Linear Regression

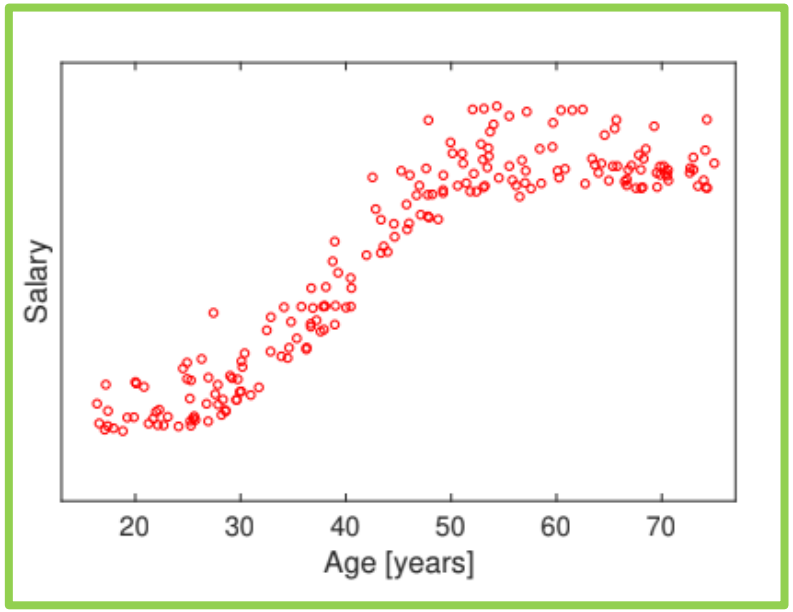
What we will Learn..

- Regression: Predictor(s) and Label
- Linear Regression: Simple Linear Regression, Multiple Linear Regression,
- Mathematical Notation of Regression (Simple Linear Regression)
- Candidate Solutions and Determining a good solution for Linear Regression
- Nature of Error in Regression
- Regression as an optimization problem
- Least Square Error
- Beyond Linearity: Polynomial Regression
- Flexibility, Interpretability, Accuracy of a Model
- Generalization: Underfitting, Overfitting and Right fitting of a model

Data as a point in a space

Let’s assume you are given with a dataset contains Salary and Age data in it and this dataset is plotted on a graph space where x axis corresponds to “Age” and y axis corresponds to “Salary”.

Age	Salary
18	12000
37	68000
66	80000
25	45000
26	30000
..	..
..	..



Predictors and Labels

Age	Salary
18	12000
37	68000
66	80000
25	45000
26	30000
..	..
..	..

In this Dataset:

- Age is the predictor, Salary is the Label
- Salary is the Predictor, Age is the Label
- Both options can be considered

Predictors and Labels Continued… Association and Causation

Sometimes, Prediction Models are interpreted as a method of indicating Causation (The predictor is a Cause and Label is the Effect). However, as the Role of “Predictor” and “Label” can be reversed, so need to use Caution.

The real reason why we can build such models is the Association between the attributes, rather than causation. Two attributes in a Dataset appear associated, if:

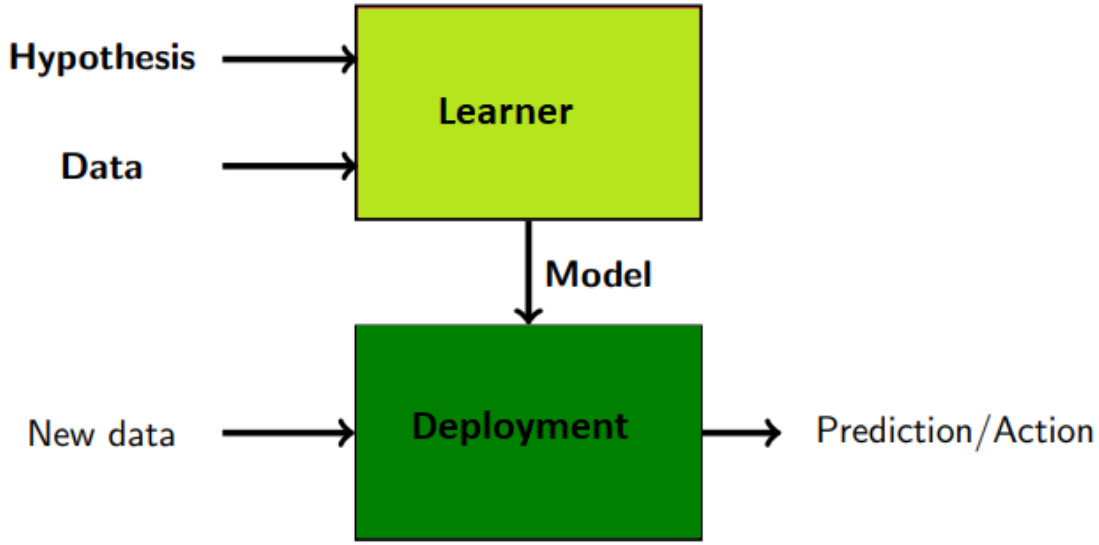
- If one causes another
- When both have common cause
- Due to Sampling

Remember: *All the Causation are Association, but not all the Association are Causation*

