

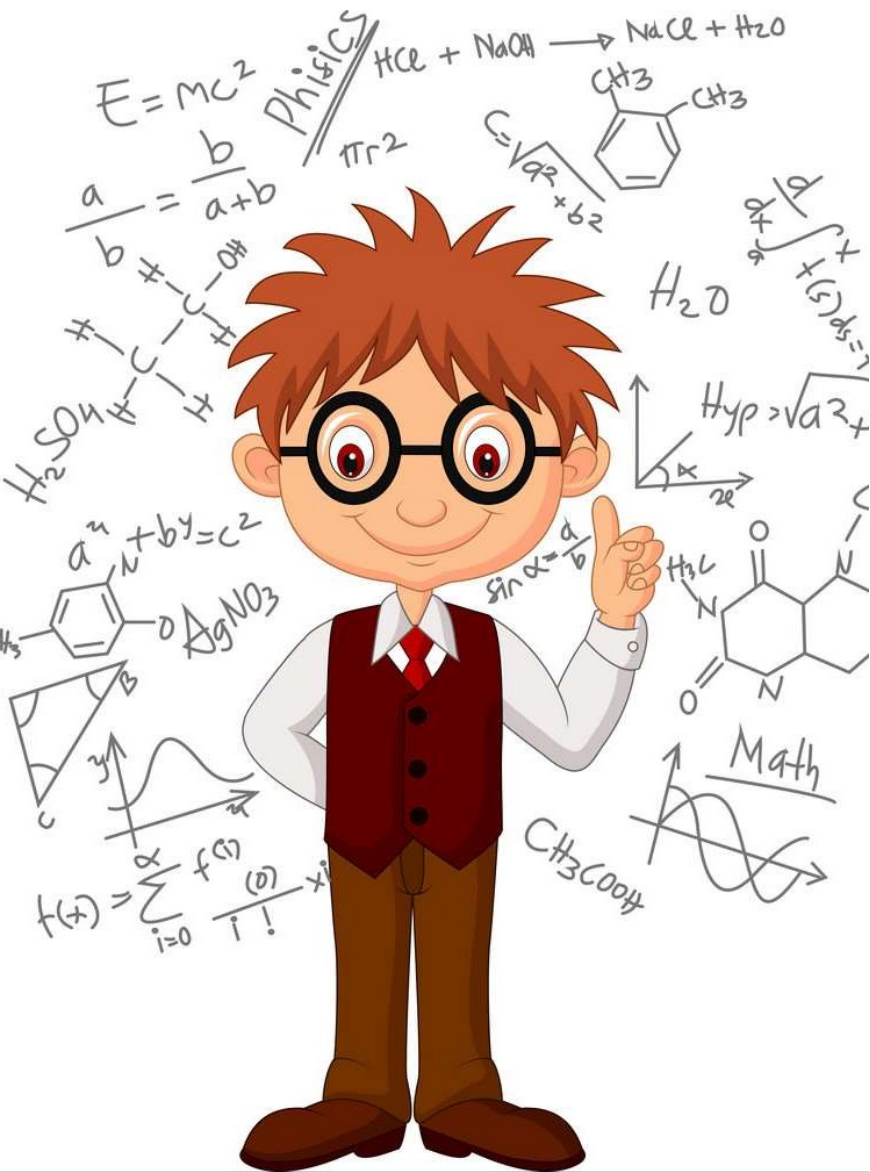
Predictive Data Analysis

Lecturer: Marina Iantorno

E-mail: miantorno@cct.ie

July 2022





In today's class we will cover:

