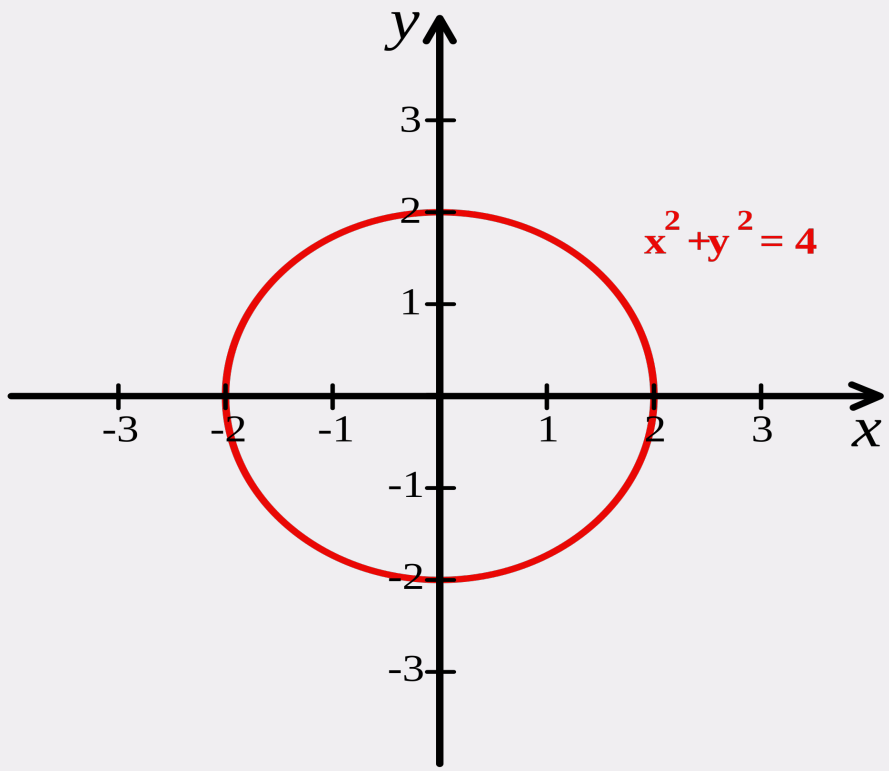




Source - https://en.wikipedia.org/wiki/Cartesian_coordinate_system

Straight line and curve





Source - https://en.wikipedia.org/wiki/Cartesian_coordinate_system

Linear regression use case

Serial No	Engine size	Cylinders	Fuel consumption	C02 Emission
1	3	6	8.3	239
2	4	12	12.6	359
3	3	12	12.6	359
4	3	8	11.3	338
5	2	8	12.2	354
6	2	8	11.3	338
7	3	8	12.2	354
8	3	12	12.6	?

The dependent variable must be continuous and not discrete.

Can we use linear regression to predict a continuous variable CO2?

Simple Vs multiple linear regression

Which one is the best approach?

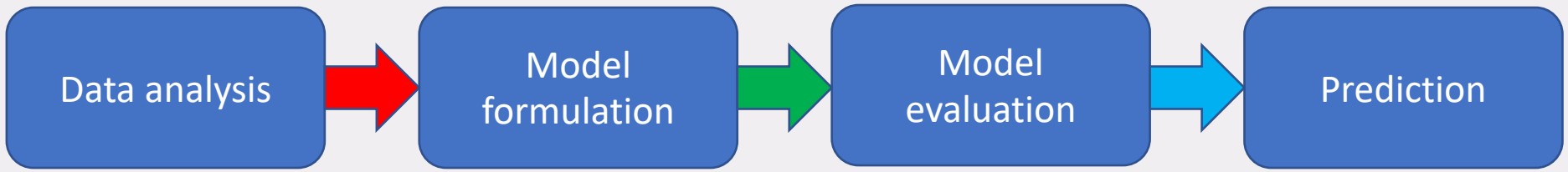
Simple linear regression

- Single input, single output (SISO)
- Dependent variable – CO2 emission
- Independent variable – engine size/cylinder/fuel consumption. Which means either of these variable. Not more than one though.

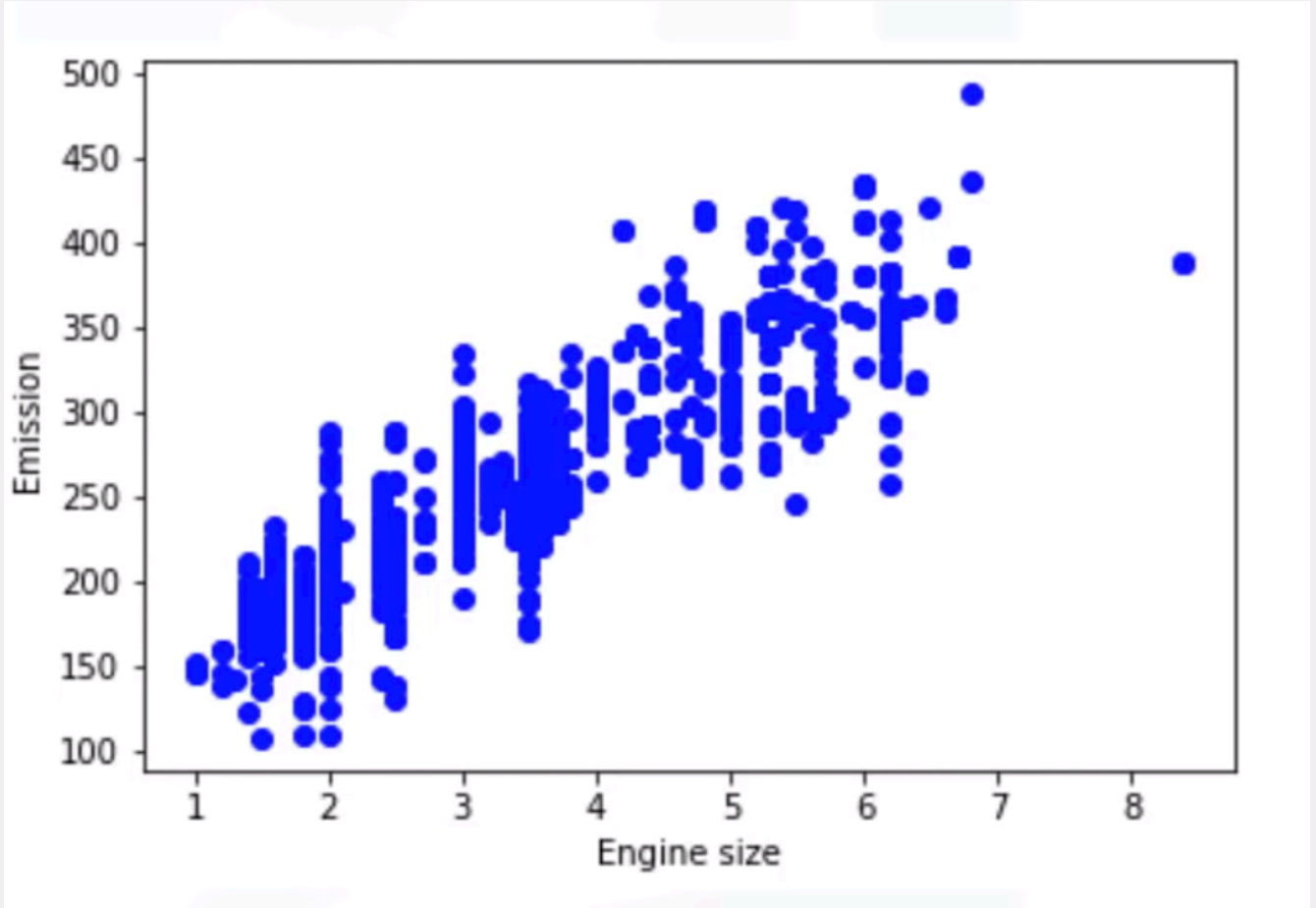
Multiple linear regression

- Multiple input, single output (MISO)
- Dependent variable – CO2 emission
- Independent variable – a combination of engine size, cylinder, fuel consumption.