Regression

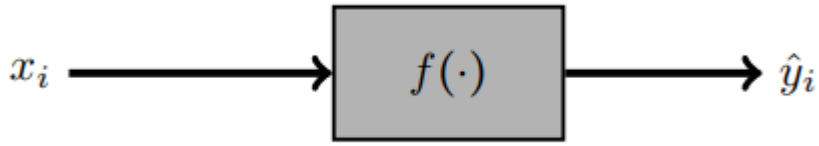
Regression is a statistical measure that determines the strength of the relationship between one “dependent variable” (Usually values, y_1 ,y_2 ,y_3 ….y_n exist in Y axis) against the another “independent variable” (Usually the values x_1 , x_2 , x_3, ……x_n exist in X axis)

Regression Learner



Hypothesis: Type of Model (e.g. Linear)
Data: Collection of samples consisting of one true (label) and one or more predictors
Model: Predict a label based on the predictors

Mathematical Notation



Dataset:

- N is the number of samples, i identifies each sample
- x_i is the **predictor** of sample i
- y_i is the (continuous) **true label** of sample i
- The dataset is $\{(x_i, y_i) : 1 \leq i \leq N\}$, and (x_i, y_i) is sample i

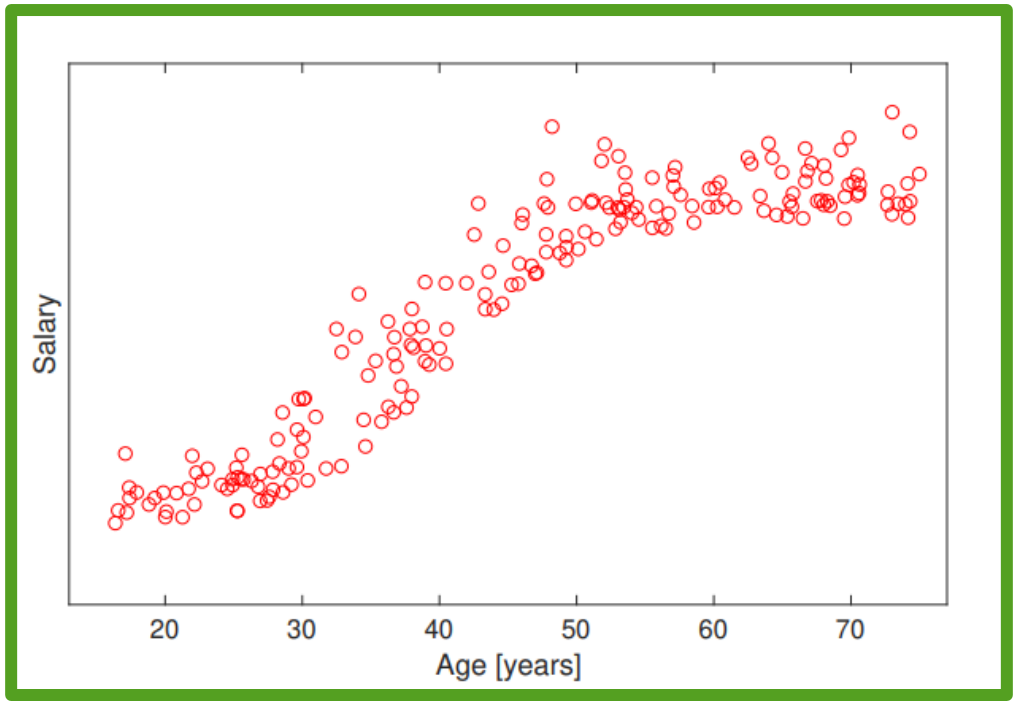
Model:

- $f(\cdot)$ denotes the model
- $\hat{y}_i = f(x_i)$ is the **predicted label** for sample i
- $e_i = y_i - \hat{y}_i$ is the **prediction error** for sample i

(*Note that we are considering* **one predictor** *here, this notation will be extended to multiple predictors when discussing multivariate models*)

Simple Regression

Simple Regression considering one predictor ***x*** and one label ***y***.



Simple Linear Regression

- There is only one Predictor, “x”
- Relationship between “x” and “y” is described by a Linear Function
- Changes in “y” are assumed to be related with the changes in “x”

Simple Linear Regression Continued…

In **Simple Linear Regression**, Candidate Models are defined by the mathematical expression:

