Polynomial regression

Polynomial regression

Types of Polynomials

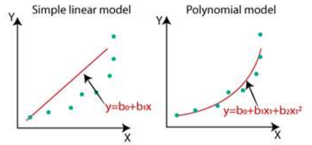
Linear —— $ax + b = 0$

Quadratic —— $ax^2 + bx + c = 0$

Cubic —— $ax^3 + bx^2 + cx + d = 0$

Polynomial regression

Why do we need Polynomial Regression?
Let's consider a case of ***Simple Linear Regression***.
•We make our model and find out that it performs very badly,
•We observe between the actual value and the best fit line, which we predicted and it seems that the actual value has some kind of curve in the graph and our line is no where near to cutting the mean of the points.
•This where polynomial Regression comes to the play, it predicts the best fit line that follows the pattern(curve) of the data, as shown in the pic below:



Polynomial regression

What is Polynomial Regression?

- Polynomial Regression is a form of. *regression analysis in which the relationship between the independent variables and dependent variables are modeled in the nth degree polynomial.*
- Polynomial Regression models are usually fit with the method of **least squares**. The least square method **minimizes the variance of the coefficients**, under the Gauss Markov Theorem.
- Polynomial Regression is a special case of Linear Regression where we *fit the polynomial equation on the data with a curvilinear relationship between the dependent and independent variables*.
- A Quadratic Equation is a Polynomial Equation of 2nd Degree. However,this degree can increase to nth values.

Polynomial regression

Assumptions of Polynomial Regression:
•The behavior of a dependent variable can be explained by a linear, or curvilinear, additive relationship between the dependent variable and a set of *k* independent variables *(xi, i=1 to k)*.

•The relationship between the dependent variable and any independent variable is linear or curvilinear (specifically polynomial).

•The independent variables are independent of each other.

•The errors are independent, normally distributed with mean zero and a constant variance (*OLS*).

Polynomial regression

- Polynomial Regression *does not require the relationship between the independent and dependent variables to be linear in the data set*, This is also one of the main difference between the Linear and Polynomial Regression.
- **Polynomial Regression is generally used when the points in the data are not captured by the Linear Regression Model and the Linear Regression fails in describing the best result clearly.**

As we increase the degree in the model, it tends to increase the performance of the model. However,increasing the degrees of the model also increases the risk of over-fitting and under-fitting the data.

Polynomial regression

How to find the right degree of the equation?
In order to find the right degree for the model to prevent over-fitting or under-fitting, we can use:

1.Forward Selection:
This method increases the degree until it is significant enough to define the best possible model.

2.Backward Selection:
This method decreases the degree until it is significant enough to define the best possible model.

Polynomial regression

Solution:
Let the quadratic polynomial regression model be
y=a0+a1*x+a2*x2
The values of a0, a1, and a2 are calculated using the following system of equations:

$$\sum y_i = na_0 + a_1\left(\sum x_i\right) + a_2\left(\sum x_i^2\right)$$
$$\sum y_i x_i = a_0\left(\sum x_i\right) + a_1\left(\sum x_i^2\right) + a_2\left(\sum x_i^3\right)$$
$$\sum y_i x_i^2 = a_0\left(\sum x_i^2\right) + a_1\left(\sum x_i^3\right) + a_2\left(\sum x_i^4\right)$$

Polynomial regression

Solving this system of equations we get
a_0=12.4285714
a_1=-5.5128571
a_2=0.7642857
The required quadratic polynomial model is
y=12.4285714 -5.5128571 * x +0.7642857 * x^2

Polynomial regression

Let there be only one independent variable x and the relationship between x, and dependent variable y, be modeled as,
y=a0+a1*x+a2*x2 +………+ +an*xn
for some positive integer n >1, then we have a polynomial regression.

Problem Definition:
Find a quadratic regression model for the following data:

X	Y
3	2.5
4	3.2
5	3.8
6	6.5
7	11.5

Polynomial regression

First, we calculate the required variables and note them in the following table.

