Regression

High School of Digital Culture
ITMO University
dc@itmo.ru

Contents

1 Simple Linear Regression

This time we are going to cover regression that is one of two fundamental problems in supervised learning. We have already noted that regression is a problem of predicting the value Y (or response) given the values of the input variables $X_1, X_2, ..., X_p$ (or predictors). We assume that the function $f(X)$ corresponding to the relationship $Y = f(X_1, X_2, ..., X_p)$ is linear. The task is to find coefficients for the linear model. As mathematicians, we call this a problem of parameter estimation. Let's get started!

1.1 Linear Regression and ML

We often need to determine the relationship between a random variable and one or more other variables. A statistical relationship is the most general one. For example, assume that $X = \xi + \eta$ is the sum of the random variables ξ and η, while $Y = \xi + \varphi$ is the sum of the random variables ξ and φ. It is clear that X and Y are dependent, but there is no explicit functional relationship. In other words, we cannot express the relationship by means of an equation $X = f(Y)$ or $Y = f(X)$.

Let's look at a real-world example. We know that the apartment price depends on the total area, the floor in the building, location, and other parameters, but it is not a function of them. There's a whole bunch of things that are impossible to take into account. For example, given the same input parameters (which is quite relative), the seller who needs money is likely to offer a lower price. However, if this seller is not in desperate need of money, the price will not be lowered. The seller can even raise it because swallows build a nest nearby each spring. Given that, how can we understand the pricing?

What can we do? How to obtain a function predicting changes in the values of interest, based on changes in the parameters? As to dependent random variables, it makes sense to consider the expected value of one of them against the fixed value of another and to find out how the change in the second value affects the mean of the first one. In the apartment example, the mean of the price can be considered a function of the parameters that affect the price.

This module introduces a quite simple concept of linear regression that is often used in supervised learning. Linear regression has been around for a long time and is found in innumerable textbooks. Though it may seem somewhat dull compared to some of the more advanced statistical learning approaches (that we will discuss later), linear regression is still widely used both in statistics and its applications in machine learning. Moreover, linear regression serves as a good jumping-off point for newer approaches: as we will see later, many statistical methods can be seen as generalizations or extensions of linear regression.

What are the questions that linear regression can answer? Let's look at the following example. Assume that we are statistical consultants in a company. We analyze data to boost sales of a specific product. Let the products be phones and planes, for example. One hundred and one distributors sell these products (the sample size is 101). The input data is the amount of funding invested in advertising of a particular product (in thousands of dollars). The output data is the volume of sold goods (in thousands of units). In the figures, the abscissa stands for the amount of advertising budget, the ordinate stands for the sales volume (in thousands of units). Phones and planes are shown in figures 1 and 2 respectively. According to the figure, advertising has an overall positive effect on the sales volume of phones, although the type of relationship is not very clear.

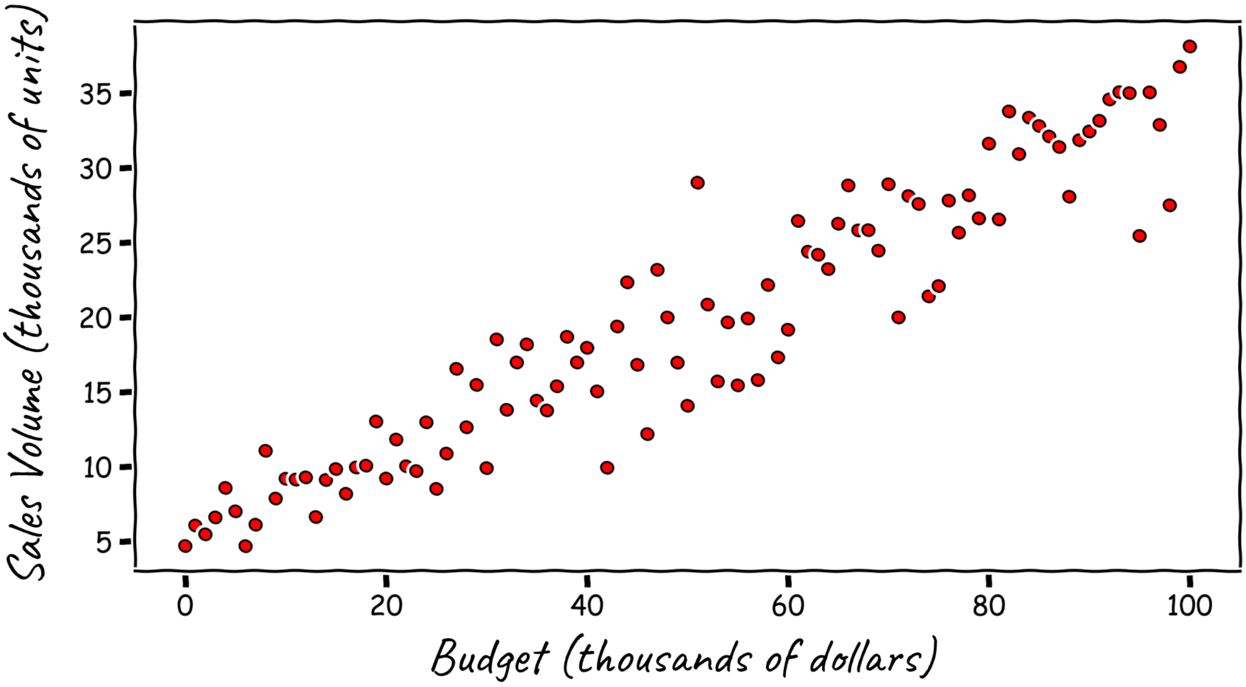

Figure 1: The relationship between phone sales volume and advertising budget

It would be naive to assume that the canonical relationship would be the one in fig. 3. The relationship was obtained when the neighboring points were simply connected by line segments. How to interpret and outline such a model? Why do sales drop or sharply rise when the funding is being raised? Maybe we miss something important here, for example, summer vacation (when the sales obviously decrease) or the New Year's Eve (when the sales sharply increase for the apparent reason). The other explanation is that the data contains some errors. In all these cases, the prediction of the proposed model will be just as bad as a random number.