$$f(x) = w_0 + w_1 x$$

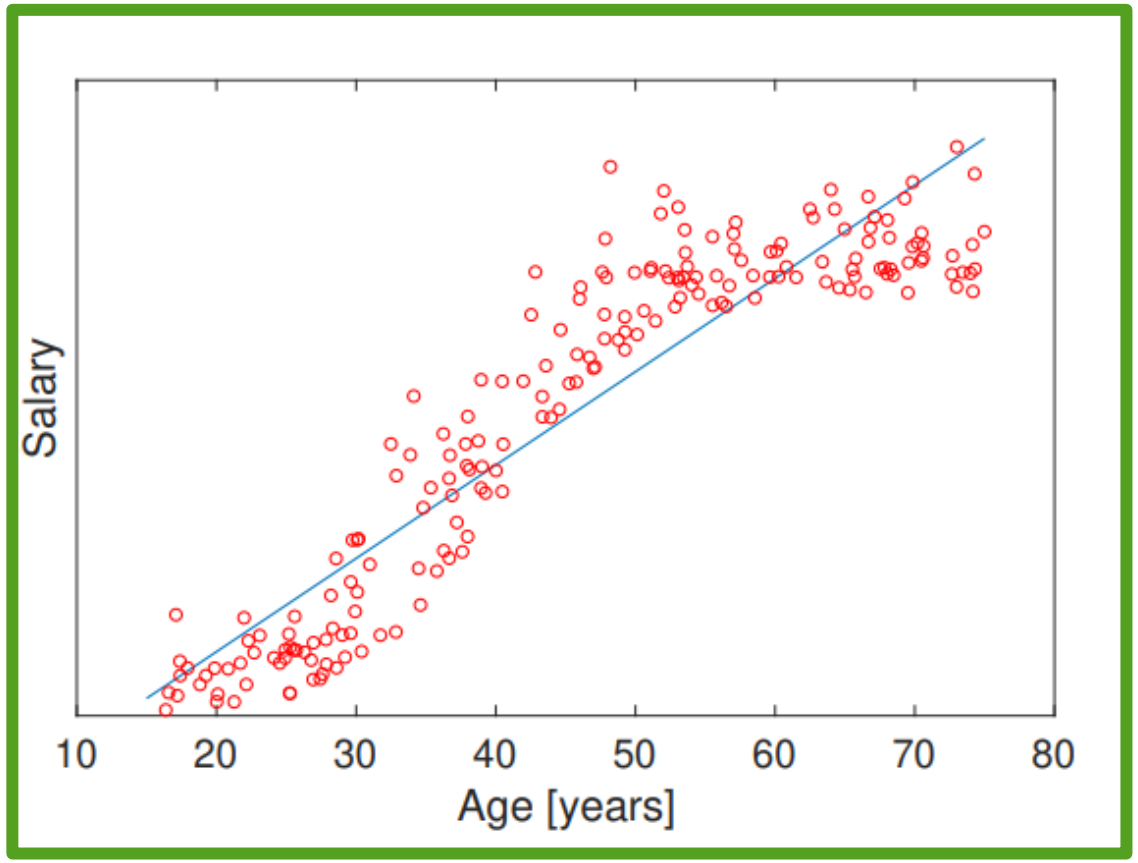
Hence, the predicted label $\hat{y}_i$ can be expressed as:

$$\hat{y}_i = f(x_i) = w_0 + w_1 x_i$$

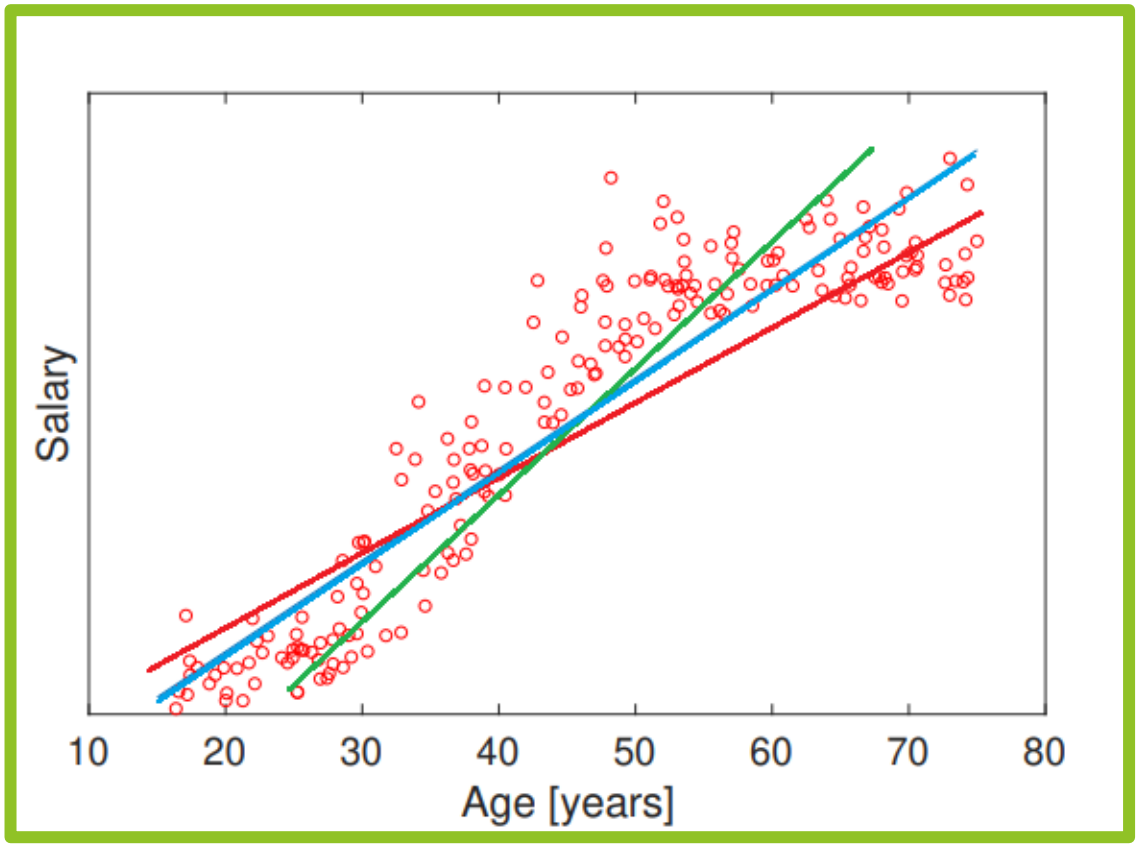
Linear models have therefore "two parameters" w_0 (intercept) and here w_1 (slope), which need to tuned to achieve the highest quality.