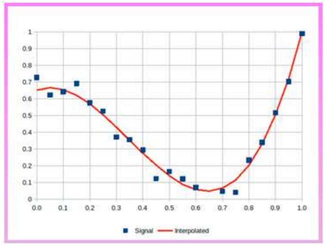
x	y	x2	x3	x4	y*x	y*x2
3	2.5	9	27	81	7.5	22.5
4	3.2	16	64	256	12.8	51.2
5	3.8	25	125	625	19	95
6	6.5	36	216	1296	39	234
7	12	49	343	2401	80.5	563.5
Σ 25	27.5	135	775	4659	158.8	966.2

Using the given data we,

$$27.5 = 5a_0 + 25a_1 + 135a_2$$
$$158.8 = 25a_0 + 135a_1 + 775a_2$$
$$966.2 = 135a_0 + 775a_1 + 4659a_2$$

Polynomial regression

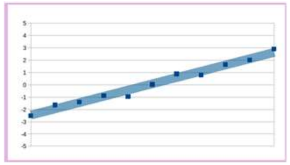
Simply put polynomial regression is an attempt to create a polynomial function that approximates a set of data points. This is easier to demonstrate with a visual example.



Polynomial regression

Linear regression is polynomial regression of degree 1, and generally takes the form
$y = m\,x + b$ where m is the slope, and b is the y-intercept.

It could just as easily by written $f(x) = c_0 + c_1 x$ with c_1 being the slope and c_0 the y-intercept.



Polynomial regression

We need a way of calculating coefficients that takes all our data points into consideration. This is where least squares come in.

To preform a least square approach we need to define a **residual** function.

For any point (x_i, y_i) on the line, the residual would be:

$$r_i(x_i) = y_i - (mx_i + b)$$

Where i is the index into the set of known data points.

We can generalize it for any function:

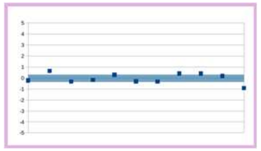
$$r_i(x_i) = y_i - f(x_i)$$

Polynomial regression

- Polynomial regression is one of several methods of **curve fitting**.
- With polynomial regression, the data is approximated using a polynomial function.
- A polynomial is a function that takes the form $f(x) = c_0 + c_1 x + c_2 x^2 \cdots c_n x^n$
  where n is the degree of the polynomial and c is a set of coefficients.

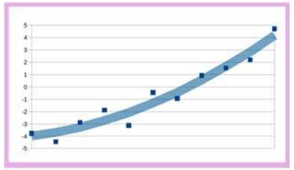
A polynomial of degree 0 is just a constant because $f(x) = c_0 x^0 = c_0$

Likewise preforming polynomial regression with a degree of 0 on a set of data returns a single constant value. It is the same as the mean average of that data. This makes sense because the average is an approximation of all the data points.



Polynomial regression

Quadratic regression is a 2nd degree polynomial and not nearly as common. Now the regression becomes non-linear and the data is not restricted to straight lines.



Polynomial regression

What does the residual tell us?
Well, for each data point the *residual denotes the amount of error between our estimation and the true data*.

How about the overall error?
To get this, we could just *sum up the error for all data points*.
However error can be *positive or negative*.
So we could end up with a lot of error uniformly distributed between negative and positive values.

$$r(x) = \sum_{i=0}^{n} \left[y_i - f(x_i)\right]^2$$

So now that we have a function that measures residual, what do we do with it?
Well, if we are trying to produce a function that models a set of data, we want the residual to be as small as possible—we want to **minimize** it.

Polynomial regression

Let's try using a 2nd degree polynomial to get the results for quadratic regression.

First, the quadratic function:

$$y = c_0 + c_1 x + c_2 x^2$$

Now the sum of squares for n known data points:

$$r(x) = \sum_{i=0}^{n} (c_0 + c_1 x_i + c_2 x^2 - y_i)^2$$