School of Electronic Engineering and Computer Science
Queen Mary University of London

ECS607/766 Data Mining
Week 2: Regression

Dr Jesús Requena Carrión

5 Oct 2018

Agenda

Recap (with some extras)

Formulation of regression problems

Basic models

Interpretability, accuracy and generalisation

Final remarks

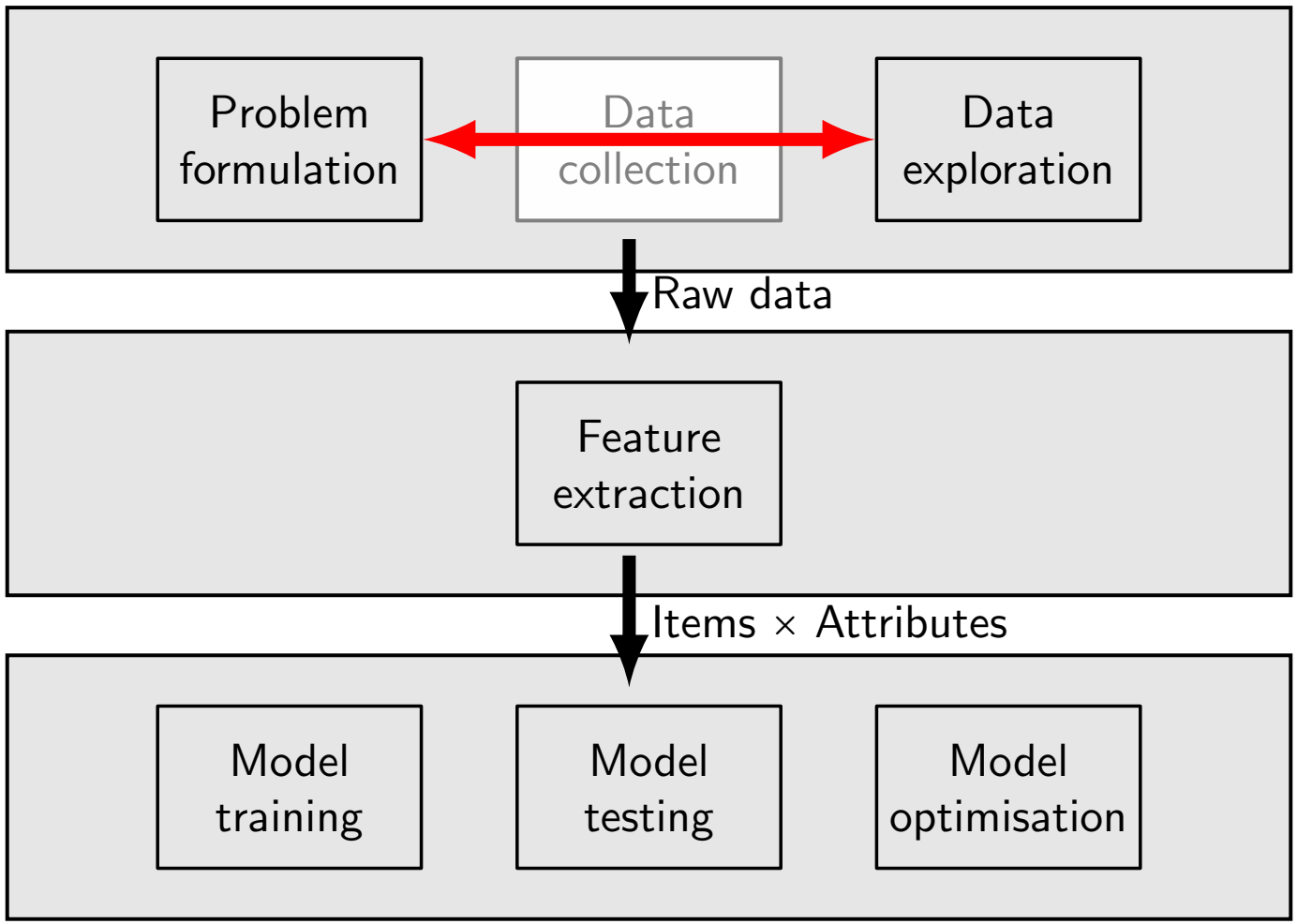
Data Mining

Data Mining is the art of extracting **knowledge** (in our case, in the form of a mathematical or computer model) from **existing data** (anything that has been recorded).

The term **mining** suggests that

- Knowledge is of **value**, specifically the models we build can be put into production and generate revenue.
- Data **already exists** somewhere waiting for us to find its hidden value, so we don't need to collect it.

A Data *Mining* pipeline



The 1936 Literary Digest poll

Alfred Landon
Republican Party

Franklin D. Roosevelt
Democratic Party

- The Literary Digest conducted one of the largest and most expensive polls ever conducted (around 2.4 million people)
- Predicted Landon would get 57 % of the vote, Roosevelt 43 %
- The actual results of the election were 62 % for Roosevelt against 38 % for Landon (19 % error, the largest ever)
- The cause: **Bad sampling**, 10 million names were taken from telephone directories, club membership lists, magazine subscribers lists, etc. The poll suffered from **selection** and **nonresponse bias**

Know thy data!

Agenda

Recap (with some extras)

