Introduction to regression

Introduction to regression

In regression, our objective is to understand some dependent variable Y based on an independent variable X. Regression is a tool for predicting the value of Y as a function of X. We can use this tool to do the following,

- Support hypotheses of causation of changes in y values due to changes in x values
- Predict y values as a function of x values
- To explain the variation of y values using x values

Regression analysis can be used to support causal hypotheses, but it **cannot, by itself be used to determine causality**.

Visualising a regression of y values against x values

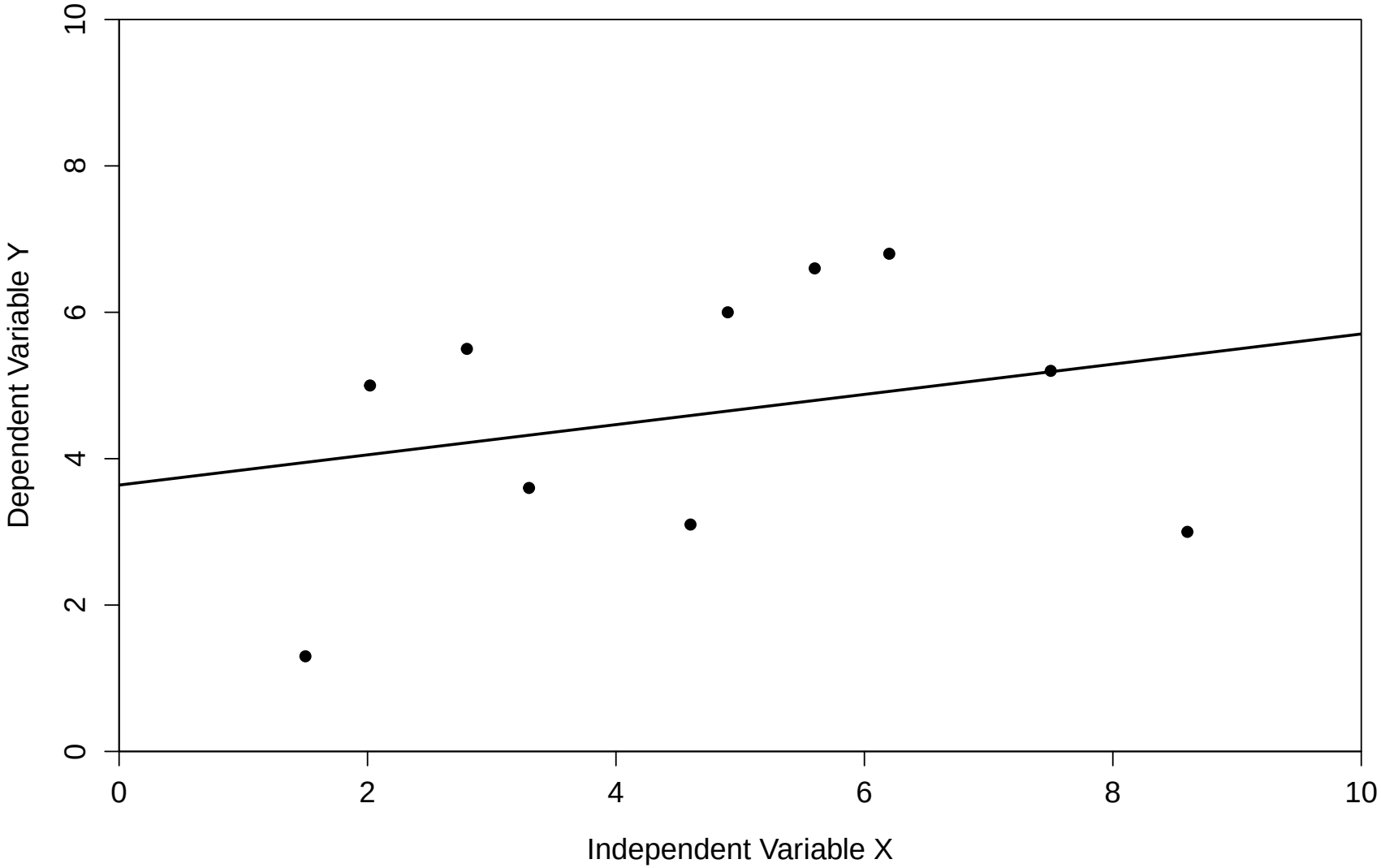

Figure 1: A regression of one dependent variable y against the independent variable x.

Regression includes independent and dependent variables

It is critical to correctly distinguish between the independent and dependent variables.