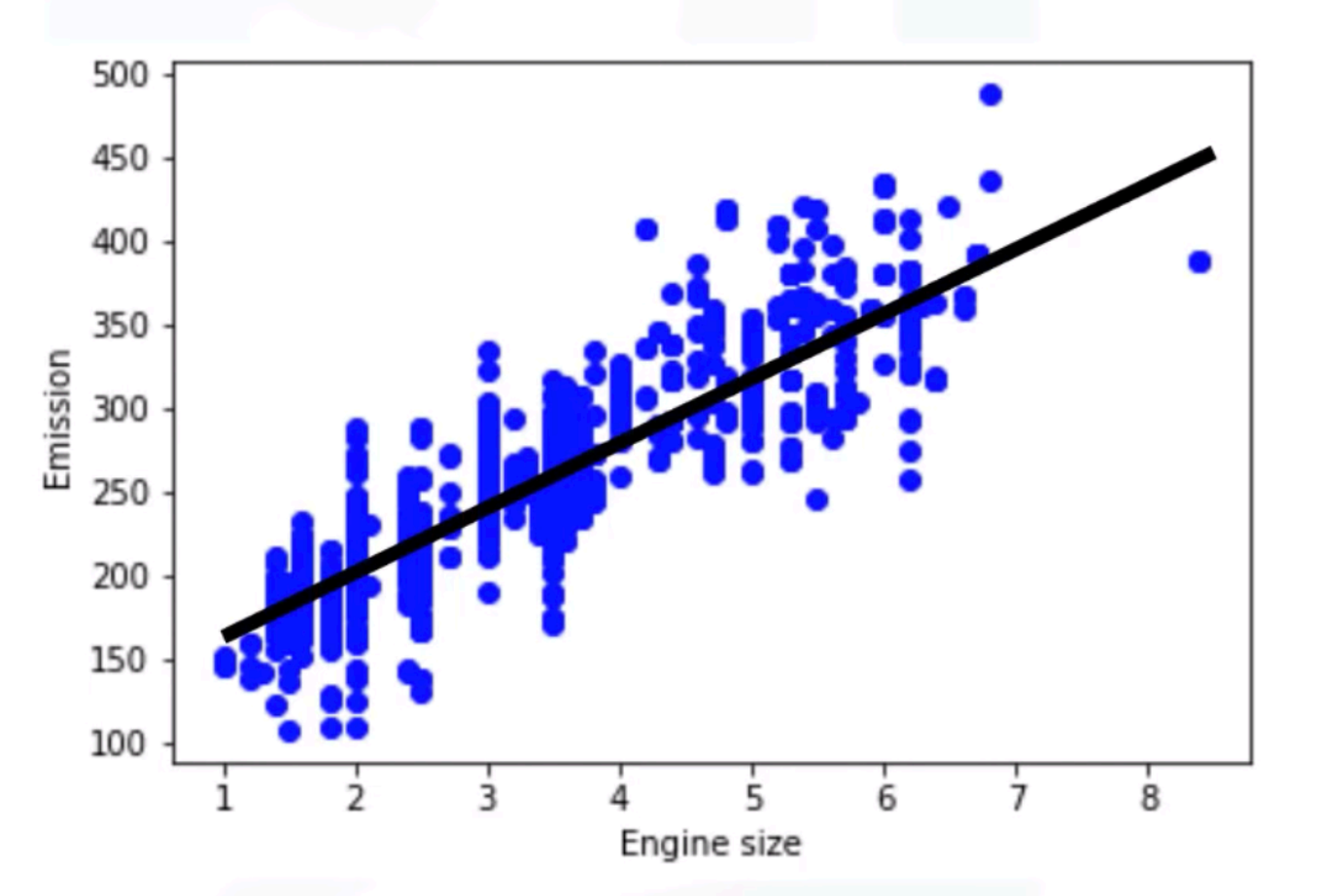
Basic steps of regression modelling



Basic data analysis



Straight line fitting



Let's assume that the independent variable, engine size is denoted by x, and the dependent variable, emission is denoted by y. The standard equation of a straight line is, y = mx + b. So essentially, we have to evaluate the values of 'm' and 'b' to correctly figure out this fitted line.

y = f(x)

Finding the best regression line (model)

- At engine size equals to 2, the actual observed value of the CO2 emission varies between 120 to 300 units approximately.
- If we try to use our model (fitted regression line), the value of CO2 emission at this particular engine size equals to 200.
- Is this a satisfactory model? If yes, then why? In no, then why?
- Residual error is defined as the difference between the model predicted value and the observed value. An ideal model will have minimum residual error.
- Mean squared error, $\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

Finding the best regression line (model)

- So the objective is to minimise the MSE equation and in order to do so, we have to find the best possible values of coefficients/parameters ‘m’ and ‘b’.
- Please note that these parameters are commonly referred as, $\theta 1$, $\theta 2$ etc.
- It’s mathematically very straightforward to evaluate these model parameter values using the optimisation techniques. But ML libraries in Python will do this job for you. It’s good to know the background though.

Salient features of linear regression

- Linear regression is very fast.
- Parameter tuning not required as optimisation is a deterministic approach.
- Model interpretation is very lucid.
- Since majority of the real world system is nonlinear so the linear description of a system is often inaccurate.

Demonstration


10 minutes break

Advanced regression

Rationale behind multiple regression

- To gauge the effectiveness of independent variables upon dependent variable. Example – how does precipitation, transpiration, temperature, humidity contribute towards soil water content? Do they contribute at all?
- To measure the sensitivity of a variable with respect to other variables present in the system. Example – which are the most and least contributing factors towards climate change. Candidate variables in this case could be, temperature, human intervention, evapotranspiration etc.

Example of multiple regression

Serial No	Engine size	Cylinders	Fuel consumption	C02 Emission
1	3	6	8.3	239
2	4	12	12.6	359
3	3	12	12.6	359
4	3	8	11.3	338
5	2	8	12.2	354
6	2	8	11.3	338
7	3	8	12.2	354
8	3	12	12.6	?

y (CO2 emission) = k0+k1*Engine size + k2*Cylinders +k3*Fuel consumption.