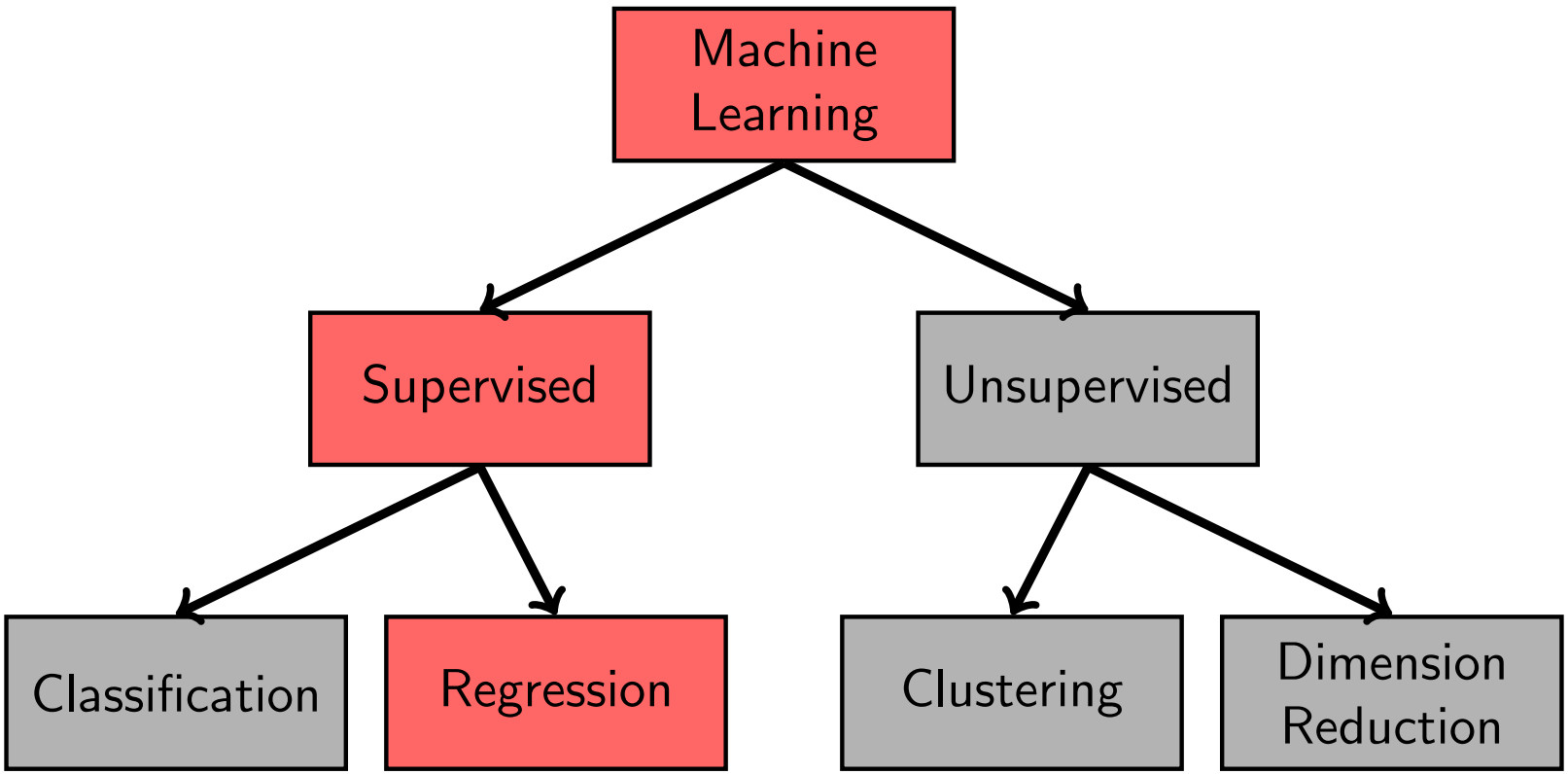
Formulation of regression problems

Basic models

Interpretability, accuracy and generalisation

Final remarks

Data science taxonomy



The dataset

Datasets can be represented as tables, where **rows correspond to instances** (a.k.a. *samples*) and **columns to attributes** (a.k.a. *features*). Therefore, number of rows corresponds to the number of instances, and the number of columns to the number of attributes.

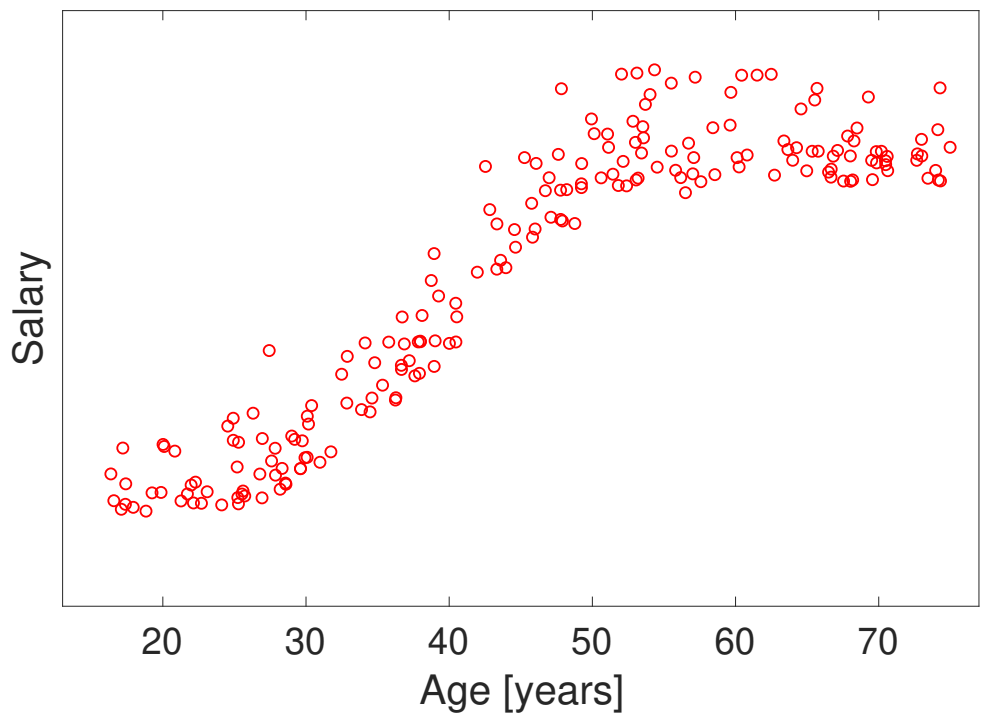
The first 5 instances of a dataset recording the age and salary of a group of people are shown below in a table form:

	Age	Salary
S_1	18	12000
S_2	37	68000
S_3	66	80000
S_4	25	45000
S_5	26	30000
...	...	...

The dataset

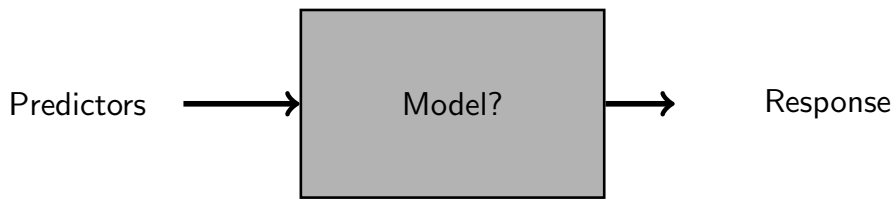
Datasets can also be represented as graphs. For every instance, the value of one feature is represented against the value of the other instances in a cartesian coordinate system.

The figure below shows 200 instances of the dataset recording the age and salary of a group of people.



Problem formulation

- **Supervised**: Our starting point is the assumption that the value of one of the attributes of our dataset (the response) can be predicted based on the value of the remaining attributes (the predictors).
- The response is a **continuous variable**.
- Our job is then to **find the** *best* **model** that relates the response attribute to the predictors.



Predictors and responses

	Age	Salary
S_1	18	12000
S_2	37	68000
S_3	66	80000
S_4	25	45000
S_5	26	30000
...	...	...

In this dataset:

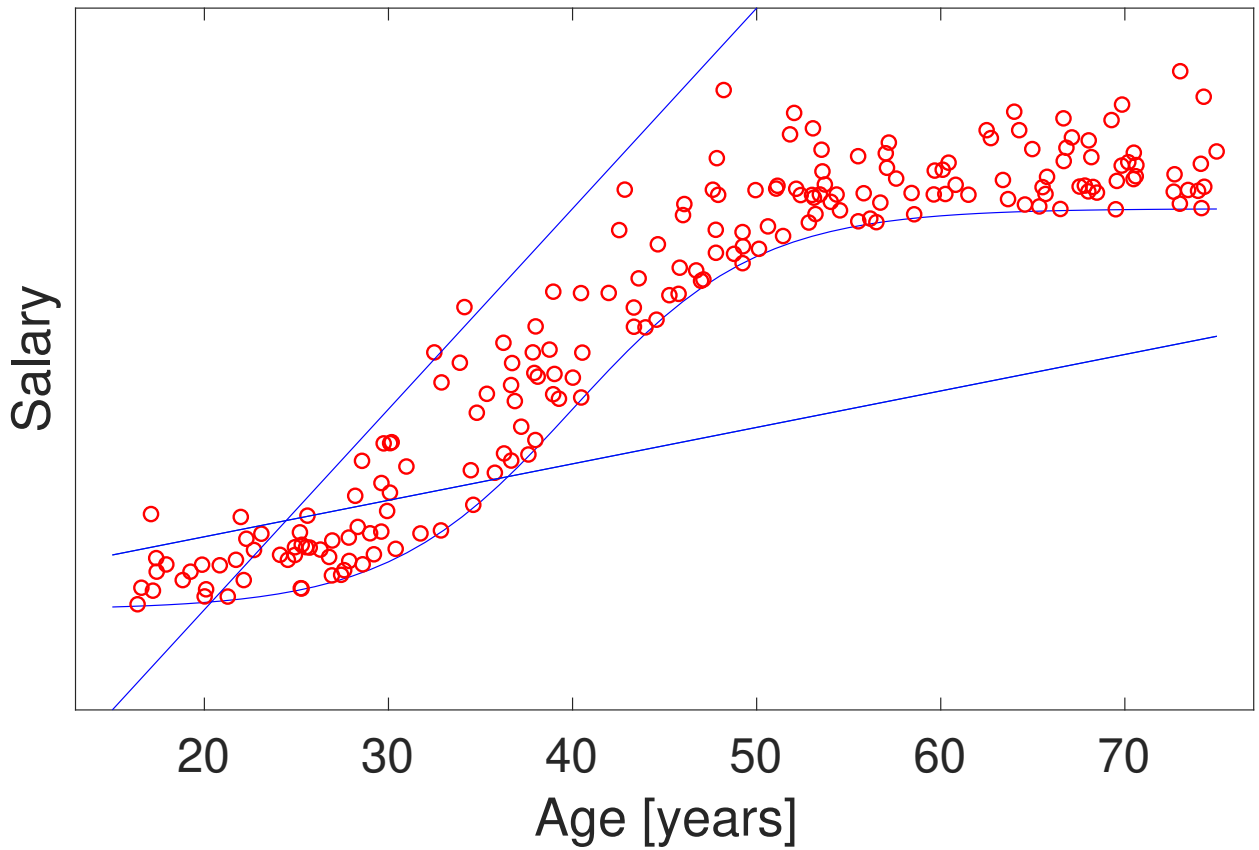
(a) *Age* is the predictor, *Salary* is the response