- Independent variable (X) is free to vary
- Dependent variable (Y) is predicted to change given a change in the independent variable
- Different results will be obtained if the two variables are confused

In an experiment, the independent variable is something that we as researchers have control over (e.g., amount of fertiliser to put down on a field), whereas the dependent variable is something that we would measure when collecting our data (e.g., total crop yield of the field).

Regression line

The line of best fit in a regression can be described mathematically with a simple equation,

$$y = a + bx.$$

This equation includes the variables x and y, and two coefficients

- a is the **intercept**; the value of y that is predicted when x = 0
- b is the **slope**; how much y changes for a change one unit of x

Note that data points rarely will sit right on the regression line. The **residual** is defined by the difference between the measured value of y (i.e., the data point) and the y value predicted by the regression line (i.e., the vertical distance between the data point and the line).

Regression line

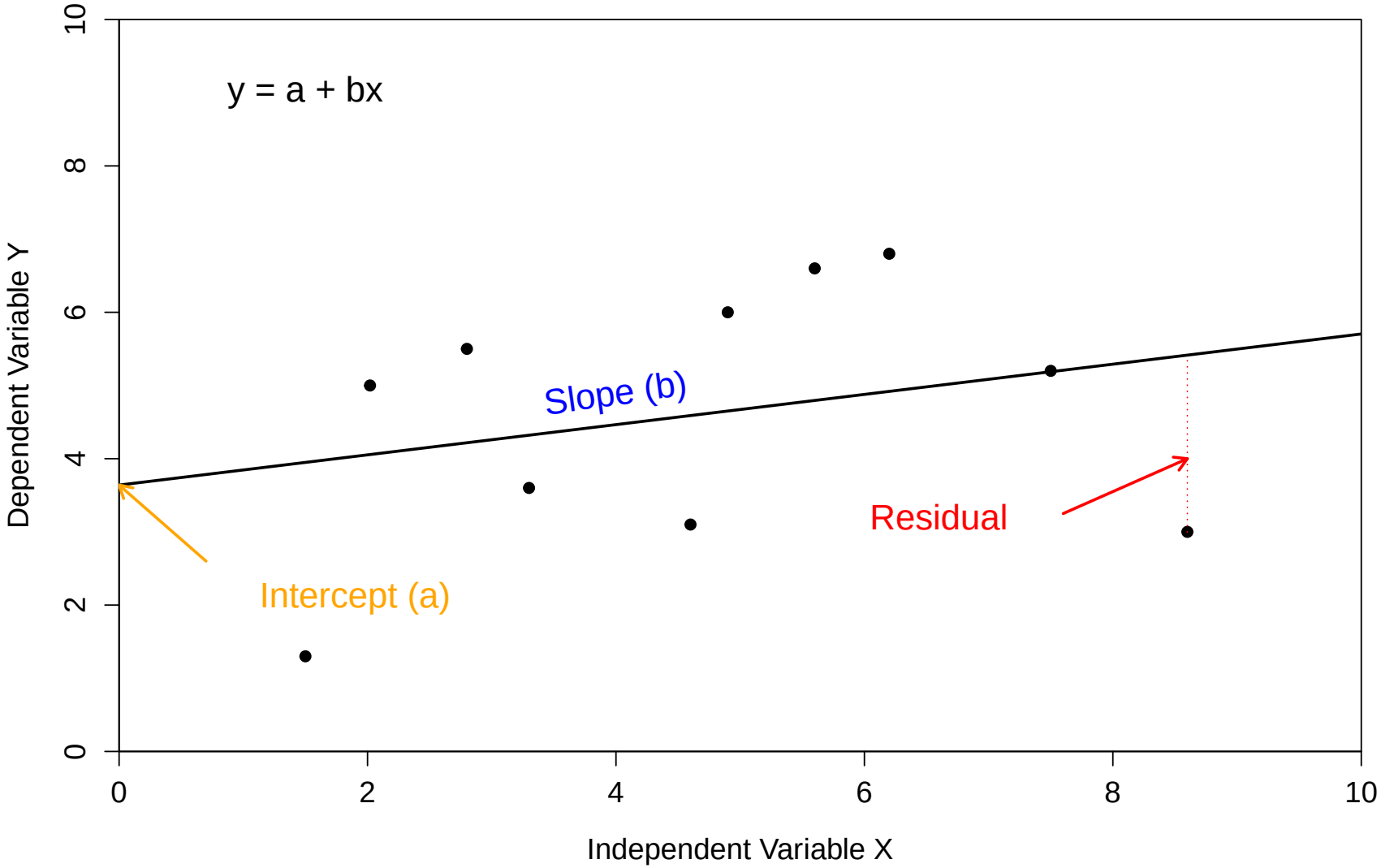

Figure 2: A regression of one dependent variable y against the independent variable x.