Parameter estimation

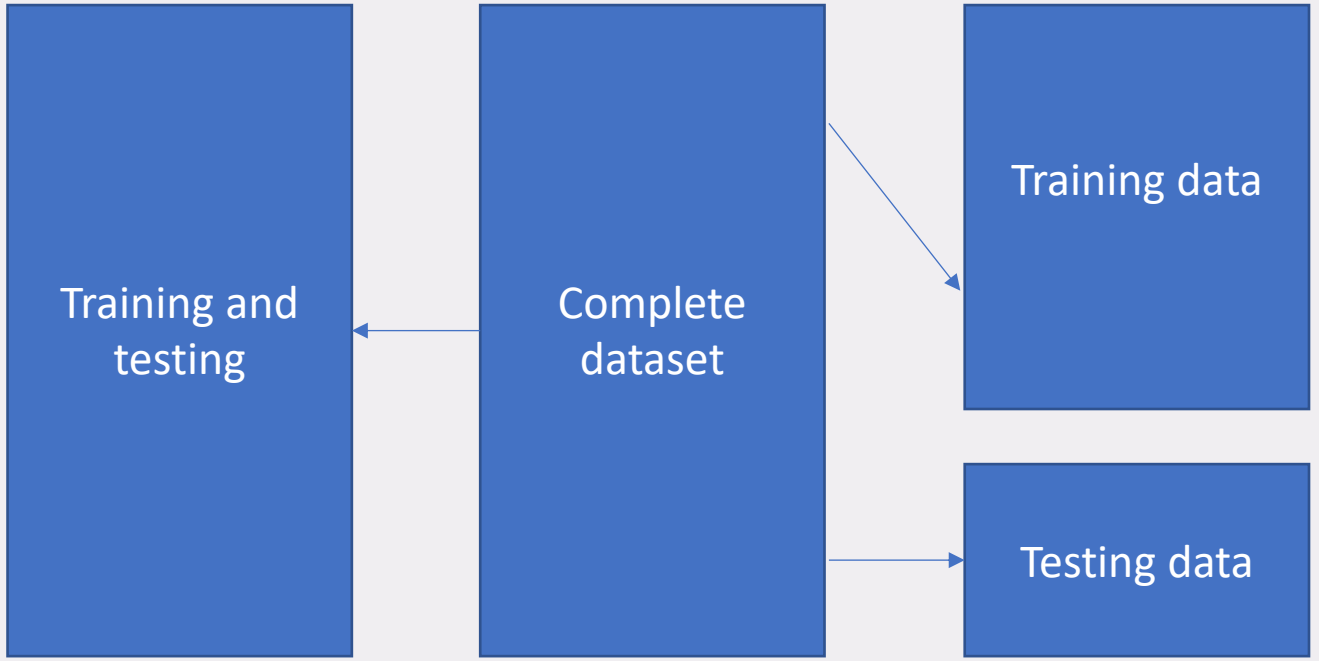
- Ordinary least square (https://en.wikipedia.org/wiki/Ordinary_least_squares) – essentially involves linear algebraic operation. Quite efficient for smaller datasets ~ 10000 data points.
- Optimisation algorithm such as gradient descent (https://en.wikipedia.org/wiki/Gradient_descent) – usually considered for a larger dataset. However, gradient descent is fundamental to many other ML and DL algorithms.

Question?

- How can we choose between simple linear regression and multiple linear regression?
- Does it mean that having many independent variables in our model gives better prediction?
- If multiple linear regression is chosen then how many exact number of variables be included?

Model evaluation

- When training and testing datasets are same – high training accuracy, low out of sample accuracy
- Training and testing dataset should be distinct.
- For time series dataset, training and testing dataset can be chopped periodically.



