Linear Regression

Keywords: combinatorics, probability and statistics, vectors, dot product, linear regression, machine learning, data processing

In practice, we often encounter situations where the values of one variable determine the values of another. Based on a set of measured or statistically obtained data, we then try to find a mathematical model that describes the functional relationship between the two variables. For example, consider data indicating the height and weight of American women aged between 30 and 39 (source: https://en.wikipedia.org/wiki/Simple_linear_regression, accessed April 12, 2024; for brevity, only half of the data is used).

height/m	1.47	1.52	1.57	1.63	1.68	1.73	1.78	1.83
weight/kg	52.21	54.48	57.20	59.93	63.11	66.28	69.92	74.46

The data is displayed in the figure on the left. It is evident from the figure that as height increases, weight also tends to increase. In such a case, it is possible to find a mathematical model that expresses weight as a function of height. Such a mathematical model is shown in color in the figure on the right. It allows us to predict a woman's weight based on her height.

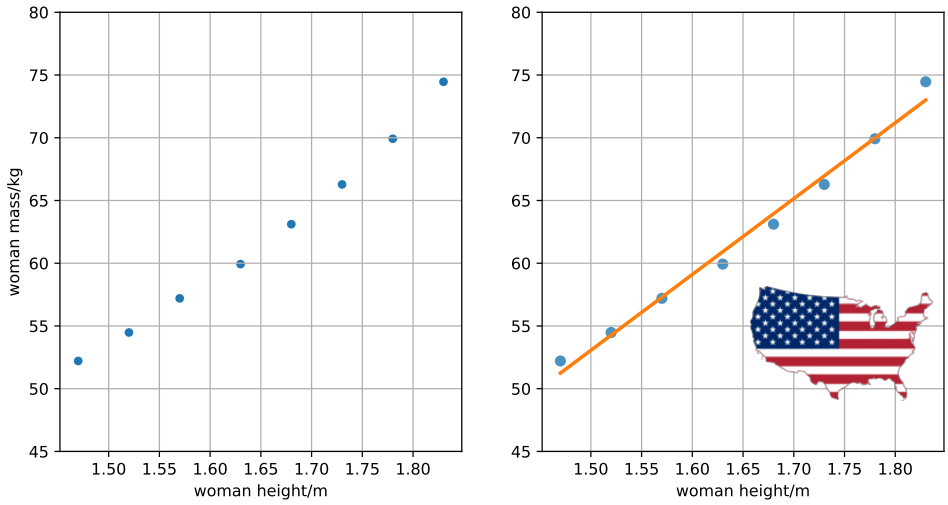

Figure 1: On the left are displayed data showing how the weight of American women depends on their height. On the right, a regression line representing a mathematical model of the functional relationship between height and weight has been added to the data.

The type of problem described above is called *linear regression*.

Linear regression is one of the fundamental methods in *machine learning*. It involves identifying a functional relationship hidden within a set of data. Once we have this relationship, we can use it to predict function values for inputs that are not included in the original dataset.

The portal Math4U is developed within the international projects of the Department of Applied Mathematics (VSB-Technical University of Ostrava) and other institutions in Czechia, Poland, Slovakia, and Spain.
Funded by the European Union

In the following section, we will show how linear regression is related to linear combinations of vectors, and how the regression line can be found using vector operations. We will proceed step by step:

- First, we will review how to express a vector as a linear combination of given vectors.
- Then, we will see how this task can be simplified if one of the vectors is perpendicular to the others.
- Next, we will explore how to find an approximate solution when an exact one does not exist.
- Finally, we will use these ideas to solve the problem of linear regression — that is, we will construct a mathematical model based on the given data that reveals the underlying trend and allows us to make predictions for values not present in the dataset.

Linear Combination of Vectors

Exercise 1. Express the vector $\vec{c} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ as a linear combination of the vectors $\vec{a} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$ and $\vec{b} = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$.