(b) *Salary* is the predictor, *Age* is the response

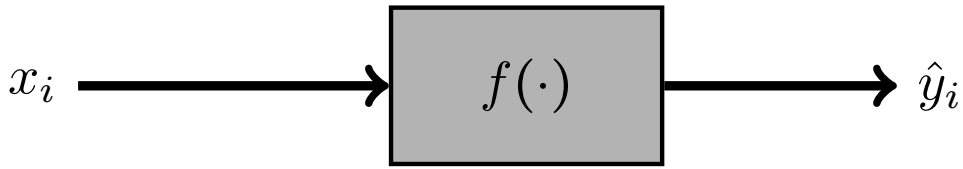
(c) Both options can be considered

Examples of candidate solutions for our dataset

Which one is the best?



Mathematical notation



Dataset:

- N is the number of instances in our dataset
- i identifies one of the instances
- x_i is the predictor of instance i
- y_i is the desired response of instance i
- The dataset is $\{(x_i, y_i) : 1 \leq i \leq N\}$

Model:

- $f(\cdot)$ denotes our model
- $\hat{y}_i = f(x_i)$ is the response produced by $f(\cdot)$ when the predictor is x_i
- $e_i = y_i - \hat{y}_i$ is the prediction error for instance i

Model evaluation: Goodness of fit

In order for us to find the **best** model we need to have a quantitative notion of **goodness**. One popular option is the **mean square error** (MSE), which is defined as follows:

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

Reading maths

$\gamma\nu\omega\theta\iota\ \sigma\epsilon\alpha\upsilon\tau\sigma\nu$

$$E_{MSE} = \frac{1}{N}\sum_{i=1}^{N} e_i^2$$

gnothi seauton

E_{MSE} is equal to 1 over N times the summation from i equals 1 to N of e sub i

know thyself

E_{MSE} is the average squared error

Problem formulation as an optimisation problem

Given a dataset $\{(x_i, y_i) : 1 \leq i \leq N\}$, each possible model f has a corresponding E_{MSE}. Our problem is then to **find the model f with the lowest MSE as computed from our dataset**.

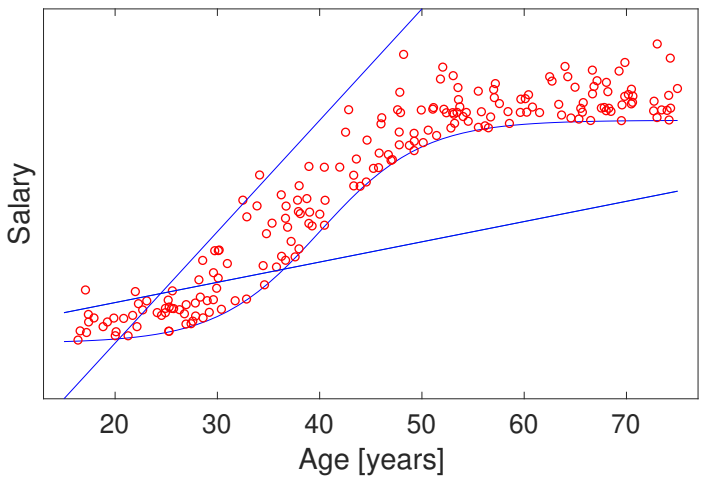
$$\arg\min_f \frac{1}{N} \sum_{i=1}^{N} (y_i - f(x_i))^2$$

This is one type of **optimisation problem** and the solution is called the **minimum mean square error** (MMSE).

The question is, how do we find such model? Do we need to explore every possible model or should we restrict ourselves to a limited collection of candidate models?

A zero-error model

Given a dataset, is it possible to find a model whose whose error is zero, $E_{MSE} = 0$? In other words, is it possible to find a model such that $\hat{y}_i = y_i$ for every instance i in the dataset?



(a) **Never**, there will always be a non-zero error

(b) It is **never guaranteed**, but might be possible for some datasets

(c) **Always**, there will always be a complex enough model achieving this

The nature of the error

Consider this: There might be two people with the same age and different salaries. Our model produces one and only one salary for each age. Therefore, our model will never be error-free, as it will never be able to predict both salaries (other factors which determine the salary of a person are unaccounted for, such as education, profession, etc)

In general, when considering a regression problem we need to be aware that:

- The chosen predictors might not reflect all the factors determining a response
- A chosen model might not be able to represent accurately the *true* relationship between response and predictor
- Random mechanisms might be involved too

We represent this discrepancy mathematically as

$$y = f(x) + e$$

Agenda

Recap (with some extras)

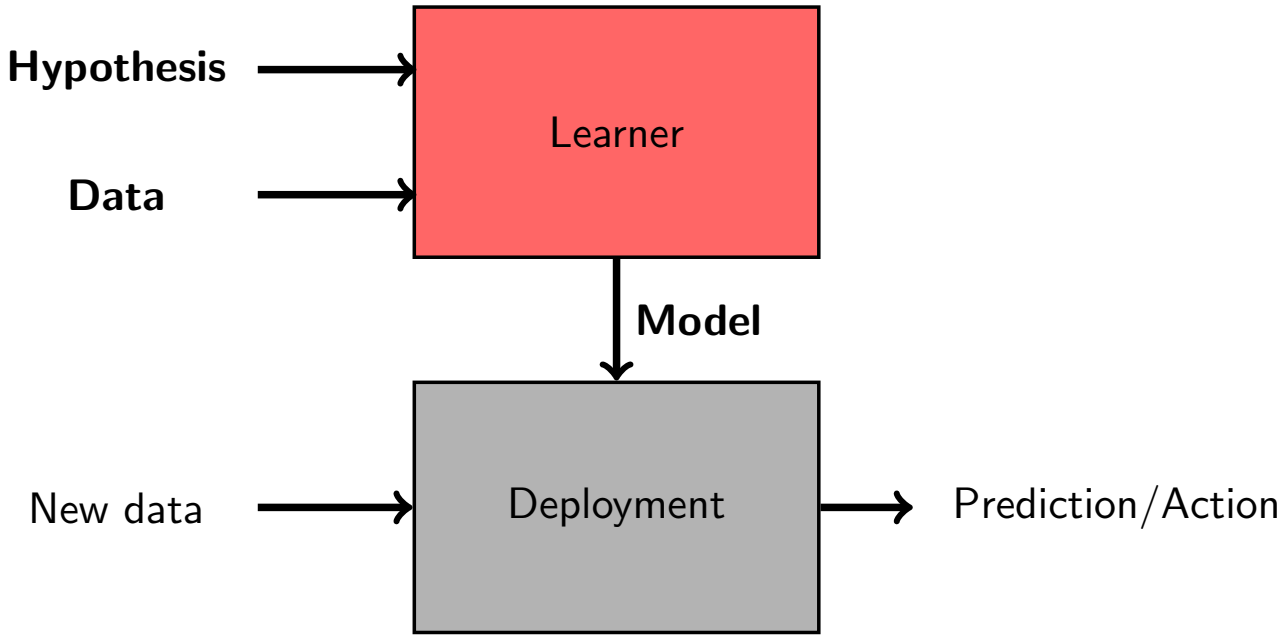
Formulation of regression problems

Basic models

Interpretability, accuracy and generalisation

Final remarks

Our regression learner



- **Hypothesis**: Type of model (linear, polynomial, etc). **Data exploration** can help us choose the type of model
- **Data**: Collection of instances. In its simplest form, each instance has two attributes (one **predictor** and one **response**, continuous)
- **Model**: Produces one response as an output for a given predictor. It is determined by searching (**optimisation**) the best model (**evaluation**) within the type of model considered.