Training accuracy Vs Out-of-sample accuracy

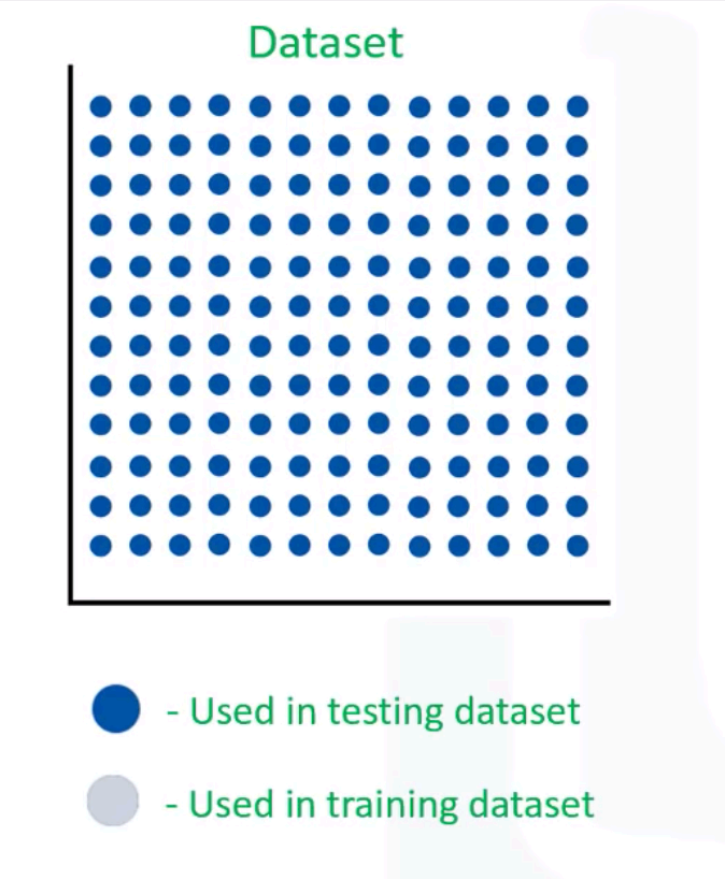
Training accuracy

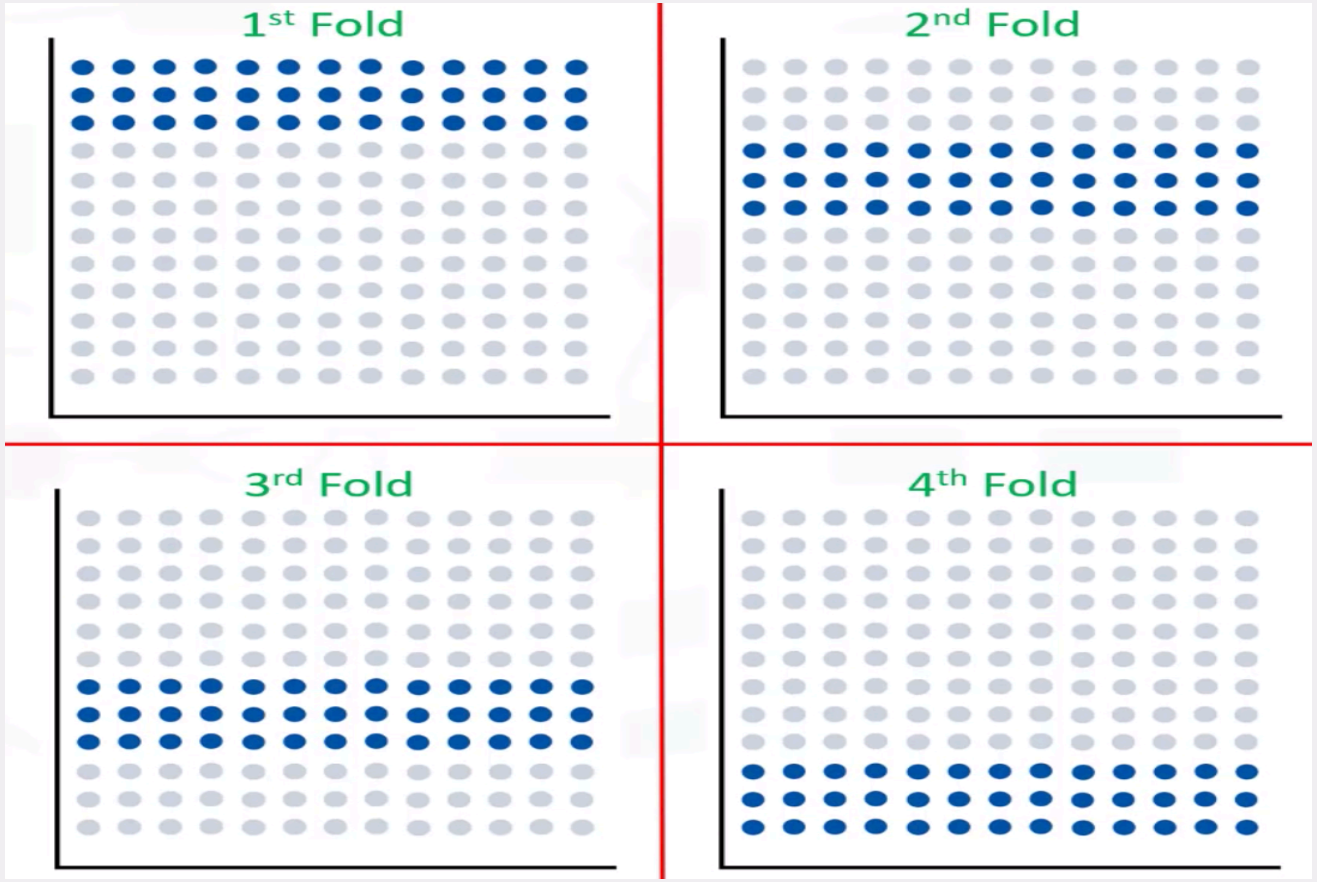
- High training accuracy often signifies model over fitting. Indeed not desired.
- Over fitting in ML refers to a situation where substantial noise has been captured and therefore the model is not generalised enough.

Out-of-sample accuracy

- High out of sample accuracy signifies that the model is not over-fitted and pretty much a generalised representation of the underlying system.
- A decent out-of-sample accuracy is a prerequisite for model fidelity and reliable prediction.

K-fold cross validation





Average accuracy (all 4 folds) are considered to be the model accuracy

Source - https://cognitiveclass.ai/

Evaluation metrics

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

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

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

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

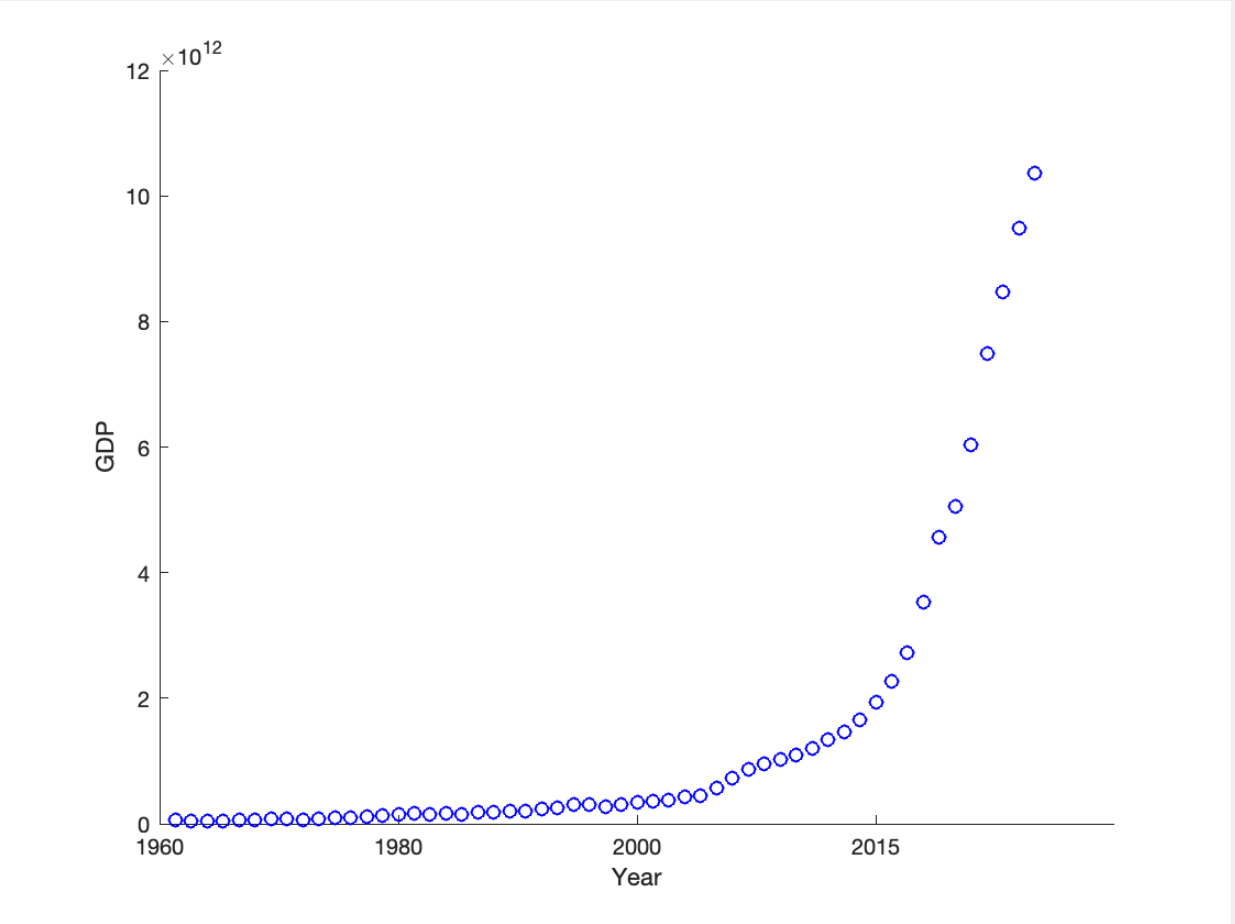
$$RSE = \frac{\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{\sum_{j=1}^{n}(y_j - \bar{y})^2}$$