Figure 2 is more difficult to interpret. We see that a small amount of funding (up to 100 thousand dollars) is associated with approximately the same sales volume, although there are also outliers in the direction of the volume increase. However, the further situation is contradictory. The increase in the advertising

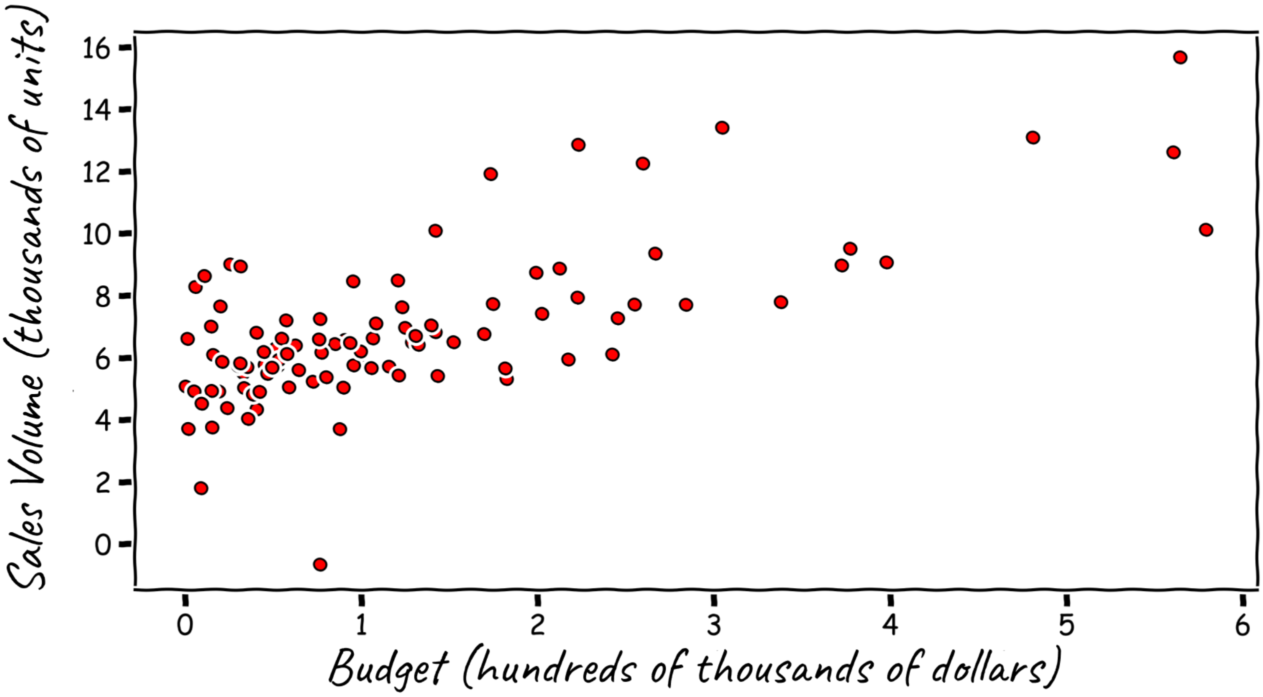

Figure 2: The relationship between plane sales volume and budget

budget by 400 or 600 thousand dollars almost doubles the sales, again, except for some outliers. By the way, the second figure also shows the obviously anomalous data relating to the negative sales volume.

The company head cannot exert a direct influence on the sales volume but can adjust the advertising budget, indirectly influencing the sales. What questions may arise?

1. Is there a relationship between the sales volume and advertising budget? Clearly, if there is no relationship, then why spend money on advertising.

2. What if there is a relationship? How strong is it? In other words, if we know the advertising budget, can we accurately predict the sales volume? If yes, then the relationship is strong, otherwise, it is weak.