Note: While Analyzing data, the data in the dataset is used to tune the parameters.

Simple Linear Regression Solution



Linear Candidate Solutions



How to determine a good solution?

To find the best model we need to know the model quality.

One popular quality metric in regression problems is the **mean squared error (MSE)**

MSE corresponds to the expected squared error of the prediction of a model during deployment.

If we are given a dataset consisting of **N** samples and a model $f(\cdot)$,we can estimate its MSE as follows:

$$
\begin{aligned}
E_{MSE} &= \frac{1}{N}\sum_{i=1}^{N} e_i^2 \\
&= \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N} (y_i - f(x_i))^2
\end{aligned}
$$

Is this possible to find a model such that E_{MSE}= 0?

The nature of error

While considering a regression problem we need to be aware that:

- The chosen predictors might not include all the factors that determine the label.
- The chosen model might not be able to represent accurately the true relationship between response and predictor (the pattern).
- Random mechanisms (noise) might be present.

Mathematically, we represent this discrepancy as,

$$\begin{aligned} y &= \hat{y} + e \\ &= f(x) + e \end{aligned}$$

There will always be some discrepancy (error e) between the true label y and our model prediction f(x).

Regression as an Optimization Problem

Given a dataset $\{(x_i, y_i) : 1 \leq i \leq N\}$, every candidate model f has its own E_{MSE}. Our goal is to find the **model with the lowest** E_{MSE}:

$$f_{best}(x) = \arg\min_{f} \frac{1}{N} \sum_{i=1}^{N} (y_i - f(x_i))^2$$

The question is, how do we find such model? Finding such a model is an **optimisation problem**.

Notice that we are looking for the model that minimises E_{MSE} **on the dataset**, however this model might not be the **minimum MSE** (MMSE) solution, i.e. the best model during **deployment**!

Optimal solution for Simple Linear Regression

In Simple Linear Regression, the training dataset can be represented by a "design matrix"

$$X = [x]^T = [1, x_1, x_2, \ldots\ldots\ldots, x_N] = \begin{bmatrix} 1 & x_{1,1} \\ 1 & x_{2,1} \\ \vdots & \vdots \\ 1 & x_{N,1} \end{bmatrix}$$

The Label Vector $\boldsymbol{y}$:

$$\boldsymbol{y} = [y_1, \ldots, y_N]^T = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

The Least Square Solution

It can be shown that the **linear model** that minimises the metric E_{MSE} on a **training dataset** defined by a design matrix $\mathbf{X}$ and a label vector $\mathbf{y}$, has the parameter vector:

$$\boldsymbol{w} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{y}$$

This is an **exact** or **analytical solution** and is known as the **least squares** solution. It is valid for simple and multiple linear regression.

Note that the inverse matrix $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ *exists when all the columns in* $\mathbf{X}$ *are independent.*

Simple Linear Regression Example

Assume, a housing company wants to determine if there's any relationship (association) between the size of the house (square feet) and their selling prices from the following sample dataset.

Price (($) 1000s)	Size (Square Feet)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Let's assume the association is as:

- Dependent Variable (**y**) = **house price in $1000s**
- Independent Variable (**x**) = **square feet**

Can you predict the price of the house of size 2000 Square Feet?

Simple Linear Regression Example Continued…

$$\hat{y}_i = f(x_i) = w_0 + w_1 x_i$$

House Price = 98.24833 + 0.10977 (Square Feet)

Here, we used Least Square Solution to determine the parameters w_0 and w_1.

House price corresponds to the predicted value, $f(x_i)$ and Square Feet corresponds to the predictor x_i.

Now, the price for the House with 2000 square feet:

House Price = 98.24833 + 0.10977 (2000) = $ 317.85

Multiple Linear Regression

If we have two or, more “predictors” then the regression become Multiple Linear Regression.