Exercise 2. Express the vector $\vec{w} = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}$ as a linear combination of the vectors

$$\vec{u}_1 = \begin{pmatrix} 2 \\ 2 \\ 1 \end{pmatrix}, \quad \vec{u}_2 = \begin{pmatrix} 3 \\ 1 \\ 2 \end{pmatrix}, \quad \vec{u}_3 = \begin{pmatrix} 3 \\ -1 \\ -4 \end{pmatrix}.$$

Linear Combination Using the Dot Product

If at least one of the given vectors is perpendicular to the remaining vectors, we can use a clever trick to obtain a simpler system of equations.

Let's go back to the previous problem. We can observe that the vector $\vec{u}_3$ is perpendicular to the vectors $\vec{u}_1$ and $\vec{u}_2$. This means it is also perpendicular to the plane defined by these two vectors. We can easily verify this by calculating the dot products:

$$\vec{u}_1 \cdot \vec{u}_3 = 2 \cdot 3 + 2 \cdot (-1) + 1 \cdot (-4) = 0$$

and

$$\vec{u}_2 \cdot \vec{u}_3 = 3 \cdot 3 + 1 \cdot (-1) + 2 \cdot (-4) = 0.$$

Thanks to this property, it is worth multiplying equation (1) using dot product by the vectors $\vec{u}_1$ to $\vec{u}_3$. This gives us the following three equations.

$$\begin{aligned} t_1(\vec{u}_1 \cdot \vec{u}_1) + t_2(\vec{u}_2 \cdot \vec{u}_1) + t_3(\vec{u}_3 \cdot \vec{u}_1) &= \vec{w} \cdot \vec{u}_1 \\ t_1(\vec{u}_1 \cdot \vec{u}_2) + t_2(\vec{u}_2 \cdot \vec{u}_2) + t_3(\vec{u}_3 \cdot \vec{u}_2) &= \vec{w} \cdot \vec{u}_2 \\ t_1(\vec{u}_1 \cdot \vec{u}_3) + t_2(\vec{u}_2 \cdot \vec{u}_3) + t_3(\vec{u}_3 \cdot \vec{u}_3) &= \vec{w} \cdot \vec{u}_3 \end{aligned}$$

The portal Math4U is developed within the international projects of the Department of Applied Mathematics (VSB-Technical University of Ostrava) and other institutions in Czechia, Poland, Slovakia, and Spain.
Funded by the European Union

By calculating the dot products, we obtain a system that is much simpler than system (2).

$$\begin{aligned} 9t_1 + 10t_2 &= 7 \\ 10t_1 + 14t_2 &= 7 \\ 26t_3 &= -3 \end{aligned}$$

From the last equation, we can directly determine one of the unknowns, and the first two equations form a system of two equations with two unknowns, t_1 and t_2.

Linear Combinations and Inconsistent Systems of Equations

Let us recall that a system of linear equations is said to be inconsistent if it has no solution.

We will now modify our previous problem involving the expression of a vector as a linear combination of given vectors. This time, we will omit one of the vectors we previously used. As a result, the problem becomes unsolvable in the classical sense.

Exercise 3. Express the vector $\vec{w} = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}$ as a linear combination of the vectors

$$\vec{u}_1 = \begin{pmatrix} 2 \\ 2 \\ 1 \end{pmatrix}, \quad \vec{u}_2 = \begin{pmatrix} 3 \\ 1 \\ 2 \end{pmatrix}.$$

Solving an Inconsistent System of Equations

Let us now introduce a reasonable generalization of what we consider a solution. Instead of searching for values of the unknowns that make the left- and right-hand sides exactly equal, we will look for values that make the two sides as close to each other as possible.

By the solution of an inconsistent system of equations we will mean a choice of the unknowns for which the length of the vector expressing the difference between the left- and right-hand sides of the system is minimal.

The accompanying figure helps us understand what this system represents and how to visualize its solution in the weakened sense described above.