3. What are the popular products to sell? Will advertising of all products be worth it?

4. How accurately can we estimate the change in sales volume with respect to the changes in the advertising budget?

5. How accurately can we predict the sales volume if we know the advertising budget?

6. Is the relationship linear?

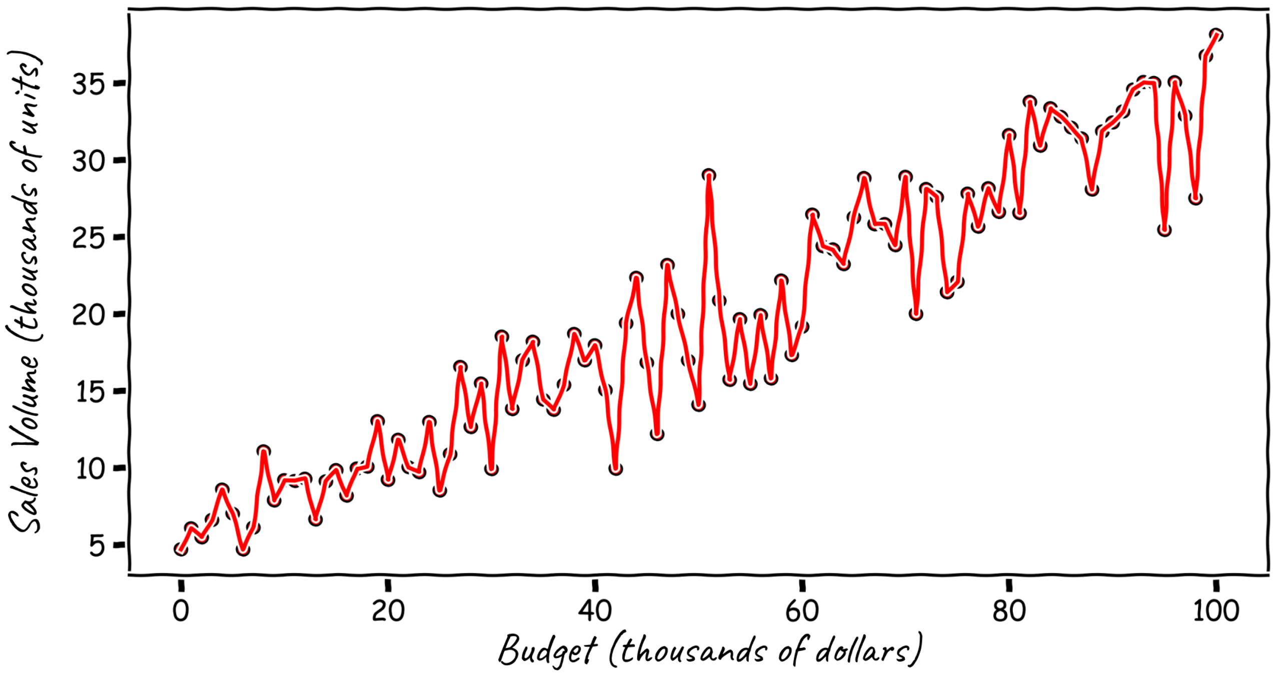

Figure 3: Naive relationship

7. Is there synergy among sales areas? Perhaps an infusion of 50.000 into phone advertising and 50.000 into plane advertising is better (because it will raise the sales volume) than the infusion of 100.000 into the phone advertising only.

It turns out that linear regression can answer each of these questions. Let's study it in a bit more detail.

1.2 Simple Linear Regression Model and Ordinary Least Squares

Assume that an observed (random) variable Y depends on some known non-random factor X_1 and a random error ε. The latter is caused by the uncertainty of measurement or model errors. A random error can also be a part of the experiment. We will consider the following linear model as the main one

$$Y = \theta_0 + \theta_1 X_1 + \varepsilon.$$

Well, the relationship between Y and X_1 is assumed to be linear with the accuracy of some error.

Definition 1.2.1 *The model described by the relationship*

$$Y = \theta_0 + \theta_1 X_1 + \varepsilon,$$

where θ_0, θ_1 are numeric parameters, X_1 is non-random parameter, which values are either set or observed (to put it differently, known), and ε is a random error, called a simple linear regression model.

The following two definitions are often used.

Definition 1.2.2 *A function*

$$f(X_1) = \theta_0 + \theta_1 X_1$$

in a simple linear regression model is called the line of regression of Y on X_1.

Definition 1.2.3 *An equation*

$$Y = \theta_0 + \theta_1 X_1$$

in a simple linear regression model is called the regression equation of Y on X_1.

An abstract model is good. But how to construct and apply it on concrete observed values of X_1 and Y? Let's describe the scheme in detail. To begin with, we can carefully write what we have here.

Assume a series of n experiments is conducted, where each (pay attention to it) **non-random** value X_1 takes values $x_1, x_2, ..., x_n$. Assume that at least two of these values are different. Depending on the values of X_1, we observe n values of $Y_1, Y_2, ..., Y_n$ of the **random** variable Y. Thus, the experiment provided us with the set of n data pairs $(x_1, Y_1), (x_2, Y_2), ..., (x_n, Y_n)$. They can easily be plotted in the plane. Fig. 4 shows the relationship between the phone sales with respect to their advertising budget, as well as 101 data pairs. The **non-random** variable X_1 is **known** the advertising budget, and the **random** variable Y is sales volume given the known budget and **unknown random** factors (or errors).

Since the experiment produced random errors, which nature we discussed earlier, then accurate equality

$$Y_i = \theta_0 + \theta_1 x_i$$

for each $i \in \{1, 2, ..., n\}$ given the same (unknown so far) parameters θ_0 and θ_1 is not possible to obtain. However, we can say that, given any i, the following relation is true:

$$Y_i = \theta_0 + \theta_1 x_i + \varepsilon_i$$

given the same parameters θ_0 and θ_1.

Remark 1.2.1 *Note that concrete values of $Y_1, Y_2, ..., Y_n$ of the random variable Y are reasonably designated by small letters $y_1, y_2, ..., y_n$. However, to disambiguate between random Y and non-random X_1, let's designate the concrete values of the random variable Y by capital letters in this module.*