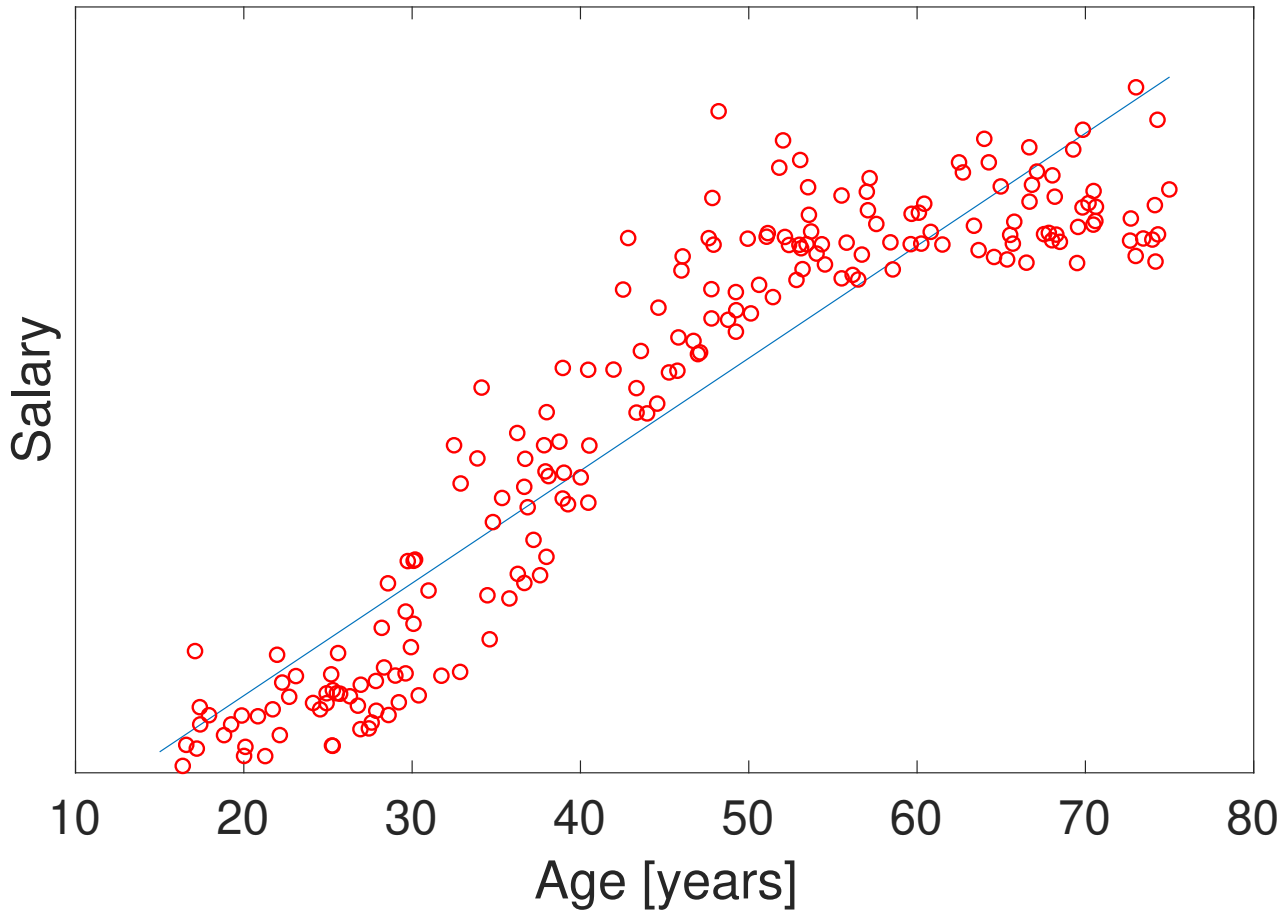
Simple linear regression

In linear regression, models of the form

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

are considered. Linear models are therefore *parametric* and have two parameters w_1 and w_0, which need to be *tuned*. The *best* values for w_1 and w_0 are the ones that minimise the MSE.

MMSE linear solution to our toy dataset



Multivariate linear regression

In multivariate regression, there are two or more predictors for a given response. For instance, in a database containing attributes such as age, gender, education and salary, we could use the attributes age, gender and education as predictors and salary as a response.

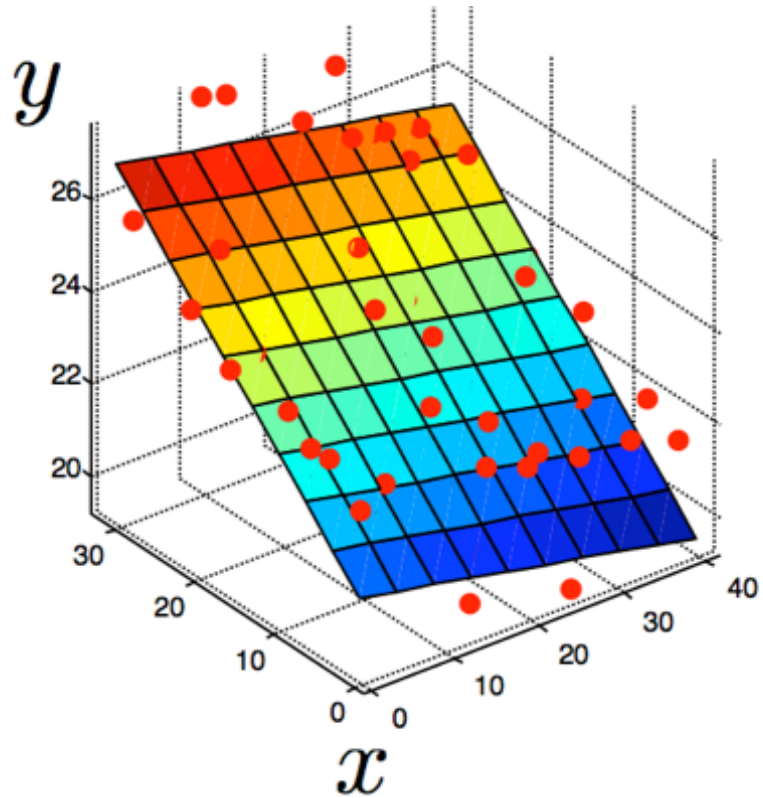
A linear multivariate model has the form:

$$f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x}$$

where $\boldsymbol{w}$ is the model's parameter vector and vector $\boldsymbol{x}$ contains multiple predictors (and a constant, usually 1).

In the simple linear regression model, $\boldsymbol{w} = [w_1, w_0]^T$ and $\boldsymbol{x} = [x, 1]^T$.

Multivariate linear regression



The MMSE solution for linear models

If we define X as the matrix containing the predictor values for every instance i, $\boldsymbol{x}_i$, and Y as the vector containing the responses for every instance, y_i, it can be shown that the model that minimises the MSE as calculated from the dataset is the one with the parameter vector

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

This is an **exact solution**. In general, there will not be an analytical expression that allows us to calculate the parameters of a model given a notion of goodness. In these scenarios, we will need to use other optimisation approaches.