The portal Math4U is developed within the international projects of the Department of Applied Mathematics (VSB-Technical University of Ostrava) and other institutions in Czechia, Poland, Slovakia, and Spain.
Funded by the European Union

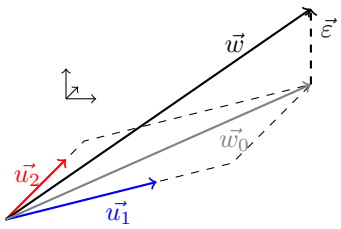

Figure 2: Vectors $\vec{u}_1$ and $\vec{u}_2$ define a plane in which vector $\vec{w}$ does not lie. Therefore, vector $\vec{w}$ cannot be expressed as a linear combination of vectors u_1 and u_2. However, it is possible to express the orthogonal projection $\vec{w}_0$ of vector $\vec{w}$ onto the plane as a linear combination of the given vectors. Among all vectors that can be written as a linear combination of $\vec{u}_1$ and $\vec{u}_2$, the vector $\vec{w}_0$ is the closest to $\vec{w}$. The quantitative criterion for this "closeness" is the length of the vector $\vec{\varepsilon}$. The fact that vector $\vec{w}_0$ is the closest to vector $\vec{w}$ of all vectors in the plane follows from the perpendicularity of vector $\vec{\varepsilon}$ to the plane defined by vectors $\vec{u}_1$ and $\vec{u}_2$.

This combination is given by the vector $\vec{w}_0$, and the difference between $\vec{w}$ and $\vec{w}_0$ is represented by the vector $\vec{\varepsilon}$. Our goal is to make the length of the vector $\vec{\varepsilon}$ as small as possible.

From a visual point of view and geometric properties, it is easy to see that this occurs when the vector $\vec{\varepsilon}$ is perpendicular to the plane defined by the vectors $\vec{u}_1$ and $\vec{u}_2$. This brings us to the same situation we encountered in the alternative solution to Exercise 3. There, we also used a trick to find the coefficients of $\vec{u}_1$ and $\vec{u}_2$ without solving the full system of equations—we took the dot product of the system with the vectors $\vec{u}_1$ and $\vec{u}_2$. In fact, we didn't even need to know the vector $\vec{\varepsilon}$ for this calculation.

Since the length of the vector $\vec{\varepsilon}$ is expressed in terms of the squares of its components, this method is called the *least squares method*.

We will demonstrate the entire procedure using the following example.

Linear Regression

Let us consider the data in the following table:

x	2	3	4
y	1	5	7

We are looking for a line $y = ax + b$ that best captures the trend in this dataset and serves as a suitable mathematical model for the data. By substituting each of the three points into the equation of the line, we obtain a system of three equations in two unknowns.

$$\begin{aligned} 2a + b &= 1 \\ 3a + b &= 5 \\ 4a + b &= 7 \end{aligned}$$

This is a system of equations that is inconsistent — also known as an overdetermined system — and it has no solution in the classical sense. The vector form of the system is:

$$a \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix} + b \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 5 \\ 7 \end{pmatrix}$$

The portal Math4U is developed within the international projects of the Department of Applied Mathematics (VSB-Technical University of Ostrava) and other institutions in Czechia, Poland, Slovakia, and Spain.
Funded by the European Union

Taking the dot product of both sides with the vectors $\begin{pmatrix}2\\3\\4\end{pmatrix}$ and $\begin{pmatrix}1\\1\\1\end{pmatrix}$ gives the following system of two equations:

$$29a + 9b = 45$$
$$9a + 3b = 13$$

The solution to this system is $a = 3$ and $b = -\frac{14}{3}$. The regression model for the given data is therefore the line

$$y = 3x - \frac{14}{3}.$$

A graph containing the given data and the regression line is shown in the figure.

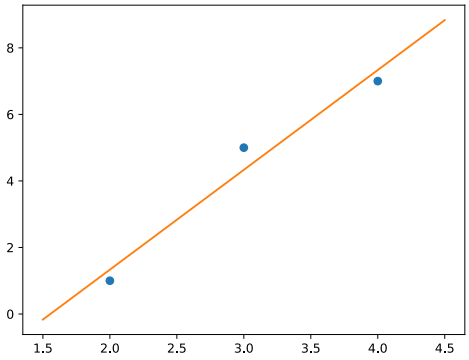

Figure 3: Three points that do not lie on a single line and a line that is a regression model for the given set of three points.