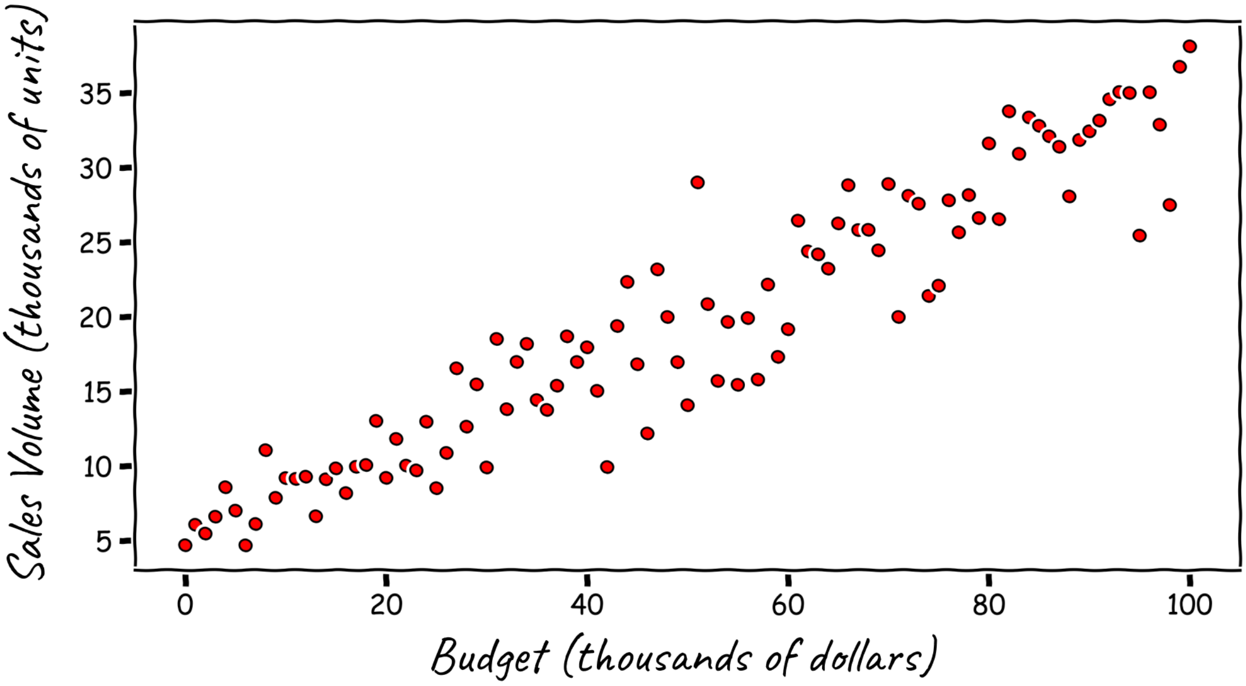

Figure 4: The relationship between phone sales volume and advertising budget

Well, we've introduced the model. But how to find estimators of the parameters θ_0 and θ_1 given the dataset $(x_1, Y_1), (x_2, Y_2), ..., (x_n, Y_n)$? If the parameters are unknown, we cannot solve a prediction problem (which is our goal).

We will use ordinary least squares (OLS) to find regression coefficients. This method allows finding such estimators $\widehat{\theta}_0$ and $\widehat{\theta}_1$ for parameters θ_0 and θ_1 that minimize the sum of squared errors $\varepsilon(\theta_0, \theta_1)$ in the observed n experiments. In other words, we minimize the function

$$\varepsilon(\theta_0, \theta_1) = \varepsilon_1^2 + \varepsilon_2^2 + ... + \varepsilon_n^2 = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \theta_0 - \theta_1 x_i)^2$$

and find the arguments minimizing the function. The following problem is solved

$$\arg\min_{\theta_0, \theta_1} \sum_{i=1}^{n} (Y_i - \theta_0 - \theta_1 x_i)^2 .$$

Definition 1.2.4 *An estimator of the ordinary least squares method for unknown parameters θ_0 and θ_1 of the regression equation is a set of parameter values minimizing the expression*

$$\varepsilon(\theta_0, \theta_1) = \sum_{i=1}^{n} (Y_i - \theta_0 - \theta_1 x_i)^2 .$$

Well then, nothing remains now on this side of the question. Technically, it is a function

$$\varepsilon(\theta_0, \theta_1) = \sum_{i=1}^{n} (Y_i - \theta_0 - \theta_1 x_i)^2,$$

of two variables θ_0 and θ_1, which we are going to minimize. To solve the minimization problem, we can use the following theorem.

Theorem 1.2.1 *The minimum of the function*

$$\varepsilon(\theta_0, \theta_1) = \sum_{i=1}^{n} (Y_i - \theta_0 - \theta_1 x_i)^2$$

is unique and attained when

$$\theta_1 = \frac{\sum\limits_{i=1}^{n}(x_i - \overline{X_1})(Y_i - \overline{Y})}{\sum\limits_{i=1}^{n}(x_i - \overline{X_1})^2}, \quad \theta_0 = \overline{Y} - \theta_1 \overline{X_1},$$

where $\overline{X_1}$ is a mean of values that the variable X_1 takes, that is,

$$\overline{X_1} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

and $\overline{Y}$ is a mean of values that the variable Y takes, that is,

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Proof. The function is differentiable. Therefore, a necessary condition for an extremum is the equality to zero of the partial derivatives of the function:

$$\begin{cases} \frac{\partial\varepsilon(\theta_0,\theta_1)}{\partial\theta_0} = 0 \\ \frac{\partial\varepsilon(\theta_0,\theta_1)}{\partial\theta_1} = 0 \end{cases} \Leftrightarrow \begin{cases} -2\sum\limits_{i=1}^{n} (Y_i - \theta_0 - \theta_1 x_i) = 0 \\ -2\sum\limits_{i=1}^{n} x_i (Y_i - \theta_0 - \theta_1 x_i) = 0 \end{cases}.$$

After solving the system (in other words, a linear system of two equations with two unknown θ_0 and θ_1), we find that

$$\theta_1 = \frac{\sum\limits_{i=1}^{n}(x_i - \overline{X_1})(Y_i - \overline{Y})}{\sum\limits_{i=1}^{n}(x_i - \overline{X_1})^2}, \quad \theta_0 = \overline{Y} - \theta_1 \overline{X_1}.$$