R square = 1 – RSE

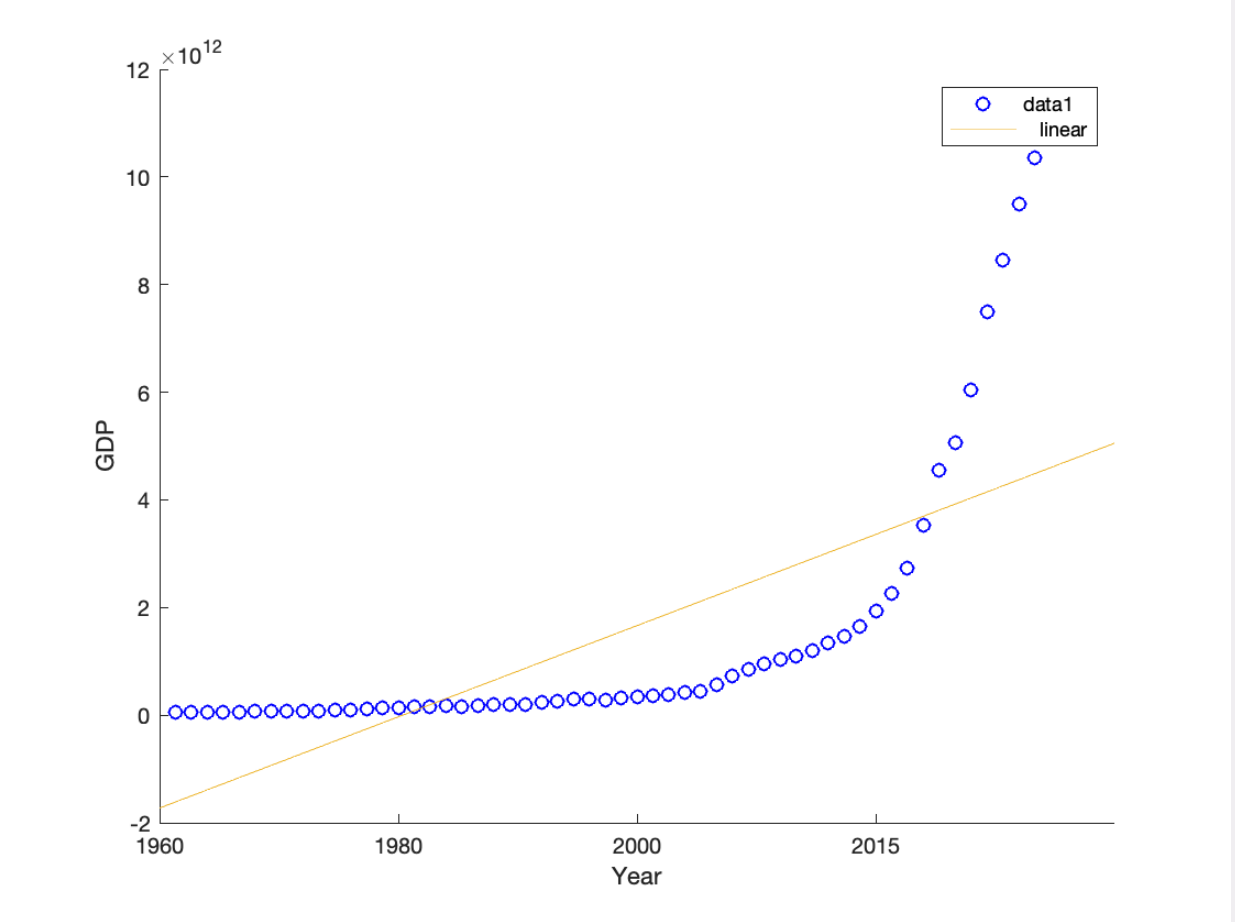
This is a very popular metric to gauge model accuracy . In ideal scenario, R square is very close to 1.

The choice of evaluation metric depends upon, the data-type, domain type and most importantly prior experience/intuition.