$$f(\boldsymbol{x}_i) = w_0 + w_1 x_{i,1} + \cdots + w_K x_{i,K}$$

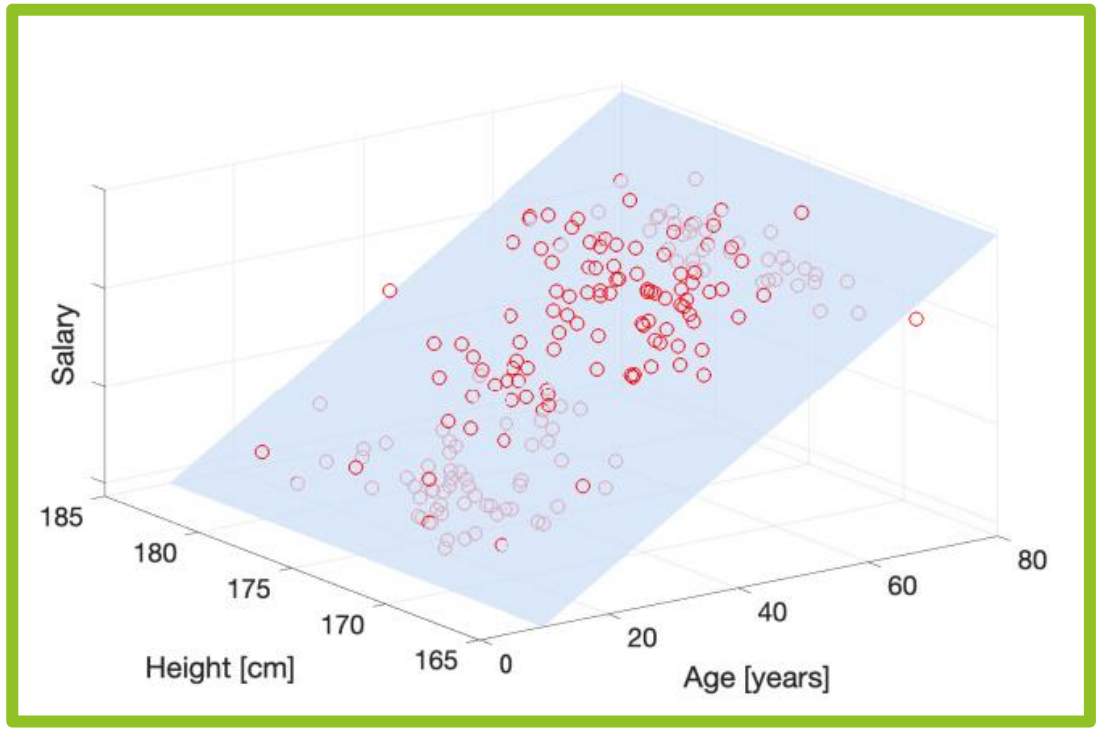
Multiple Linear Regression Example

Age (Years)	Height (cm)	Salary
18	175	12000
37	180	68000
66	158	80000
25	168	45000
.. ..	.. ..	.. ..

Assuming, In this Dataset:

- Age and Height are the predictors, Salary is the Label

Multiple Linear Regression Example Continued…



Multiple Linear Regression models are planes (or, hyper planes)

Using a Vector Notation to represent a Multiple Linear Regression

Consider a dataset consisting of 4 samples described by three attributes, namely age, height and salary:

Age (Years)	Height (cm)	Salary
18	175	12000
37	180	68000
66	158	80000
25	168	45000

We decide to build a linear model that maps age and height (As predictors) to salary (label):

- Use vector notation to represent the resulting linear regression model
- Obtain the design matrix X and response vector y.

Using a Vector Notation to represent a Multiple Linear Regression Continued…

In multiple linear regression, the **training dataset** can be represented by the **design matrix X**:

$$\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^T = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,K} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N,1} & x_{N,2} & \cdots & x_{N,K} \end{bmatrix}$$

and the **label vector y**:

$$\boldsymbol{y} = [y_1, \ldots, y_N]^T = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

Beyond Linearity: Simple Polynomial Regression

The general form of a polynomial regression model is:

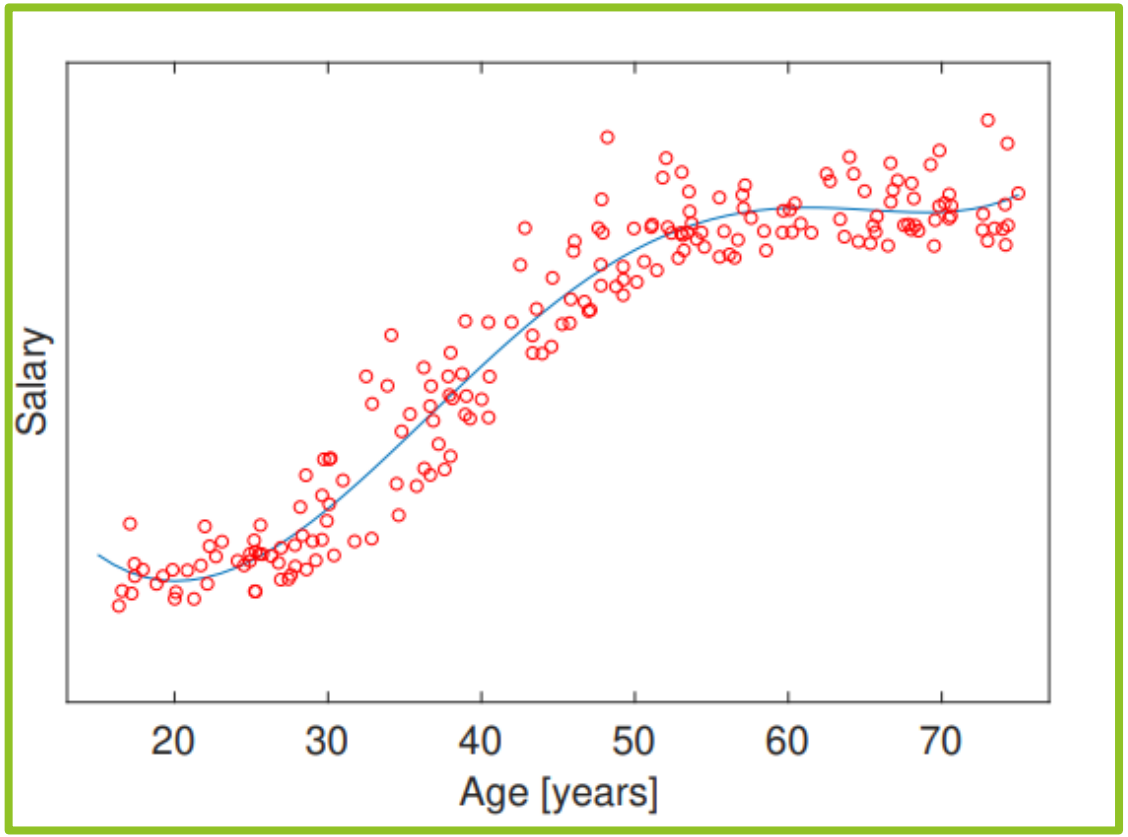
$$f(x_i) = w_0 + w_1 x_i + w_2 x_i^2 + \cdots + w_D x_i^D$$