Polynomial regression

Polynomial regression uses non-linear models in which the predictor is raised to a power. The general form of a polynomial model is:

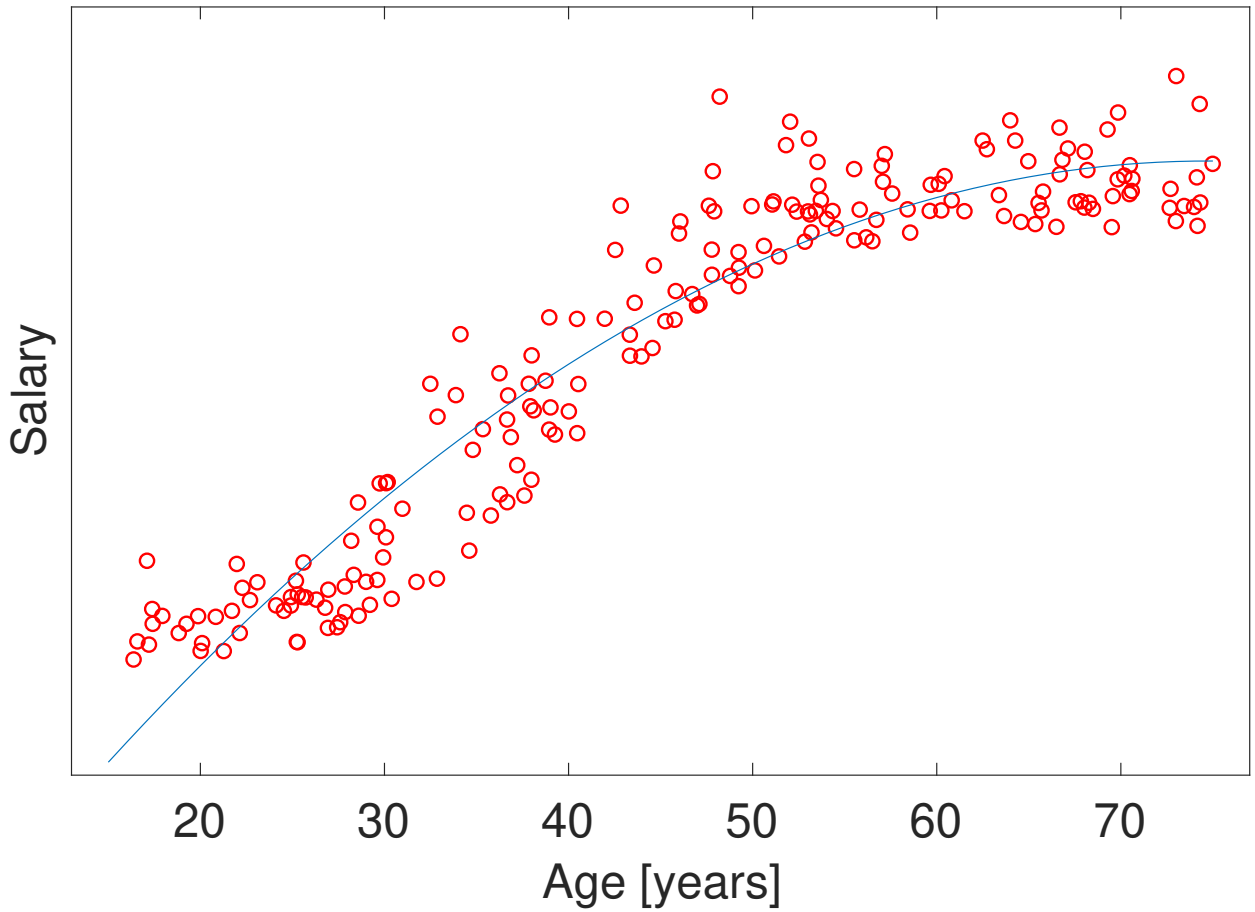
$$f(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + \ldots$$

A polynomial model with one predictor can be expressed as a multivariate linear model by treating the powers of the predictor as predictors themselves. For instance:

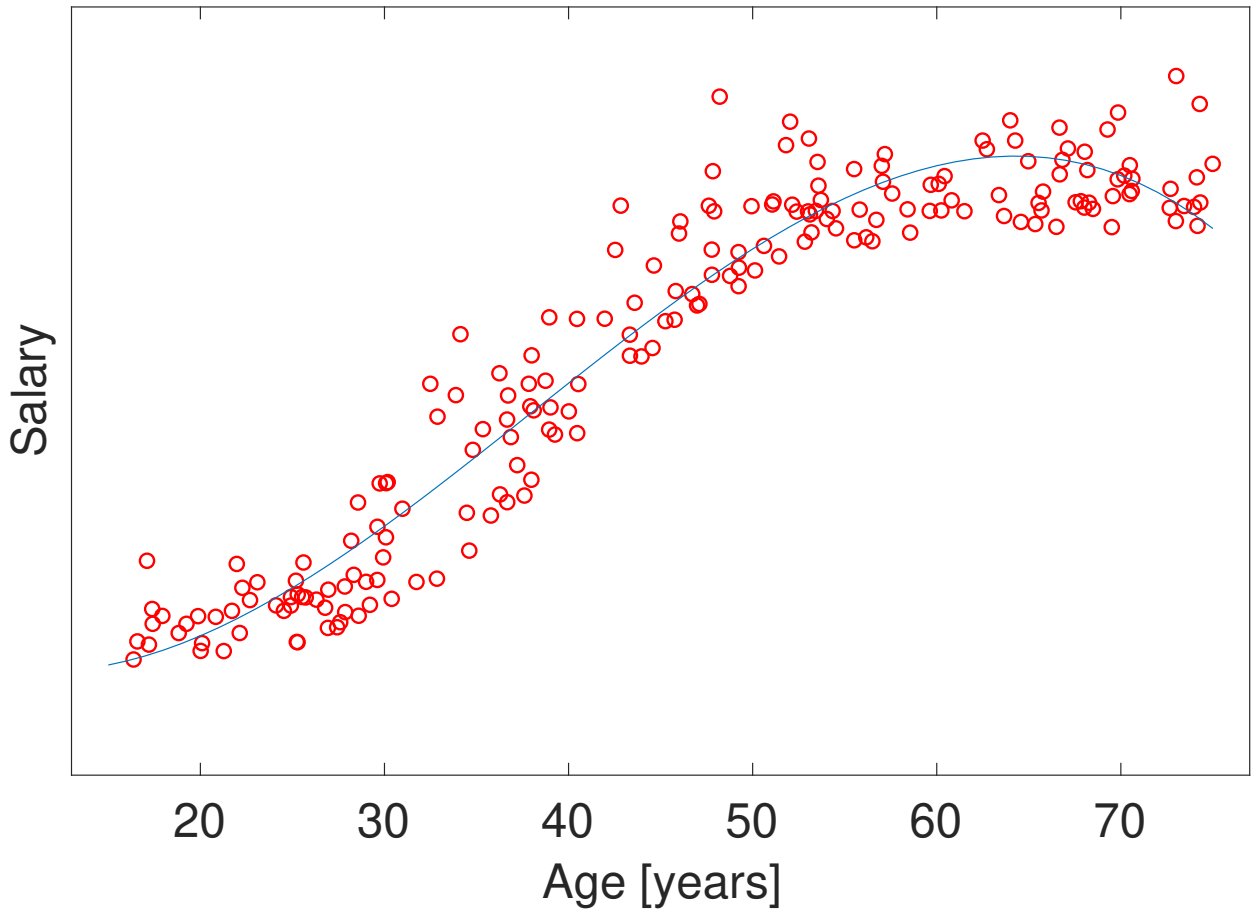
$$f(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 = \boldsymbol{w}^T \boldsymbol{\phi}$$

where $\boldsymbol{\phi} = [1, x, x^2, x^3]^T$. Therefore, there exists an exact MMSE solution (see previous slide).

