. Linear Regression is a **Linear Model**. Which means, we will establish a linear relationship between the input variables(**X**) and single output variable(**Y**). When the input(**X**) is a single variable this model is called **Simple Linear Regression** and when there are mutiple input variables(**X**), it is called **Multiple Linear Regression**.

**Regression**

Regression analysis is one of the most important fields in statistics and machine learning. There are many regression methods available. Linear regression is one of them.

What Is Regression?

Regression searches for relationships among variables.

For example, you can observe several employees of some company and try to understand how their salaries depend on the **features**, such as experience, level of education, role, city they work in, and so on.

This is a regression problem where data related to each employee represent one **observation**. The presumption is that the experience, education, role, and city are the independent features, while the salary depends on them.

Similarly, you can try to establish a mathematical dependence of the prices of houses on their areas, numbers of bedrooms, distances to the city center, and so on.

Generally, in regression analysis, you usually consider some phenomenon of interest and have a number of observations. Each observation has two or more features. Following the assumption that (at least) one of the features depends on the others, you try to establish a relation among them.

In other words, **you need to find a function that maps some features or variables to others sufficiently well**.

The dependent features are called the **dependent variables**, **outputs**, or **responses**.

The independent features are called the **independent variables**, **inputs**, or **predictors**.

Regression problems usually have one continuous and unbounded dependent variable. The inputs, however, can be continuous, discrete, or even categorical data such as gender, nationality, brand, and so on.

It is a common practice to denote the outputs with 𝑦 and inputs with 𝑥. If there are two or more independent variables, they can be represented as the vector 𝐱 = (𝑥₁, …, 𝑥ᵣ), where 𝑟 is the number of inputs.

When Do You Need Regression?

Typically, you need regressio

### When Do You Need Regression?

Typically, you need regression to answer whether and how some phenomenon influences the other or **how several variables are related**. For example, you can use it to determine ifand to what extent the experience or gender impact salaries.

Regression is also useful when you want **to forecast a response** using a new set of predictors. For example, you could try to predict electricity consumption of a household for the next hour given the outdoor temperature, time of day, and number of residents in that household.

Regression is used in many different fields: economy, computer science, social sciences, and so on. Its importance rises every day with the availability of large amounts of data and increased awareness of the practical value of data.

## Linear Regression

Linear regression is probably one of the most important and widely used regression techniques. It’s among the simplest regression methods. One of its main advantages is the ease of interpreting results.

### Problem Formulation

When implementing linear regression of some dependent variable 𝑦 on the set of independent variables 𝐱 = (𝑥₁, …, 𝑥ᵣ), where 𝑟 is the number of predictors, you assume a linear relationship between 𝑦 and 𝐱: 𝑦 = 𝛽₀ + 𝛽₁𝑥₁ + ⋯ + 𝛽ᵣ𝑥ᵣ + 𝜀. This equation is the **regression equation**. 𝛽₀, 𝛽₁, …, 𝛽ᵣ are the **regression coefficients**, and 𝜀 is the **random error**.

Linear regression calculates the **estimators** of the regression coefficients or simply the **predicted weights**, denoted with 𝑏₀, 𝑏₁, …, 𝑏ᵣ. They define the **estimated regression function** (𝐱) = 𝑏₀ + 𝑏₁𝑥₁ + ⋯ + 𝑏ᵣ𝑥ᵣ. This function should capture the dependencies between the inputs and output sufficiently well.

The **estimated** or **predicted response**, (𝐱ᵢ), for each observation 𝑖 = 1, …, 𝑛, should be as close as possible to the corresponding **actual response** 𝑦ᵢ. The differences 𝑦ᵢ - (𝐱ᵢ) for all observations 𝑖 = 1, …, 𝑛, are called the **residuals**. Regression is about determining the **best predicted weights**, that is the weights corresponding to the smallest residuals.