where D is the degree of the polynomial.

By treating the powers of the predictor x as predictors themselves, simple polynomial models can be expressed as multiple linear models:

$$f(x_i) = w_0 + w_1 x_i + w_2 x_i^2 + w_3 x_i^3$$

Beyond Linearity: Simple Polynomial Regression



Flexibility

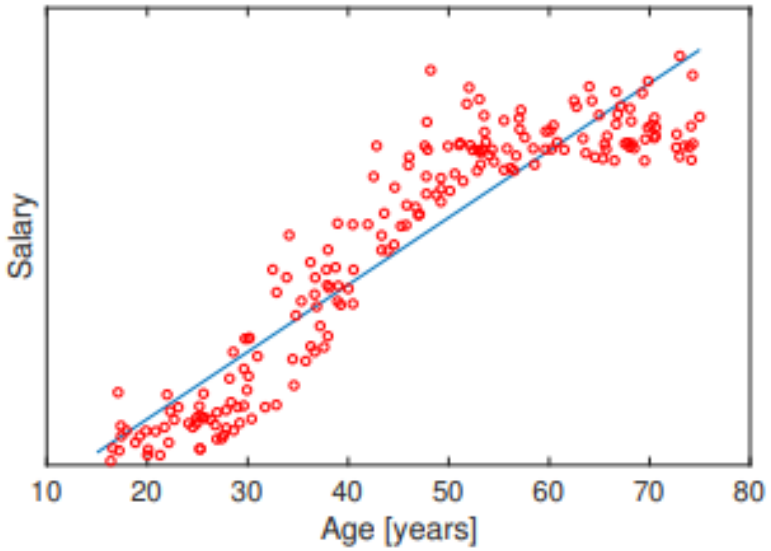
Models allow us to generate multiple shapes by tuning their parameters. Degrees of freedom or the complexity of a model describe its ability to generate different shapes, e.g., its flexibility.

The degrees of freedom of a model are in general related to the number of parameters of the model:

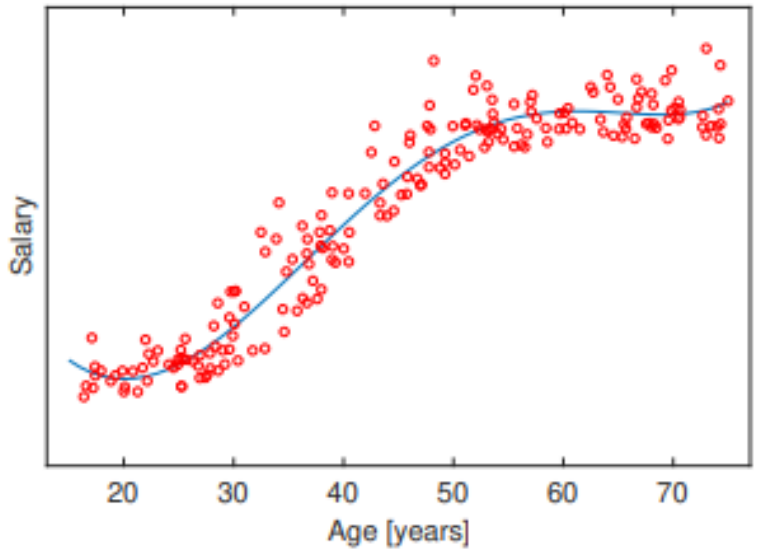
- A Linear Model has two parameters and is inflexible, as it generates only a straight line.
- A Polynomial Model has multiple parameters and is more flexible than a Linear one.

The flexibility of a model is related to its **interpretability** and **accuracy** and there is a trade-off between the two.