Of course, we can call the obtained values of θ_0 and θ_1 the estimates only after verifying that the obtained point is a minimum point. We can do this by using some sufficient condition for an extremum of a two-variable function or use the following: the function $\varepsilon(\theta_0, \theta_1)$ is convex downward, therefore, the obtained critical point is truly the minimum point. □

Now we can write that

$$\widehat{\theta}_1 = \frac{\sum\limits_{i=1}^{n}(x_i - \overline{X_1})(Y_i - \overline{Y})}{\sum\limits_{i=1}^{n}(x_i - \overline{X_1})^2}, \quad \widehat{\theta}_0 = \overline{Y} - \widehat{\theta}_1\overline{X_1},$$

where

$$\overline{X_1} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

As a result, we get the values $\widehat{\theta}_0 \approx 4.88$ and $\widehat{\theta}_1 \approx 0.30$ based on the phone data. The function

$$f(X_1) = 4.88 + 0.30X_1$$

is a desired linear regression line. Let's create its graph. It is shown in the figure 5 in blue.

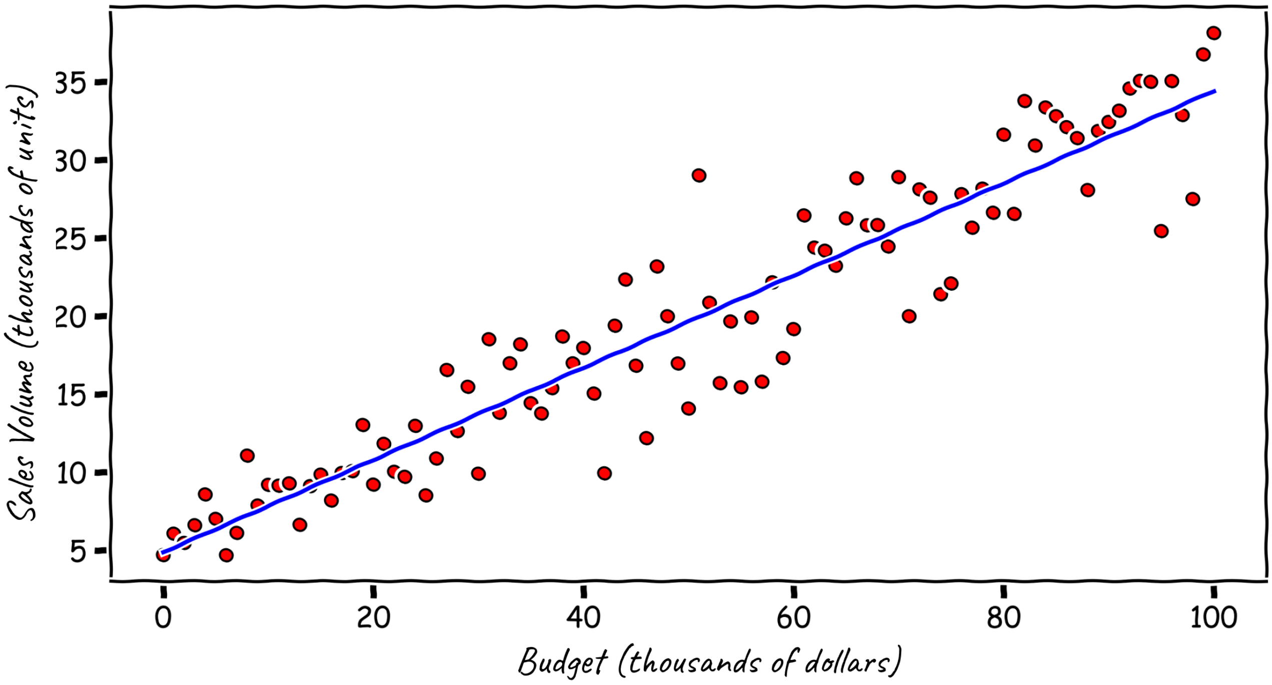

Figure 5: The relationship between phone sales volume and advertising budget as well as regression

The graph shows that the obtained model quite well approximates the data (actually, we will discuss the bad/not bad criterion later, again, returning to these examples). Moreover, if we draw the vertical green lines (that are parallel to the axis Oy) from the red dots to the blue line, we will get the errors ε_i (the minimized sum of the squares). They show the deviation of the model from the real-life data. See fig. 6.