Regression for Larger Data Sets

The procedure described in the previous section for three points can be generalized to any number of points. It is not uncommon to work with data sets containing hundreds of points.

If the vector $\vec{X}$ contains the values of the independent variable[1] and the vector $\vec{Y}$ contains the values of the dependent variable, we consider the model[2]

$$\vec{Y} = a\vec{X} + b.$$

We determine the coefficients a and b by rewriting this equation as a vector equation:

$$\vec{Y} = a\vec{X} + b\vec{1},$$

where $\vec{1}$ is a vector consisting of ones. We then take the dot product of both sides with the vectors $\vec{X}$ and $\vec{1}$. We thus get a system:

[1] We will use a notation commonly used in data processing, where data sets (vectors) are denoted by capital letters and a vector that has all components equal to the same number is written as a given number with an arrow to indicate the vector.

[2] Strictly speaking, this operation does not make mathematical sense, because we are adding a vector to a real number. This operation must be interpreted component-wise, where this addition means that the real number is changed to a vector of the appropriate dimension so that the operation is defined. We call this adaptation *broadcasting*. As a result, we add the value b to each component of the vector $a\vec{X}$.

The portal Math4U is developed within the international projects of the Department of Applied Mathematics (VSB-Technical University of Ostrava) and other institutions in Czechia, Poland, Slovakia, and Spain.
Funded by the European Union

$$\begin{aligned} a(\vec{X}\cdot\vec{X}) + b(\vec{X}\cdot\vec{1}) &= \vec{X}\cdot\vec{Y} \\ a(\vec{1}\cdot\vec{X}) + b(\vec{1}\cdot\vec{1}) &= \vec{1}\cdot\vec{Y} \end{aligned} \tag{3}$$

When working with more than three points, we deal with vectors of dimension greater than three. As a result, we lose the visual geometric interpretation. Apart from that, the procedure remains unchanged. The dot product of two vectors is still calculated by multiplying the corresponding components and then summing the products.

Exercise 4. Find a suitable linear model for the data table from the introductory section.

Exercise 5. Find the regression line for the given data.

x	1.0	2.0	3.0	4.0
y	2.3	2.5	3.1	3.3

Final Notes

- In statistics, the method described above is one of the fundamental tools used to predict whether one variable influences the values of another. For this reason, there are methods that evaluate the quality of the approximation and assess whether the chosen model is suitable for a given set of data points.
- There are also multivariable (multivariate) versions of the least squares method, where the predicted value depends not on a single, but on several independent variables.
- The problem of solving an inconsistent system of linear equations also appears in image reconstruction in *acoustic tomography*. This allows studying the composition of geological layers or the health of wood or a tree based on information about the speed at which elastic deformation waves pass through the material. A series of articles on the blog `https://tomroelandts.com/` can serve as an introduction to the issue.
- It is also possible to derive direct formulas for calculating the coefficients of linear regression from the given data — this approach avoids computing dot products and solving systems of equations. See for example `https://en.wikipedia.org/wiki/Simple_linear_regression\#Expanded_formulas`.

Literature and References

Literature

- Wikipedie, Simple linear regression, https://en.wikipedia.org/wiki/Simple_linear_regression, 12.4.2024
- Tom Roelandts, https://tomroelandts.com/articles/tomography-part-1-projections, https://tomroelandts.com/arti sirt-algorithm, 13.4.2024

The portal Math4U is developed within the international projects of the Department of Applied Mathematics (VSB-Technical University of Ostrava) and other institutions in Czechia, Poland, Slovakia, and Spain.

Funded by the European Union

Sources of figures

- https://commons.wikimedia.org/wiki/File:Flag-map_of_the_United_States.pdf

The portal Math4U is developed within the international projects of the Department of Applied Mathematics (VSB-Technical University of Ostrava) and other institutions in Czechia, Poland, Slovakia, and Spain.

Funded by the European Union