Interpretability

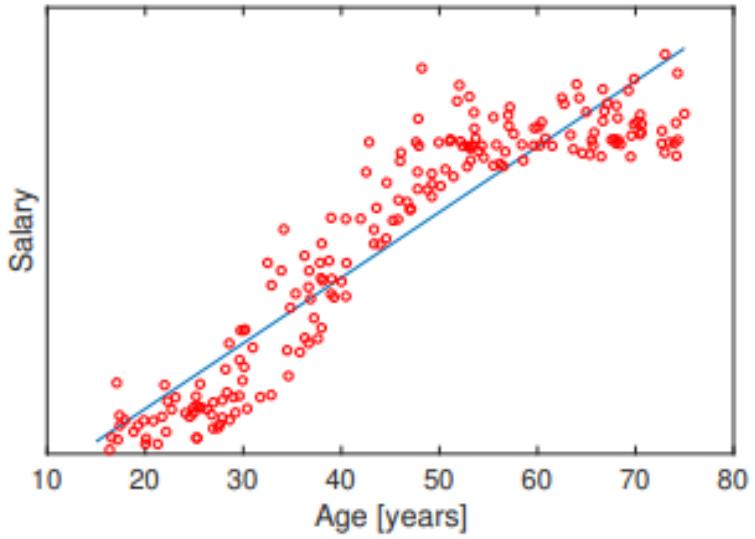


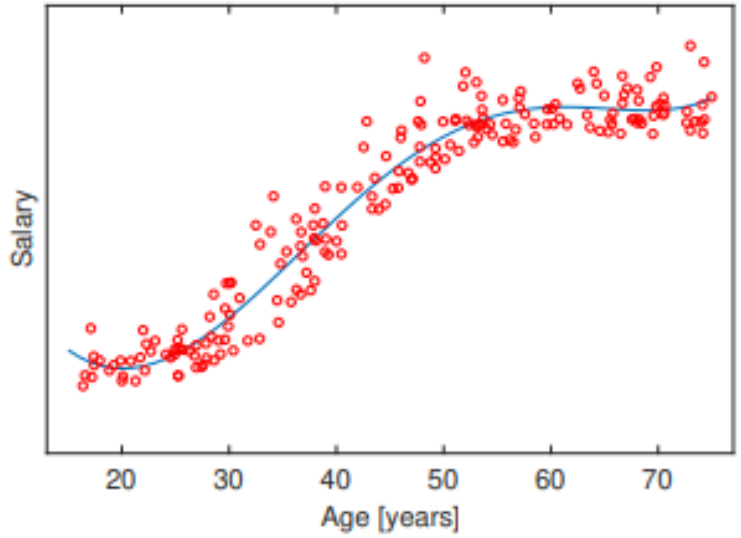
According to this linear model, the older you get, the more money you make



According to this polynomial model, our salary remains the same as teenagers, then increases between our 20s and 50s, then...

Accuracy



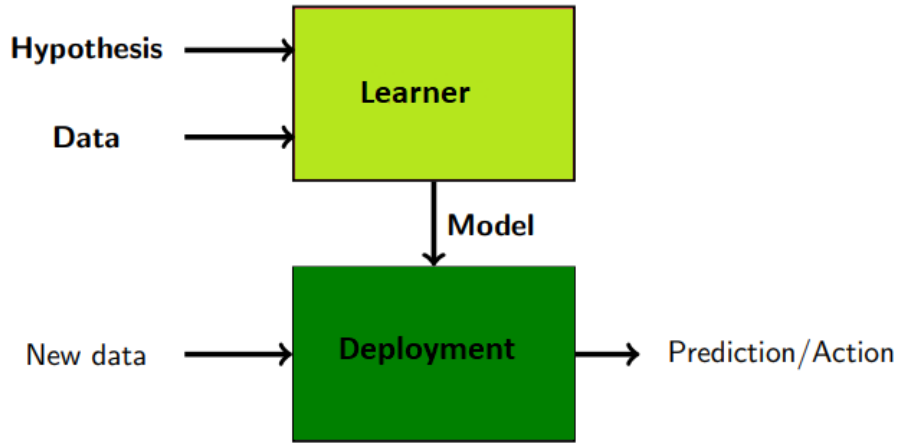


The training error of the best linear model is E_{MSE}= 0.0983

The training error of the best polynomial model is E_{MSE} = 0.0379

Generalization

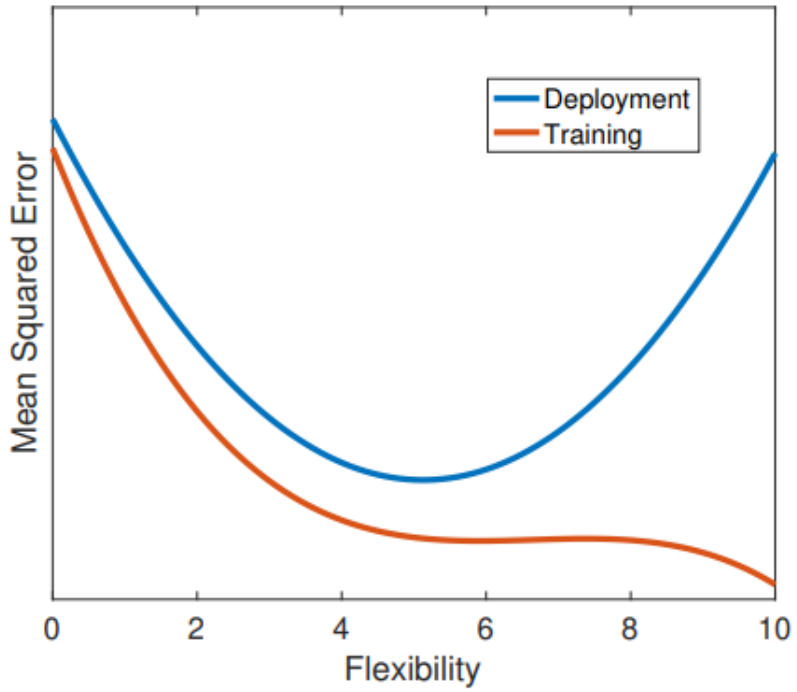
Assume, we are considering “training” MSE, means quality of Regression models on training dataset.