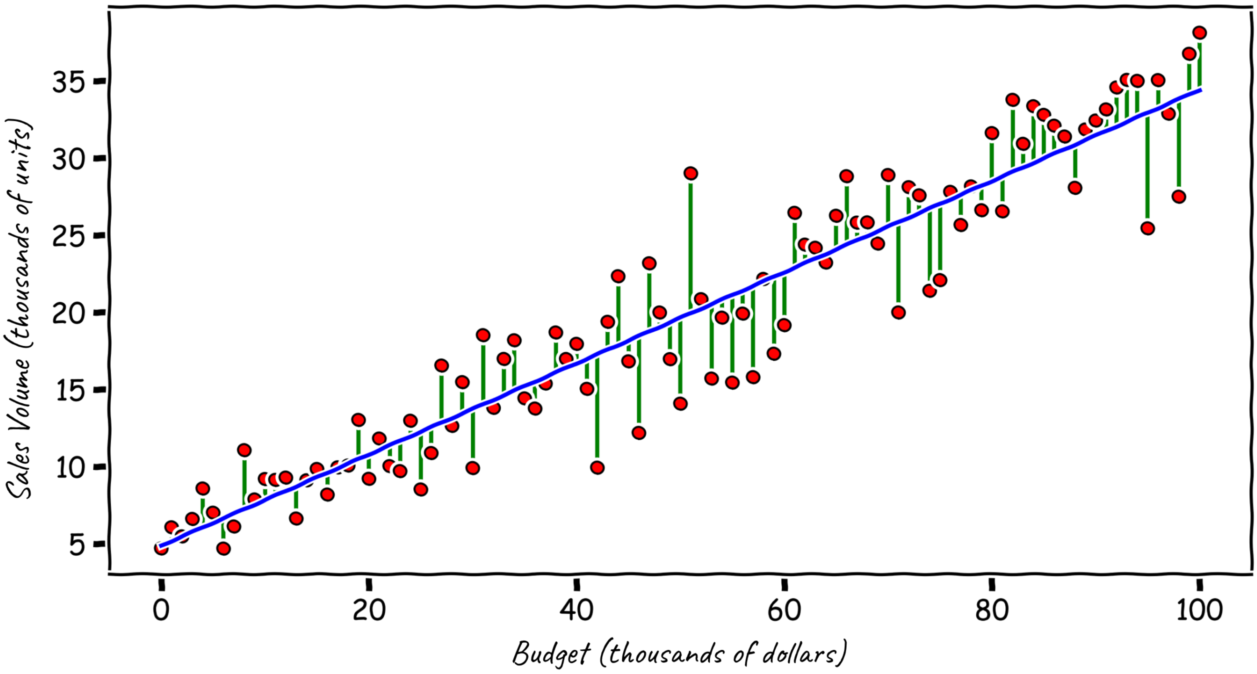

Figure 6: The relationship between phone sales volume and advertising budget, as well as regression and errors

We get the values $\widehat{\theta_0} \approx 5.13$ and $\widehat{\theta_1} \approx 1.34$ based on plane sales data. Thus, the function

$$f(X_1) = 5.13 + 1.34X_1$$

is a desired linear regression line. Let's create a graph. It is shown in the figure 7 in blue. A little bit later, we will discuss the accuracy of this and previous models.

Note that, after finding the estimators $\widehat{\theta_0}$ and $\widehat{\theta_1}$, the prediction is obtained using the regression equation:

$$Y = \widehat{\theta_0} + \widehat{\theta_1}X_1.$$

For example, the phone model with the equation

$$Y = 4.88 + 0.30X_1$$

(given the advertising budget of 150 thousand dollars) makes a prediction with respect to the sales volume of

$$Y = 4.88 + 0.30 \cdot 150 = 49.88$$

thousand units.

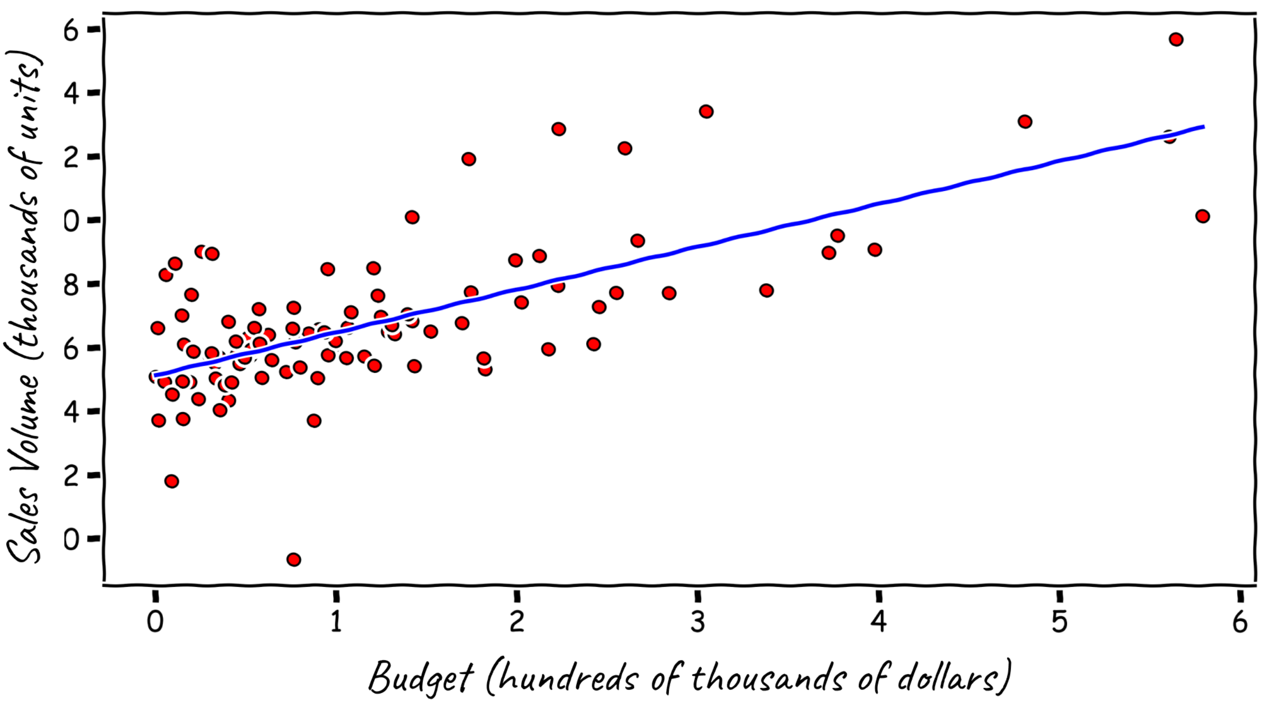

Figure 7: The relationship between plane sales volume and advertising budget as well as regression

1.3 Shopping Time Example

Don't be intimidated by the formulas. To make everything less abstract, we will use a specific simple example to perform the calculations. Assume we have data on how many minutes a person spends in a grocery. The time depends on the number of items purchased by the person. The data is presented in the table.