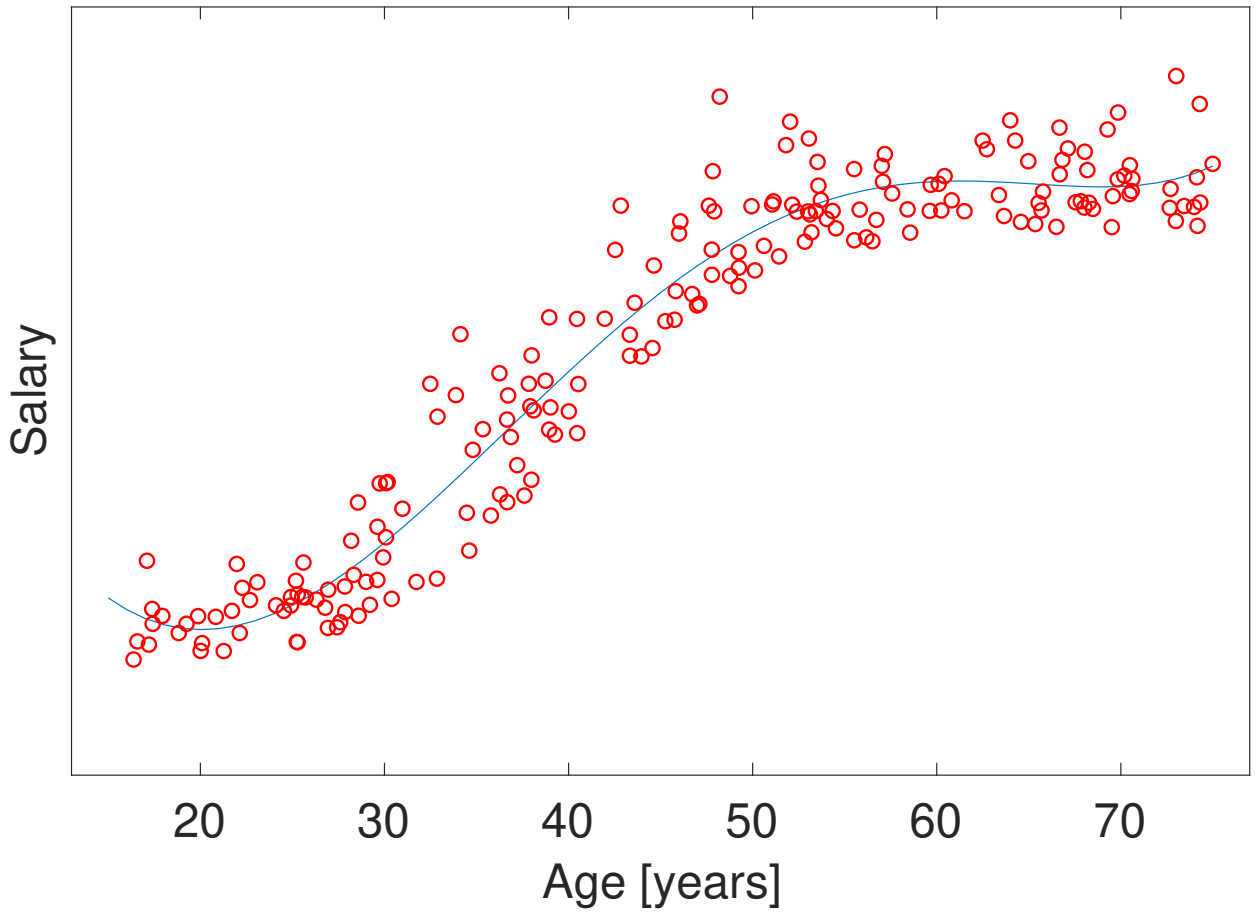
MMSE quadratic solution



MMSE cubic solution



MMSE 5-power solution



Agenda

Recap (with some extras)

Formulation of regression problems

Basic models

Interpretability, accuracy and generalisation

Final remarks

Flexibility

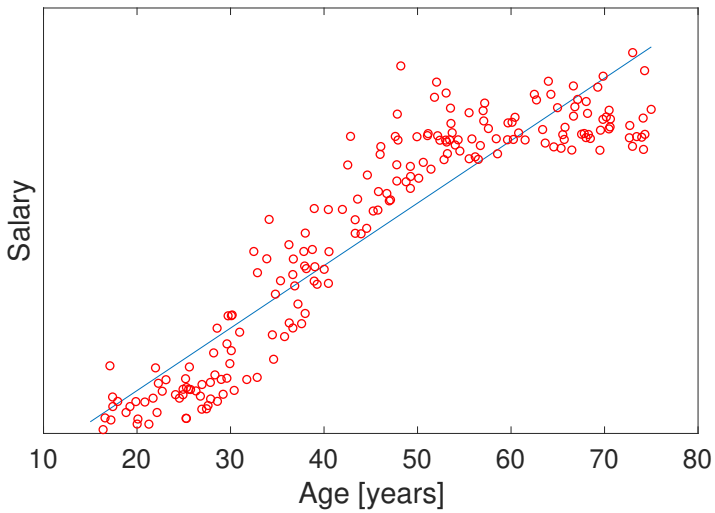
Flexible models allow us to generate multiple shapes by tuning their parameters. Sometimes, we talk about the **degrees of freedom** or **complexity** to describe their flexibility. The degrees of freedom of a model are in general related to their number of parameters.

- Linear models are inflexible, as they can only generate straight lines. They have only 2 parameters.
- Cubic models are more flexible and are characterised by 4 parameters.

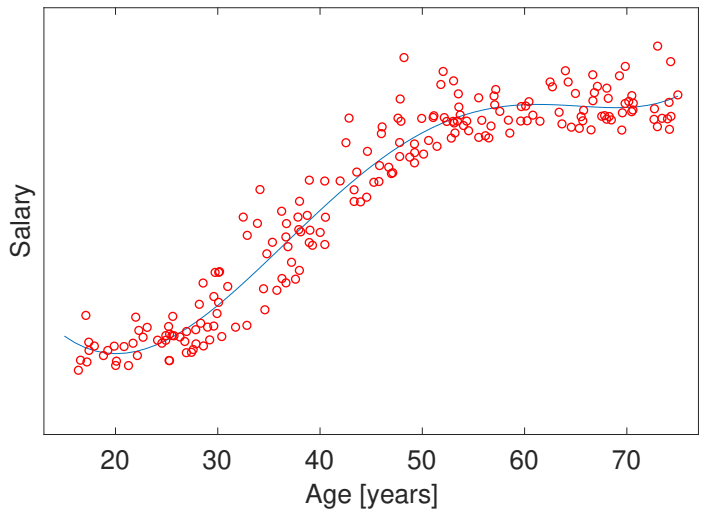
The flexibility of a model is related to their **interpretability** and **accuracy** and there is a trade-off between the two.

Interpretability

Model interpretability is crucial for us, as humans, to understand in a qualitative manner how a predictor determines a response. Inflexible models produce solutions that are usually simpler and more interpretable.