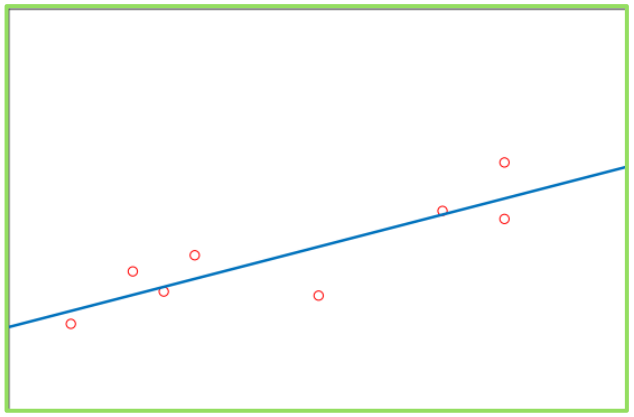
Will our model work well during deployment, when presented with new data?

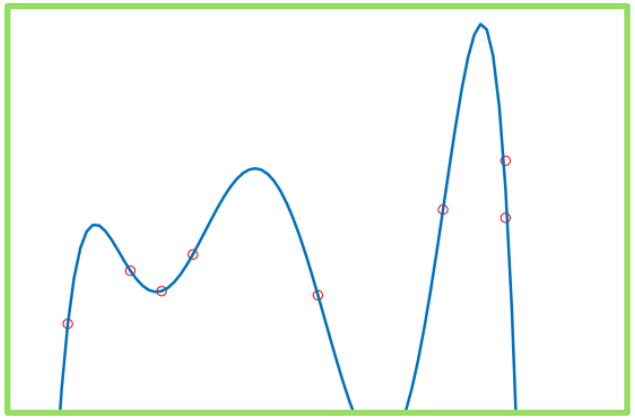
Generalization

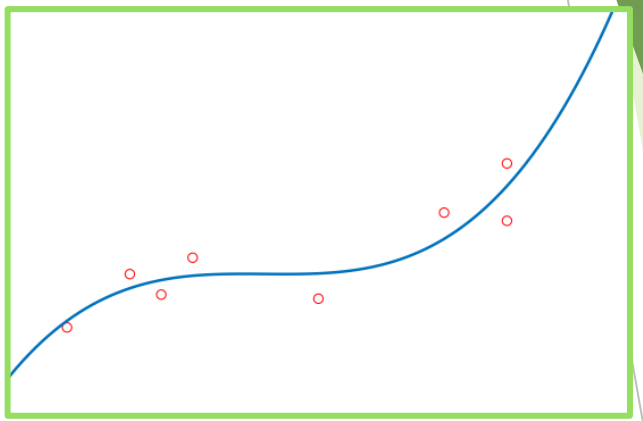
In this figure, the red curve represents the training MSE of different models of increasing complexity, whereas the blue curve represents the deployment MSE for the same models. What's happening?



Underfitting, Overfitting and Right Fitting







Underfitting: Large training and deployment errors are produced. The model is unable to reflect the underlying pattern. Too rigid models lead to underfitting

Overfitting: Small errors are produced during training, large errors during deployment. The model is memorizing irrelevant details. Too complex models and not enough data lead to overfitting.

Just right: Low training and deployment errors. The model is capable of reproducing the underlying pattern and ignores irrelevant details

Other Models for Regression

- Exponential
- Sinusoids
- Radial Basis Functions
- Splines
- And Many more..

The mathematical Formulation s identical and only the expression for $f(.)$ changes.

Other Quality Metrics

In addition to the MSE, we can consider other quality metrics:

- Root mean squared error: Measures the sample standard deviation of the prediction error.

$$E_{RMSE} = \sqrt{\frac{1}{N}\sum e_i^2}$$

- Mean absolute error: Measures the average of the absolute prediction error.

$$E_{MAE} = \frac{1}{N}\sum |e_i|$$

- R-squared. Measures the proportion of the variance in the response that is predictable from the predictors.

$$E_R = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2}, \text{ where } \bar{y} = \frac{1}{N}\sum y_i$$

Important Notes about Regression

- Regression is a model that predicts continuous model
- Among several candidate models/solutions that we produce by using the training set, we identify the best solution
- Models have different degrees of Flexibility
- The final quality of a model can be seen during deployment, and we need models capable of generalizing
- Three terms describe the ability of models to generalize:
  - Underfitting: unable to describe the underlying pattern
  - Overfitting: memorization of irrelevant details
  - Just right: reflects underlying pattern and ignores irrelevant details


Do you have any
QUESTIONS?