Observation No.	Number of purchased items	Shopping time (min)
1	10	15
2	5	12
3	12	18
4	25	30
5	1	3
6	18	20
7	11	14
8	7	10
9	19	20
10	15	13

Assume that you should purchase specific items, but you are running out of time. Can you estimate how much time you will need to purchase a different number of items based on the data about your previous shopping trips? Before modeling, we should determine the known variable and response. The number of items purchased by a person will be the known variable X_1, and the time spent in the store will be the response Y. Therefore, we obtain the following set of pairs (x_1, Y_1), (x_2, Y_2), $\ldots$, (x_{10}, Y_{10}) of input data (given in the table for convenience):

Observation No.	Number of purchased items	Shopping time (min)
1	$x_1 = 10$	$Y_1 = 15$
2	$x_2 = 5$	$Y_2 = 12$
3	$x_3 = 12$	$Y_3 = 18$
4	$x_4 = 25$	$Y_4 = 30$
5	$x_5 = 1$	$Y_5 = 3$
6	$x_6 = 18$	$Y_6 = 20$
7	$x_7 = 11$	$Y_7 = 14$
8	$x_8 = 7$	$Y_8 = 10$
9	$x_9 = 19$	$Y_9 = 20$
10	$x_{10} = 15$	$Y_{10} = 13$

To illustrate, let's plot the data in the fig. 8. We plot the x values along the horizontal axis and the y values along the vertical axis.

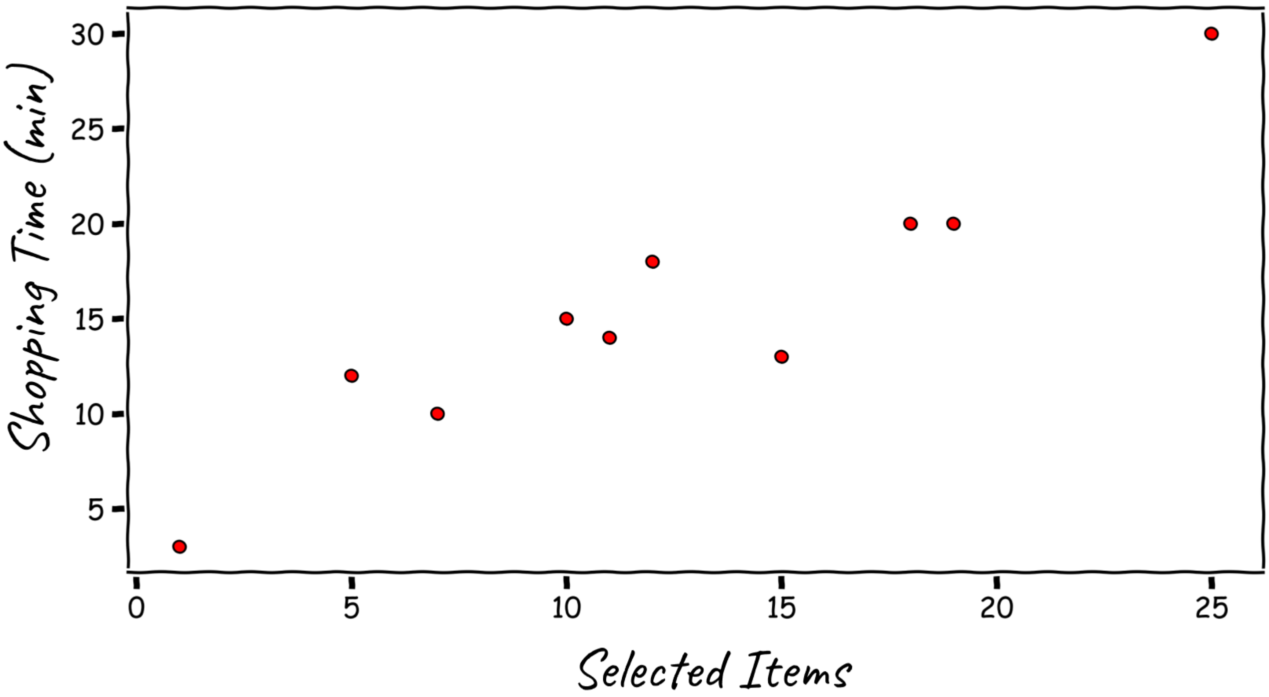

Figure 8: The relationship between shopping time and the number of purchased items

In this case, $n = 10$, so the formulas for calculating $\widehat{\theta_0}$ and $\widehat{\theta_1}$ take the following form:

$$\widehat{\theta_1} = \frac{\sum\limits_{i=1}^{10}(x_i - \overline{X_1})(Y_i - \overline{Y})}{\sum\limits_{i=1}^{10}(x_i - \overline{X_1})^2}, \quad \widehat{\theta_0} = \overline{Y} - \widehat{\theta_1}\overline{X_1},$$

where

$$\overline{X_1} = \frac{1}{10}\sum_{i=1}^{10} x_i, \quad \overline{Y} = \frac{1}{10}\sum_{i=1}^{10} Y_i.$$

Let's start with calculating the last ones. Well,

$$\overline{X_1} = \frac{1}{10}(10 + 5 + 12 + 25 + 1 + 18 + 11 + 7 + 19 + 15) = \frac{123}{10} = 12.3,$$

$$\overline{Y} = \frac{1}{10}(15 + 12 + 18 + 30 + 3 + 20 + 14 + 10 + 20 + 13) = \frac{155}{10} = 15.5.$$