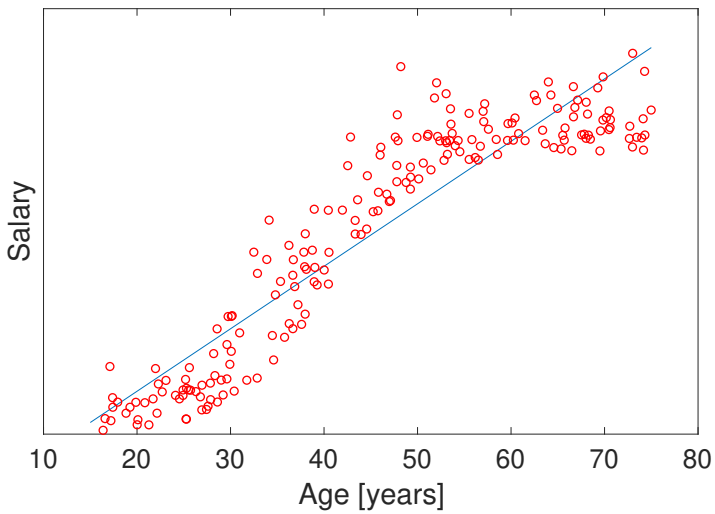
According to this linear model, the older you get, the more money you make



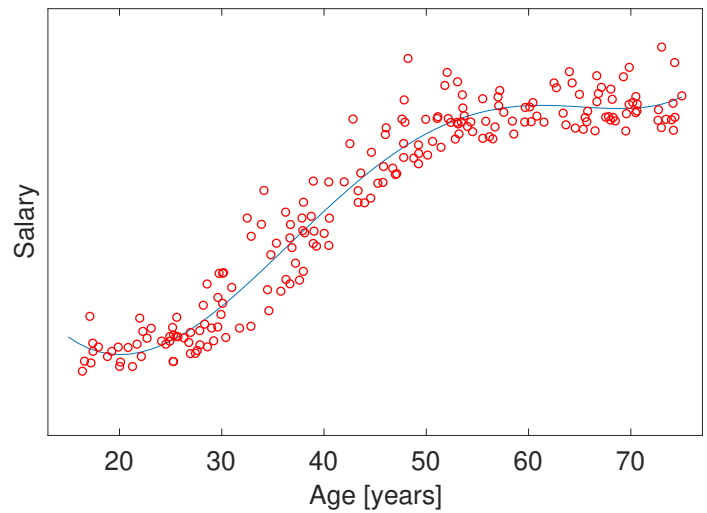
According to this polynomial model, our salary remains the same as teenagers, then increases between our 20s and 50s, then...

Accuracy

The accuracy of a model is also related to its flexibility. Flexible models will have in general lower MSE than inflexible models.



The error for the MMSE linear model is $E_{MSE} = 0.0983$



The error for this MMSE polynomial model is $E_{MSE} = 0.0379$

Generalisation

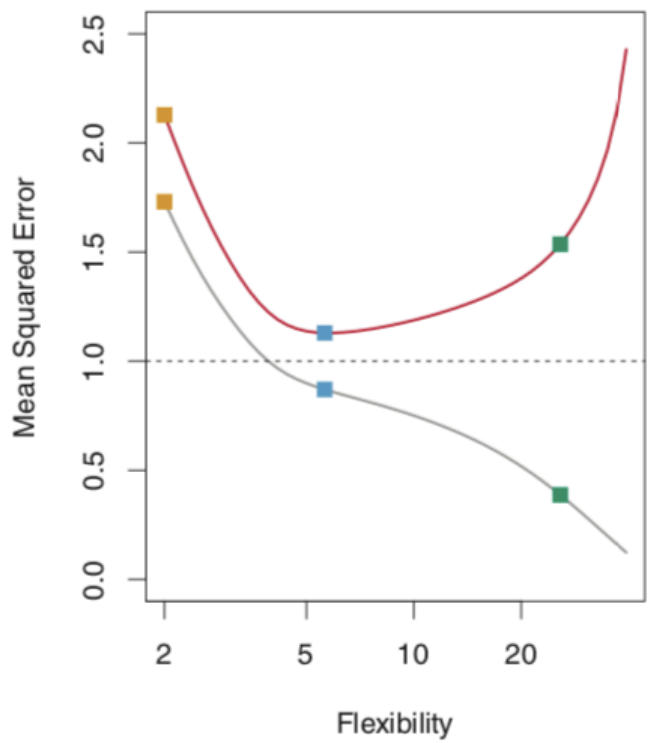
As data miners, we will use a dataset to build a model that will be deployed. Hence, our aim is not to create a model that works well for our dataset, but during production.

So far, we have assessed the quality of our model by computing the E_{MSE} on our dataset (which we have already used to build the model). How can we be sure our model works well during production?

Generalisation is the ability of our model to work well in production, in other words, to successfully **translate what we have learnt during the learning stage to the production stage**.

Generalisation

In the following figure, the grey curve represents the MSE as calculated from our dataset, whereas the red curve represents the MSE during production. What's happening?



Taken from *An Introduction to Statistical Learning* by G. James et al.

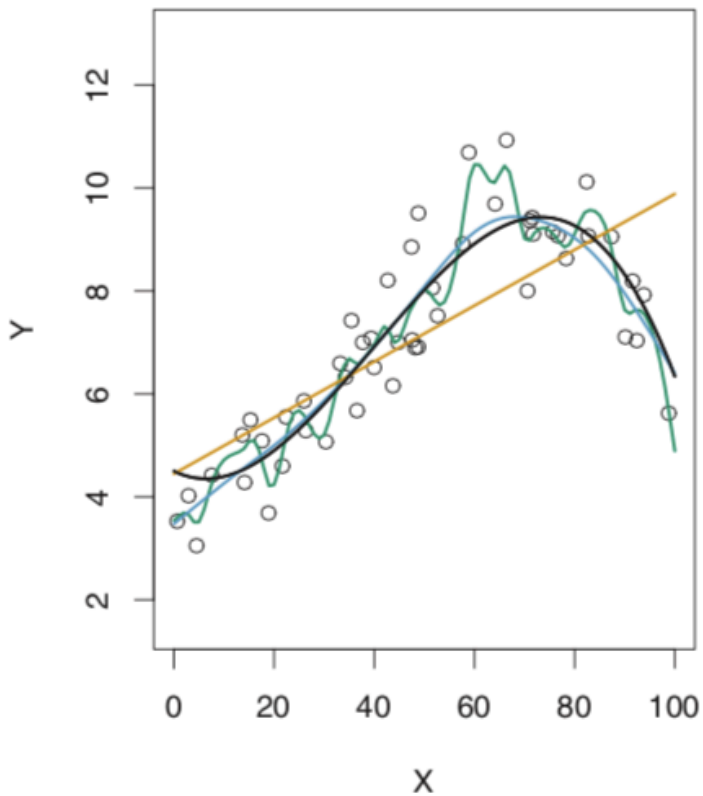
Training and test

In order to evaluate the quality of our model, we split our dataset into two disjoint datasets, the **training** dataset and the **test dataset**:

- The training dataset is used to build our model
- The test dataset is used to evaluate its performance

Overfitting

Our model is **overfitting** when it produces small training MSE and large test MSE and its causes are too complex models and too little data. By contrast, **underfitting** is characterised by large training and test MSE.



Taken from *An Introduction to Statistical Learning* by G. James et al.

Who is overfitting / underfitting in the following figure?

(a) Green is underfitting, black is overfitting

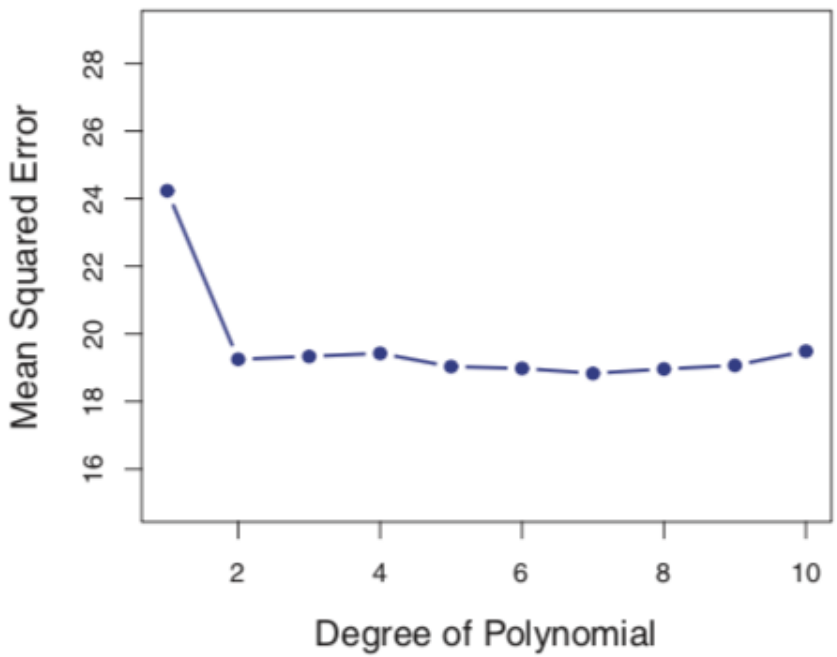
(b) Yellow is underfitting, black is overfitting

(c) Yellow is underfitting, green is overfitting

Validation

As data miners we have a wide range of models to chose from. Take the family of polynomial models, which degree should we choose?