To get the best weights, you usually **minimize the sum of squared residuals** (SSR) for all observations 𝑖 = 1, …, 𝑛: SSR = Σᵢ(𝑦ᵢ - 𝑓(𝐱ᵢ))². This approach is called the **method of ordinary least squares**.

### Regression Performance

The variation of actual responses 𝑦ᵢ, 𝑖 = 1, …, 𝑛, occurs partly due to the dependence on the predictors 𝐱ᵢ. However, there is also an additional inherent variance of the output.

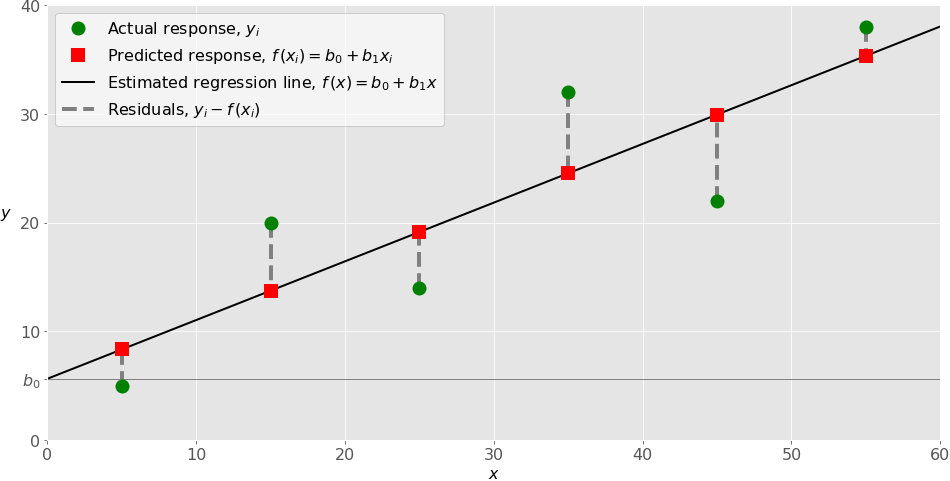
The **coefficient of determination**, denoted as 𝑅², tells you which amount of variation in 𝑦 can be explained by the dependence on 𝐱 using the particular regression model. Larger 𝑅² indicates a better fit and means that the model can better explain the variation of the output with different inputs.

The value 𝑅² = 1 corresponds to SSR = 0, that is to the **perfect fit** since the values of predicted and actual responses fit completely to each other.

### simple Linear Regression

Simple or single-variate linear regression is the simplest case of linear regression with a single independent variable, 𝐱 = 𝑥.

The following figure illustrates simple linear regression:



The estimated regression function (black line) has the equation (𝑥) = 𝑏₀ + 𝑏₁𝑥. Your goal is to calculate the optimal values of the predicted weights 𝑏₀ and 𝑏₁ that minimize SSR and determine the estimated regression function. The value of 𝑏₀, also called the **intercept**, shows the point where the estimated regression line crosses the 𝑦 axis. It is the value of the estimated response (𝑥) for 𝑥 = 0. The value of 𝑏₁ determines the **slope** of the estimated regression line.

The predicted responses (red squares) are the points on the regression line that correspond to the input values. For example, for the input 𝑥 = 5, the predicted response is (5) = 8.33 (represented with the leftmost red square).

The residuals (vertical dashed gray lines) can be calculated as 𝑦ᵢ - (𝐱ᵢ) = 𝑦ᵢ - 𝑏₀ - 𝑏₁𝑥ᵢ for 𝑖 = 1, …, 𝑛. They are the distances between the green circles and red squares. When you implement linear regression, you are actually trying to minimize these distances and make the red squares as close to the predefined green circles as possible.

### Multiple Linear Regression

Multiple or multivariate linear regression is a case of linear regression with two or more independent variables.