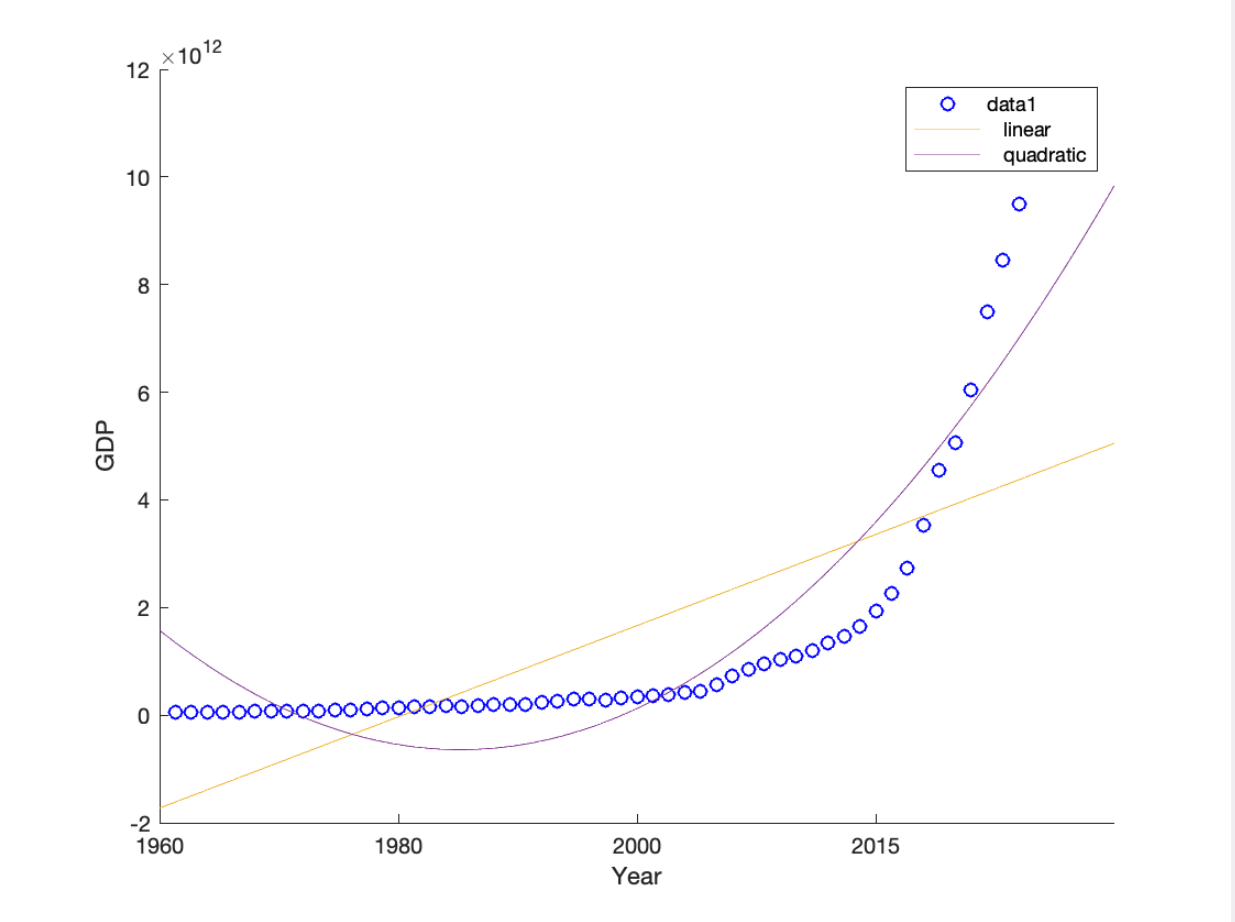
Nonlinear regression



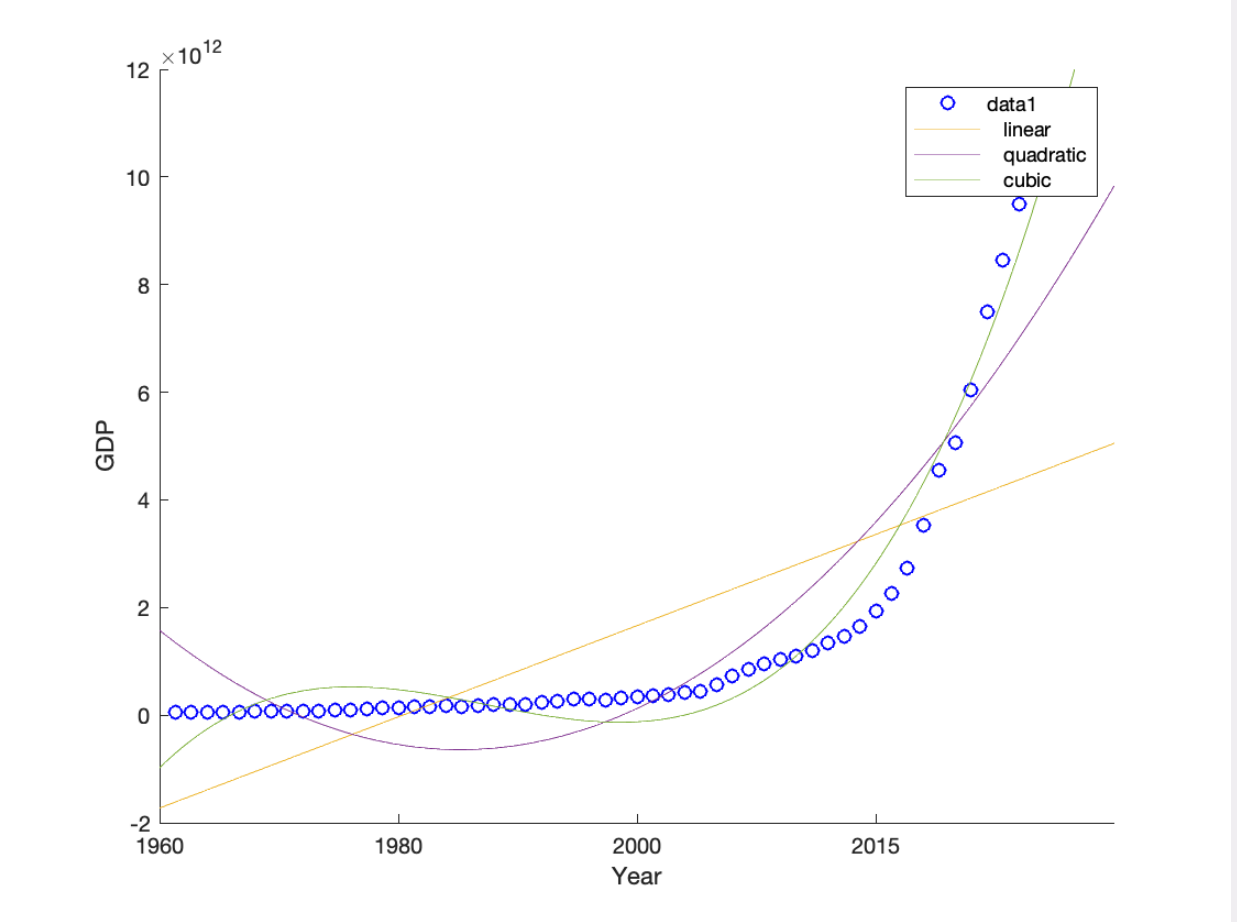
Nonlinear regression



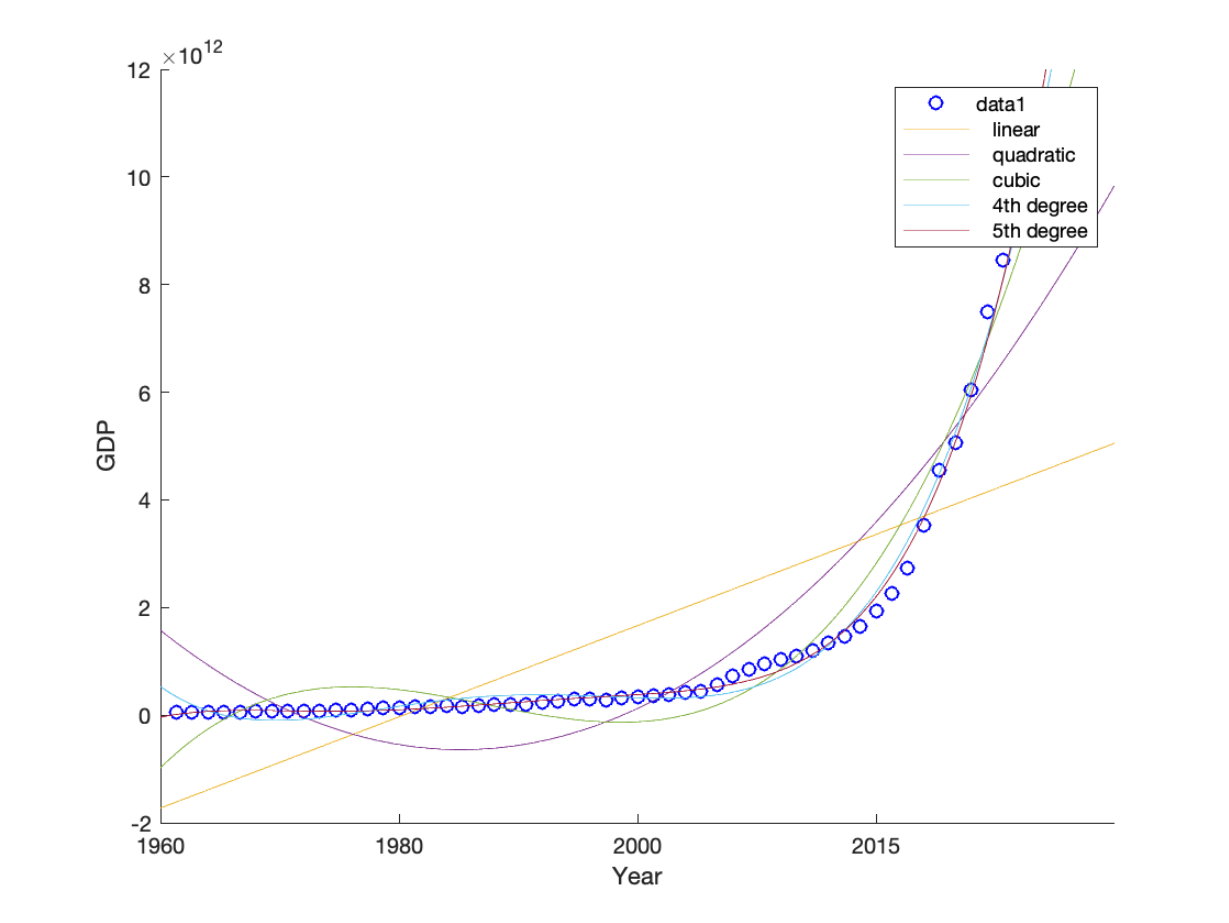
Nonlinear regression



Nonlinear regression



Nonlinear regression




```
y = p1*x^5 + p2*x^4 +
      p3*x^3 + p4*x^2 +
      p5*x + p6

Coefficients:
  p1 = 2.1925e+05
  p2 = -2.1021e+07
  p3 = 7.2754e+08
  p4 = -1.0417e+10
  p5 = 6.1721e+10
  p6 = -3.9209e+10

Norm of residuals =
     1.1668e+12
```

Polynomial regression – relationship between ‘x’ and ‘y’ are modelled using nth degree polynomial in x.

Polynomial regression

- y= p1*x^5 + p2*x^4 + p3*x^3 + p4*x^2 + p5*x + p6 is a polynomial regression of degree 5.
- A polynomial regression can be converted into linear regression for the ease of computation. Such as, if x^5 = x1, x^4 = x2, x^3 = x3, x^2 = x4, x = x5, then, y= p1*x1 + p2*x2 + p3*x3 + p4*x4 + p5*x5+ p6. This is now essentially a multiple linear regression.
- Least square can be used to estimate parameter in multiple linear regression. LS essentially works by minimising the sum of the squares of the differences between observed ‘y’ and model predicted ‘y’.