- ☐ Polynomial Regression Model
- ☐ Practice in Python

Regression Models

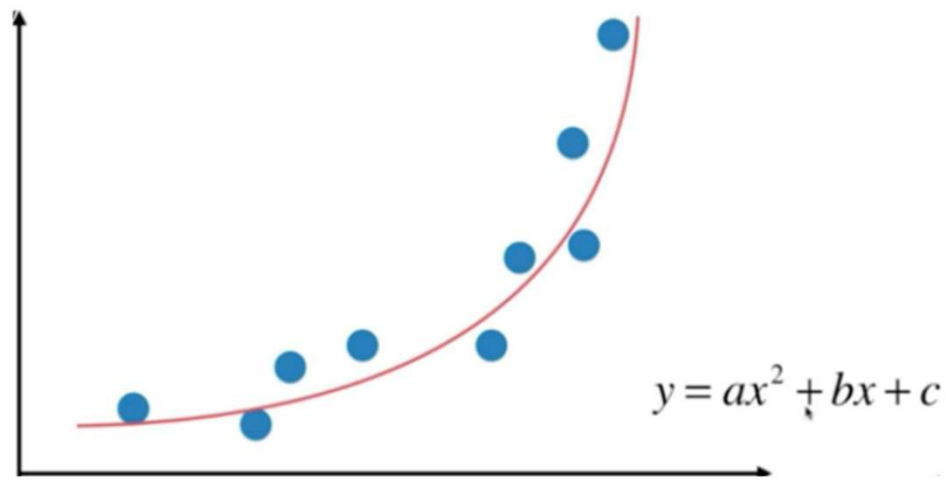
POLYNOMIAN REGRESSION



Polynomial Regression Model

Sometimes, we discard the multiple linear regression because there is no linear correlation between the variables, however, it could be that another kind of connection exists. Here is when the Polynomial regression appears. The most general example could be associated with a quadratic function, which has only one variable on the power of another one.

Polynomial Regression Model



As it happens with the other models, the datapoints are approximately drawing a curve with a parabolic effect due to the X squared, and the one that we can see in the graph is the curve that better fits with these observations.

Polynomial Regression Model

This kind of model is commonly used for those events that shouldn't be considered linear. For example:

- How does certain disease spread?
- How does a pandemic spread?
- Immigration vs population and resources.
- Territory vs population.

Polynomial Regression Model

We should keep in mind that the function of this model is a polyonomy, and therefore we should get something as follow:

$$y = b_0 + b_1x_1 + b_2x_1^2 + \dots + b_nx_1^n$$

In the end, our goal is still to find the coefficients to be able to predict y based on what happens to the different independent variables.

Polynomial Regression Model

Suppose that we are working in the HR department of a company, and there is a candidate who is very experienced and we want to hire, but when the moment to talk about salary arises, the candidate says that he expects to earn \$160,000 yearly. We ask him why and he says that this was his salary at the previous job. We will use then a Polynomial Regression Model to predict the salary of this candidate in his previous job.

Polynomial Regression Model

We collected data to proceed. We looked on Job websites for the company where the candidate worked, and we collect the data for different positions, from Business Analyst to the CEO. This is the data that is in our dataset called “Position_Salaries.csv”.

Position	Level	Salary
Business Analyst	1	45000
Junior Consultant	2	50000
Senior Consultant	3	60000
Manager	4	80000
Country Manager	5	110000
Region Manager	6	150000
Partner	7	200000
Senior Partner	8	300000
C-level	9	500000
CEO	10	1000000

Polynomial Regression Model

Suppose that this person was a Regional Manager for two years, his salary should be between \$150,000 and \$200,000.

Position	Level	Salary
Business Analyst	1	45000
Junior Consultant	2	50000
Senior Consultant	3	60000
Manager	4	80000
Country Manager	5	110000
Region Manager	6	150000
Partner	7	200000
Senior Partner	8	300000
C-level	9	500000
CEO	10	1000000

We will consider that the person worked on a level between 6 and 7, so we will consider that he was on the level 6.5

Polynomial Regression Model

What are our steps then?

1. Data pre-processing.
2. Build the Polynomial Regression Model.
3. Train the model to understand the correlation levels between the positions and the salaries.
4. Make the prediction

Polynomial Regression Model

Something important to consider in this case, is that we will not split the data into the training set and test set in this occasion during the Data Pre-Processing. We have just a few observations. Besides, we want to get the prediction straight away for the level between 6 and 7, and we need to use as much data as possible to get an accurate result.

Polynomial Regression Model

Another observation is that we don't really need to do the encoding step, because the categorical variables are represented by levels, therefore the second column could be considered as the data already encoded. We will know the position according to the level.

Level	Salary
1	45000
2	50000
3	60000
4	80000
5	110000
6	150000
7	200000
8	300000
9	500000
10	1000000


Polynomial Regression Model

Let's try it in Python!

Polynomial Regression Model

Some important outputs

We need to build the model, so we could think of our model as follow:

$$y = b_0 + b_1x_1 + b_2x_1^2 + \dots + b_nx_1^n$$


Salary Position Level

THAT'S ALL FOR TODAY

THANK YOU