Validation is a procedure that allows us to choose the complexity of our model: we split our dataset into a training and a validation dataset. The validation MSE is used to determine the model complexity.



Regularisation

Regularisation is a procedure for reducing the risk of model overfitting. The idea is to create a goodness function that penalises large coefficients in a linear model $f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x}$:

$$E_{MSER} = \frac{1}{N} \sum_{i=1}^{N} e_i^2 + \lambda w^T w$$

The solution w that minimises E_{MSER} is

$$\boldsymbol{w} = \left(X^T X + N\lambda I \right)^{-1} X^T Y$$

When $\lambda = 0$, we obtain the solution for the usual MSE problem. As λ increases, the complexity of the resulting solution decreases and so does the risk of overfitting. However, notice that the parameter λ still needs to be determined (by using some form of validation).

Agenda

Recap (with some extras)

Formulation of regression problems

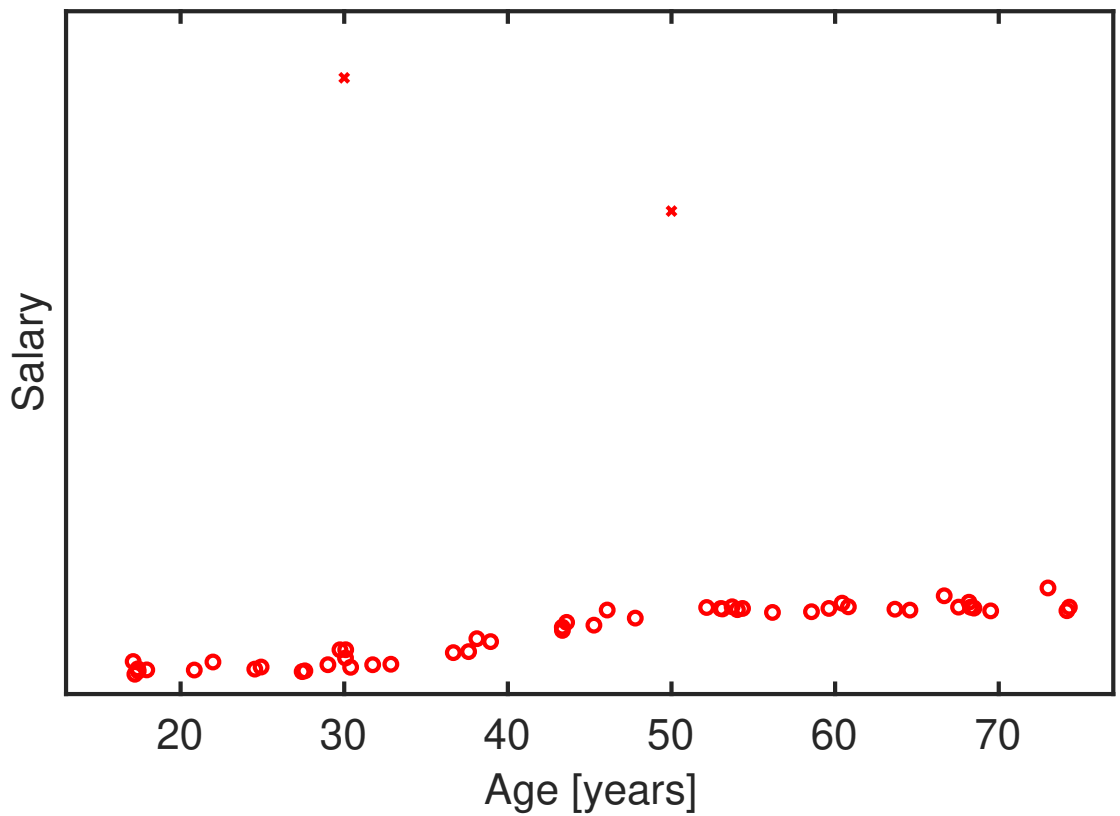
Basic models

Interpretability, accuracy and generalisation

Final remarks

Outliers

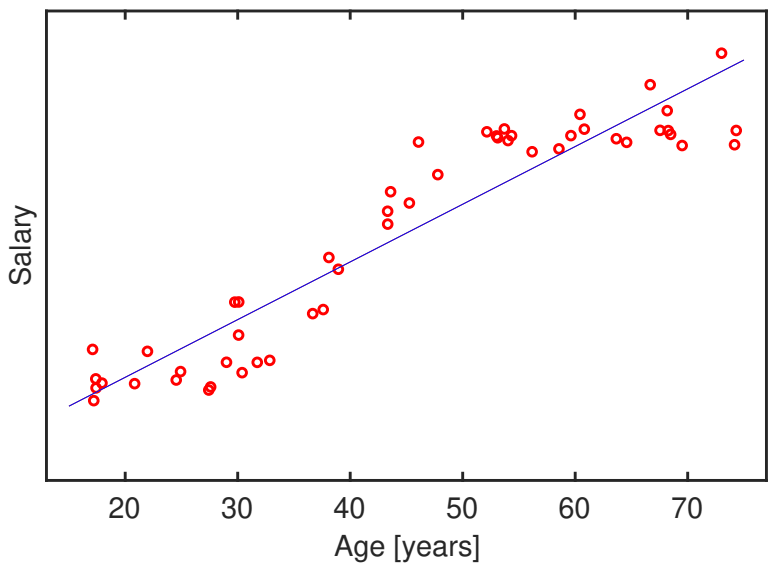
Outliers are instances in our dataset that do not follow the general pattern of your dataset. In a graph representation, they can be seen abnormally distant for the rest of the samples.



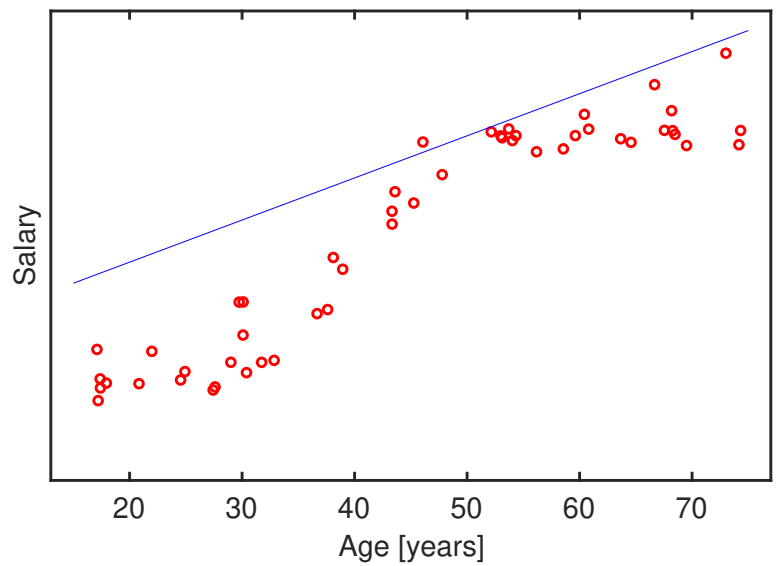
Outliers

Outliers can have a big impact on our solution and therefore, they are identified and eliminated during the stage of data exploration.

Without outliers



With outliers



Other evaluations

In addition to the MSE, there are other quality measures:

- 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.

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

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

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

Other models

Other models that can be used include:

- Exponential
- Sinusoids
- Neural networks
- Radial basis functions
- Splines
- ...

Do we always need data-driven approaches?

Think about this: Would you use a data-driven approach to build a model that predicts the distance driven by a car at a constant speed during a given time interval?