Points to ponder

- How to deduce that the problem (system) is linear and nonlinear in a straightforward way? (By visual inspection of the scatter plot among dependent and independent variable. By calculating the correlation coefficient https://en.wikipedia.org/wiki/Correlation_coefficient)
- Which degree of polynomial should be chosen if a polynomial regression is to be used. Again the principle of model overfitting comes into play. It requires lots of experience to figure out the correct order of polynomial in regression.
- Static Vs Dynamic system!!

Demonstration

Regression for dynamical system

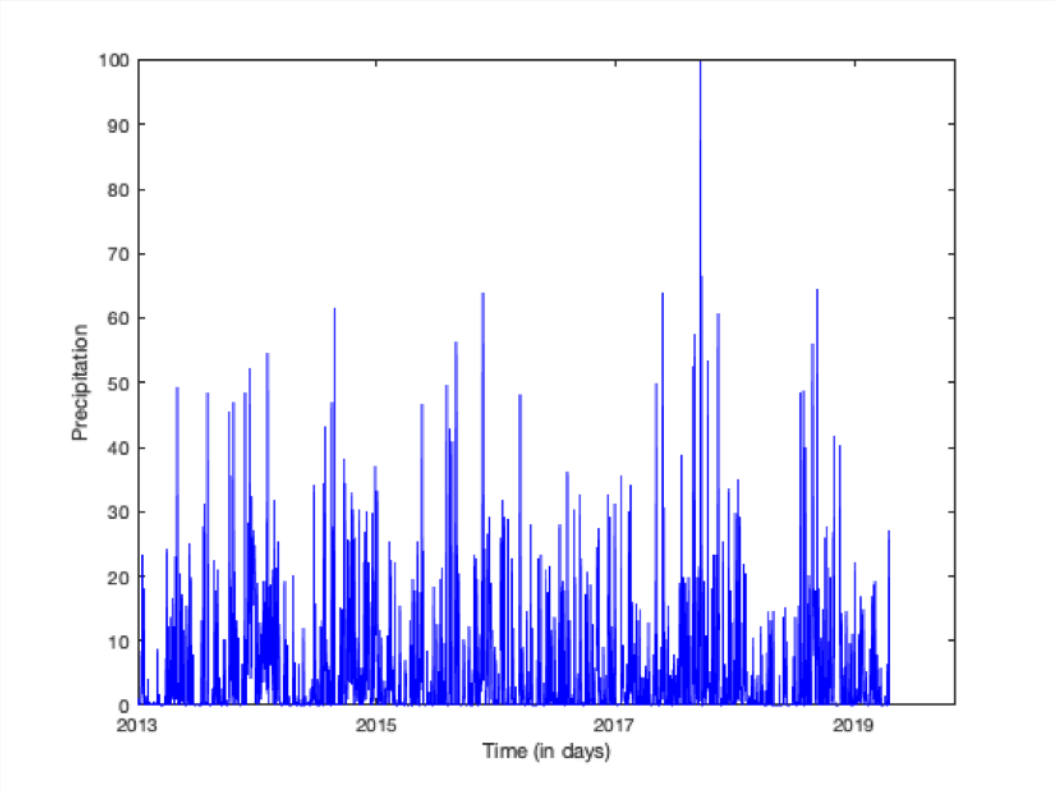
NARMAX model structure

$$y(k) = F[y(k-1), y(k-2), \ldots, y(k-n_y), u(k-d), u(k-d-1), \ldots, u(k-d-n_u), e(k-1), e(k-2), \ldots e(k-n_e)] + e(k)$$

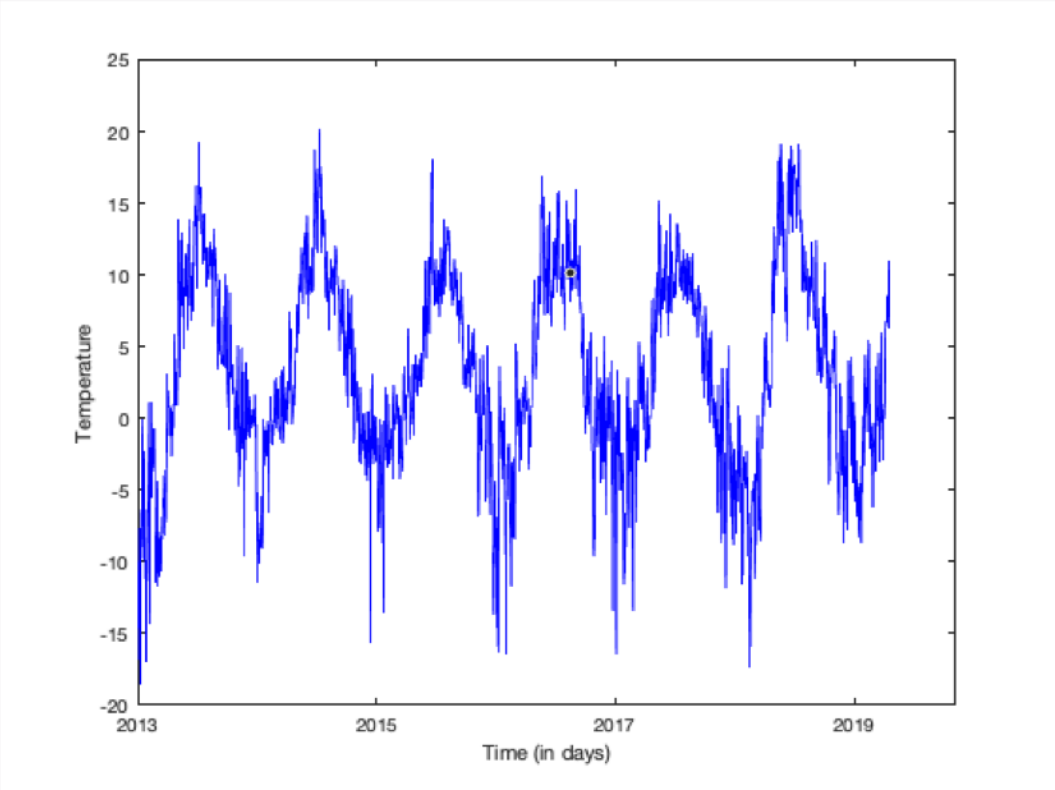
y(k) = system output, u(k) = system input, e(k) = noise sequence, ny = maximum lag for system output nu = maximum lag for system input, n(e) = maximum lag for noise, F = a nonlinear function, d = time delay