If there are just two independent variables, the estimated regression function is (𝑥₁, 𝑥₂) = 𝑏₀ + 𝑏₁𝑥₁ + 𝑏₂𝑥₂. It represents a regression plane in a three-dimensional space. The goal of regression is to determine the values of the weights 𝑏₀, 𝑏₁, and 𝑏₂ such that this plane is as close as possible to the actual responses and yield the minimal SSR.

### Polynomial Regression

You can regard polynomial regression as a generalized case of linear regression. You assume the polynomial dependence between the output and inputs and, consequently, the polynomial estimated regression function.

In other words, in addition to linear terms like 𝑏₁𝑥₁, your regression function 𝑓 can include non-linear terms such as 𝑏₂𝑥₁², 𝑏₃𝑥₁³, or even 𝑏₄𝑥₁𝑥₂, 𝑏₅𝑥₁²𝑥₂, and so on.

The simplest example of polynomial regression has a single independent variable, and the estimated regression function is a polynomial of degree 2: (𝑥) = 𝑏₀ + 𝑏₁𝑥 + 𝑏₂𝑥².

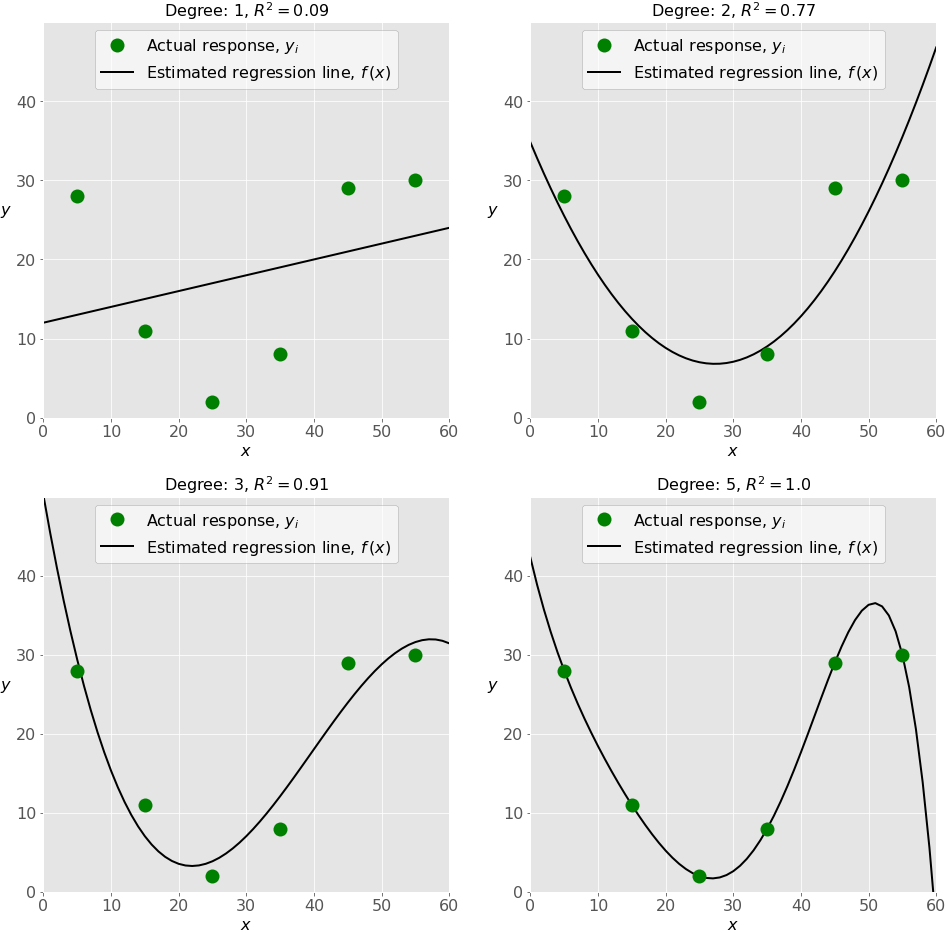
### Underfitting and Overfitting

One very important question that might arise when you’re implementing polynomial regression is related to **the choice of the optimal degree of the polynomial regression function**.