How do we decide what is the best fit line?

Now we can turn to how we calculate where the regression line should be through our data.

- How do we know what our intercept (a) and slope (b) should be?
- Use the method of **least squares regression**
- Minimise the sum of squares of all the residual values

To get an intuitive sense for how the regression line minimises the sum of squares, use **[this interactive application]** to adjust the slope and intercept to try to find the line of best fit (it will turn blue when you succeed).

Assumptions of regression

Regression is a widely used, but also often misused, statistical technique. It is important to be aware of the assumptions underlying linear regression.

1. **The independent variable X is measured without error**
2. **The relationship between X and Y is linear**
3. **For any value of X, Y is normally distributed**
4. **For all values of X, the variance of the residuals is identical**

Note that even if our assumptions are not perfectly met (indeed, they rarely if ever will be), this does not completely invalidate the method of linear regression. But large violations of one or more of these assumptions might indeed be problematic.

Assumptions of regression

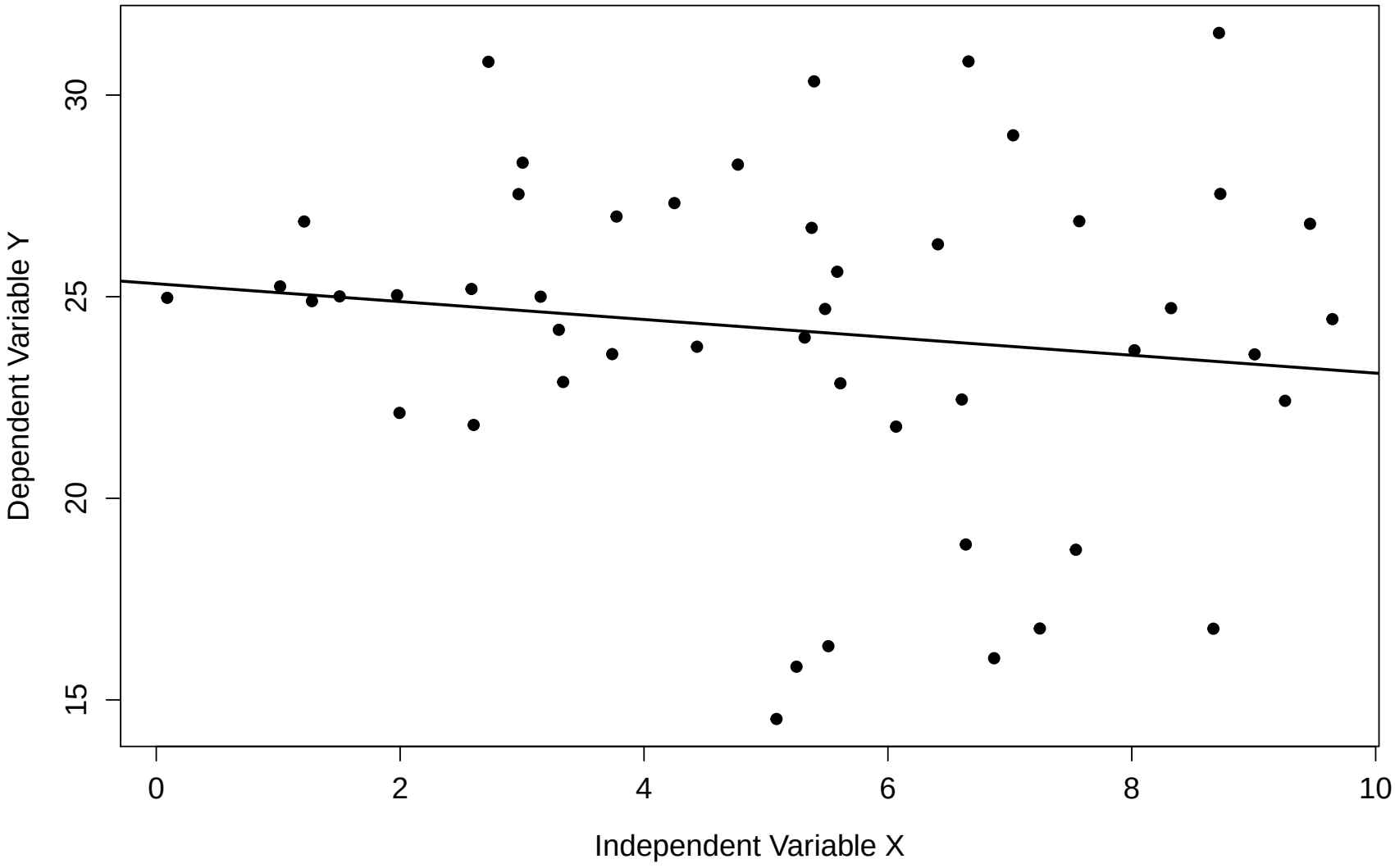

Figure 3: A regression of one dependent variable y against the independent variable x in which there is clear heteroscedasticity.