NARMAX model structure is very useful for dynamical system, wherein, current output state depends on the lagged versions of past inputs-outputs

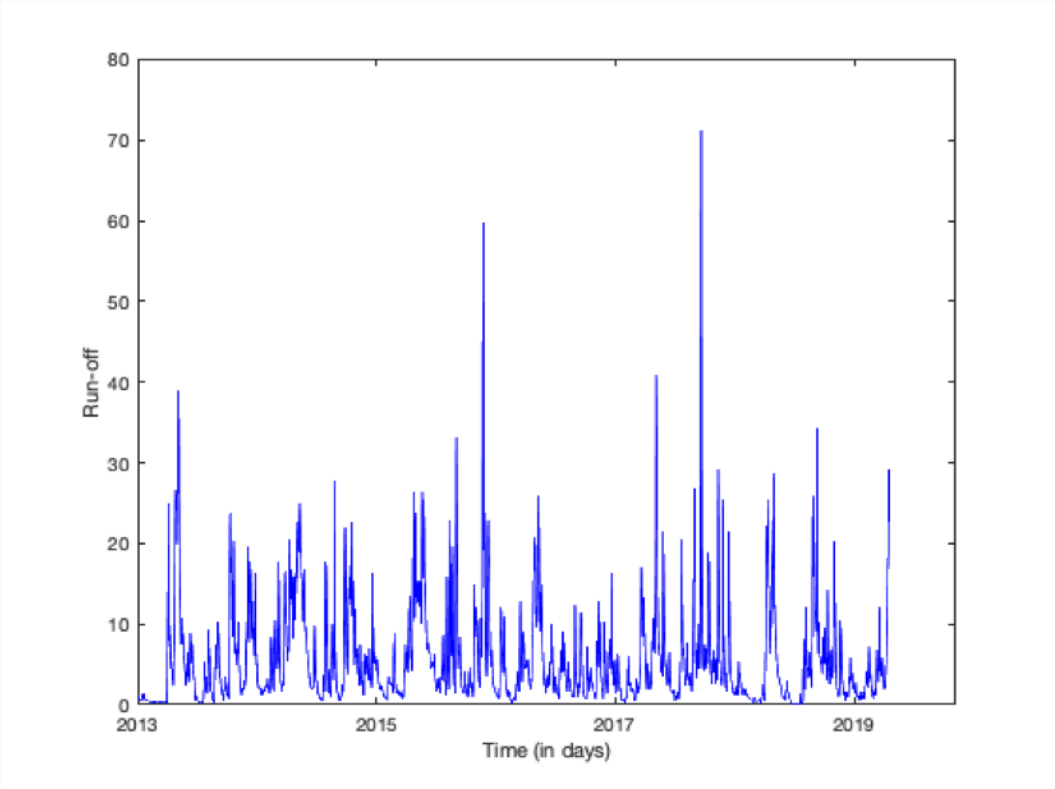
Precipitation (south Norway catchment)



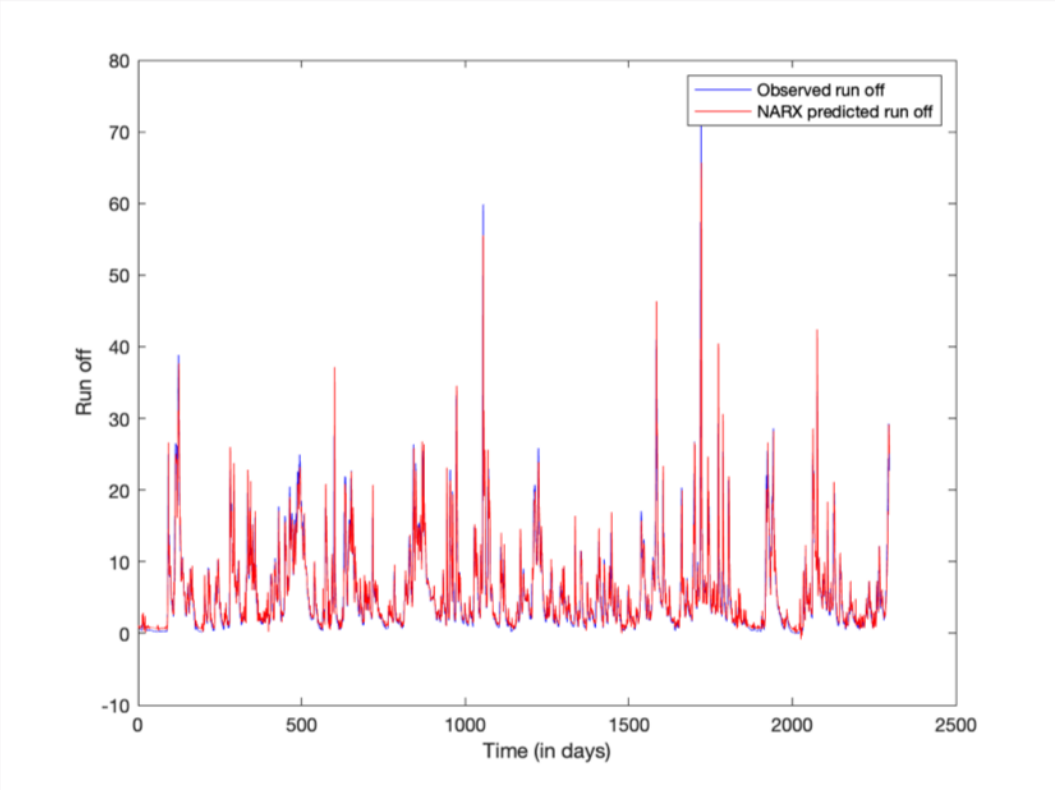
Average temperature (south Norway catchment)



Run-off (south Norway catchment)



NARMAX model fitting (south Norway catchment)



Assignment

Q & A

Cardiff Metropolitan University | Prifysgol Metropolitan Caerdydd