There is no straightforward rule for doing this. It depends on the case. You should, however, be aware of two problems that might follow the choice of the degree: **underfitting** and **overfitting**.

**Underfitting** occurs when a model can’t accurately capture the dependencies among data, usually as a consequence of its own simplicity. It often yields a low 𝑅² with known data and bad generalization capabilities when applied with new data.

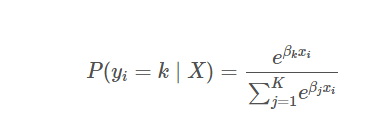
**Overfitting** happens when a model learns both dependencies among data and random fluctuations. In other words, a model learns the existing data too well. Complex models, which have many features or terms, are often prone to overfitting. When applied to known data, such models usually yield high 𝑅². However, they often don’t generalize well and have significantly lower 𝑅² when used with new data.



**Multinomial Logistic Regression**

20 Dec 2017

In multinomial logistic regression (MLR) the logistic function is replaced with a softmax function:



where P(yi=k∣X)P(yi=k∣X) is the probability the iith observation’s target value, yiyi, is class kk, and KK is the total number of classes. One practical advantage of the MLR is that its predicted probabilities using the predict\_proba method are more reliable (i.e. better calibrated).

* **Sigmoid function:** used in the logistic regression model for binary classification.
* **Softmax function:** used in the logistic regression model for multiclassification.

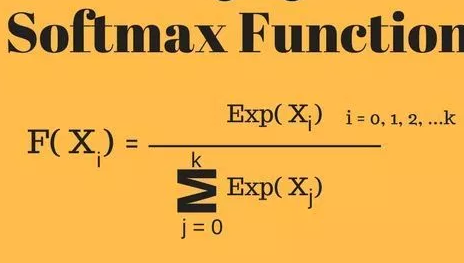
# DIFFERENCE BETWEEN SOFTMAX FUNCTION AND SIGMOID FUNCTION

In mathematical definition way of saying the sigmoid function take any range real number and returns the output value which falls in the range of **0 to 1.**Based on the convention we can expect the output value in the range of **-1 to 1.**

The sigmoid function produces the curve which will be in the Shape **“S.”**These curves used in the statistics too. With the cumulative distribution function (The output will range from 0 to 1)

**Sigmoid Function Usage**

* The Sigmoid function used for **binary classification** in logistic regression model.
* While creating artificial neurons sigmoid function used as the **activation function**.
* In statistics, the **sigmoid function graphs** are common as a cumulative distribution function.



Softmax function calculates the probabilities distribution of the event over ‘n’ different events. In general way of saying, this function will calculate the probabilities of each target class over all possible target classes. Later the calculated probabilities will be helpful for determining the target class for the given inputs.

The main advantage of using Softmax is the output probabilities range. The range will **0 to 1**, and the sum of all the probabilities will be **equal to one**. If the softmax function used for multi-classification model it returns the probabilities of each class and the target class will have the high probability.

The formula computes the **exponential (e-power)** of the given input value and the **sum of exponential values** of all the values in the inputs. Then the ratio of the exponential of the input value and the sum of exponential values is the output of the softmax function.

**Difference Between Sigmoid Function and Softmax Function**

The below are the tabular differences between Sigmoid and Softmax function.

|  |  |  |
| --- | --- | --- |
|  | **Softmax Function** | **Sigmoid Function** |
| 1 | Used for multi-classification in logistic regression model. | Used for binary classification in logistic regression model. |
| 2 | The probabilities sum will be 1 | The probabilities sum need not be 1. |
| 3 | Used in the different layers of neural networks. | Used as activation function while building neural networks. |
| 4 | The high value will have the higher probability than other values. | The high value will have the high probability but not the higher probability. |