Assumptions of regression

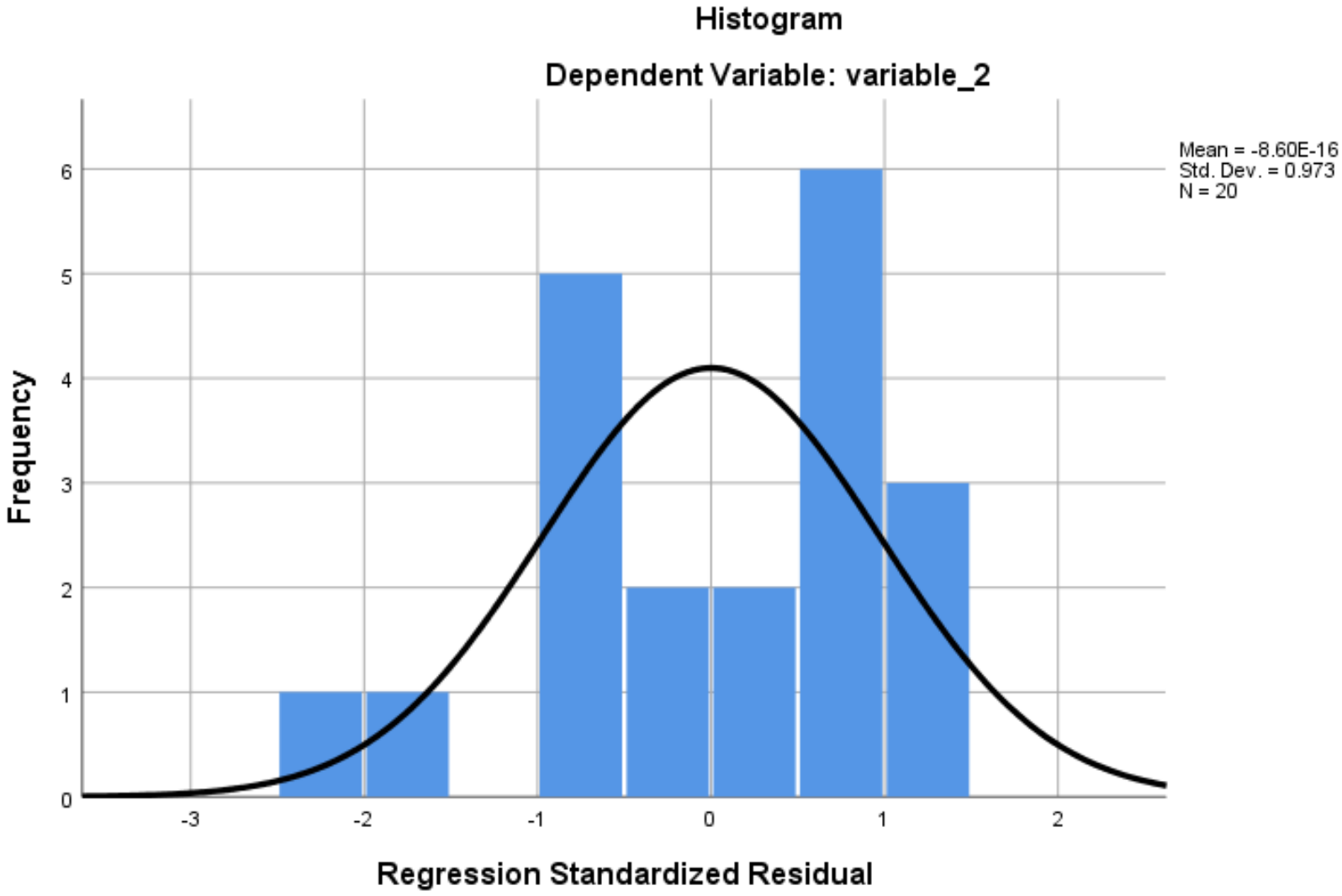

Figure 4: Example of a histogram of the residual values of a model produced in SPSS.