Now we can calculate $\widehat{\theta_1}$:

$$\widehat{\theta_1} = \frac{(x_1 - \overline{X_1})(Y_1 - \overline{Y}) + (x_2 - \overline{X_1})(Y_2 - \overline{Y}) + ... + (x_{10} - \overline{X_1})(Y_{10} - \overline{Y})}{(x_1 - \overline{X_1})^2 + (x_2 - \overline{X_1})^2 + ... + (x_{10} - \overline{X_1})^2} =$$

$$= \frac{(10 - 12.3)(15 - 15.5) + (5 - 12.3)(12 - 15.5) + ... + (15 - 12.3)(13 - 15.5)}{(10 - 12.3)^2 + (5 - 12.3)^2 + ... + (15 - 12.3)^2} \approx 0.93.$$

Hence,

$$\widehat{\theta_0} \approx 15.5 - 0.93 \cdot 12.3 \approx 4.06.$$

The common practice is not to round the values and plug in the value of $\widehat{\theta_1}$ with as many digits after a decimal point as possible while calculating $\widehat{\theta_0}$. We've rounded $\widehat{\theta_1}$ and found the approximate value of $\widehat{\theta_0}$ for clarity.

Hence, the equation for linear regression takes the following form:

$$Y = 4.06 + 0.93X_1.$$

The constructed straight line is shown in figure 9. Let's return to the prediction problem. We pose a question, how long it takes to shop for 27 products. To make a prediction, we just need to calculate the value of the function $Y = 4.06 + 0.93X_1$ at $X_1 = 27$. Hence,

$$4.06 + 0.93 \cdot 27 = 29.17,$$

it will take a little bit more than 29 minutes. The illustrated prediction is shown in figure 10.

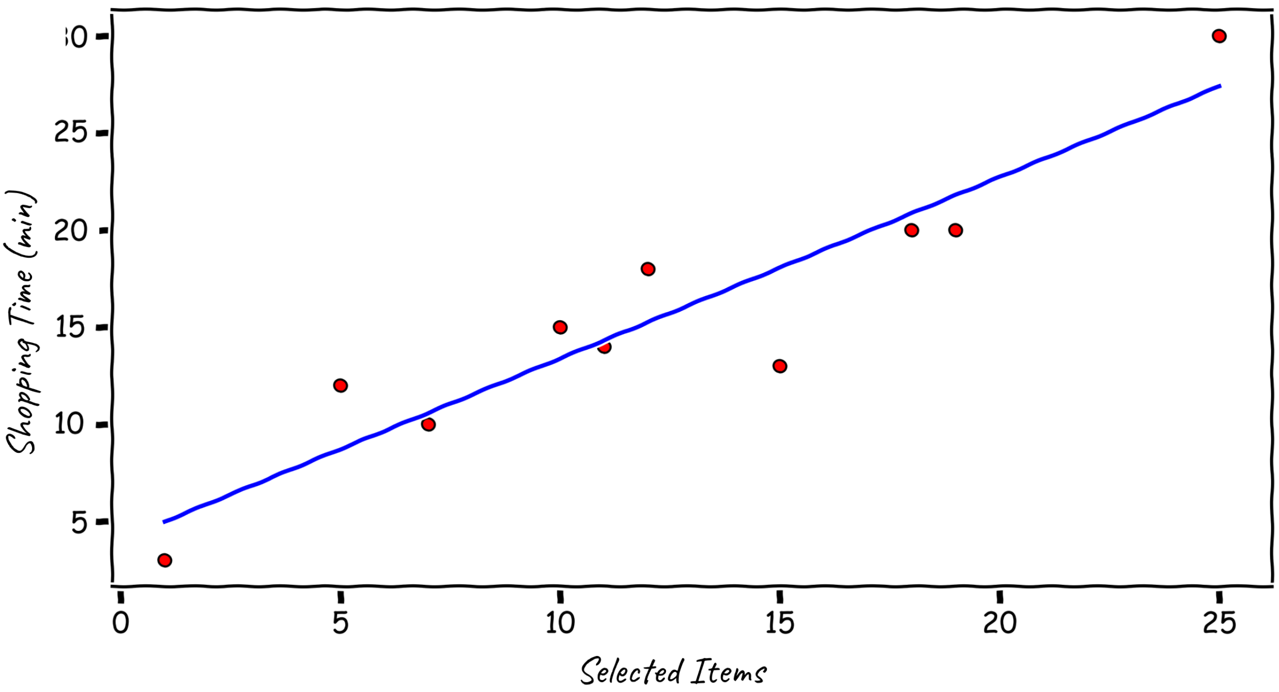

Figure 9: The relationship between shopping time and the number of purchased items as well as regression

2 Some Statistical Characteristics of Parameters of Simple Linear Regression Model

Earlier we solved the problem of regression construction rather at the heuristic level because we didn't explain why the described scheme provided a good model, or why the ordinary least squares method was used. Moreover, we didn't discuss model applicability and left some questions unanswered. What are these random errors ε? Can they really be any? Let's discuss the constructed model in detail.

2.1 Parameters θ_0 and θ_1 as Random Variables

Recall the problem setting. After n experiments, where the value X_1 has taken the values $x_1, x_2, ..., x_n$ (at least two of which are different), we observe n values $Y_1, Y_2, ..., Y_n$ of our random variable Y. Since measurements led to random errors in the experiment, we assume that, given any i, the following relation is true:

$$Y_i = \theta_0 + \theta_1 x_i + \varepsilon_i$$

given the same parameters θ_0 and θ_1.

Now we will make the following important assumptions (the first three are often called Gauss–Markov assumptions):

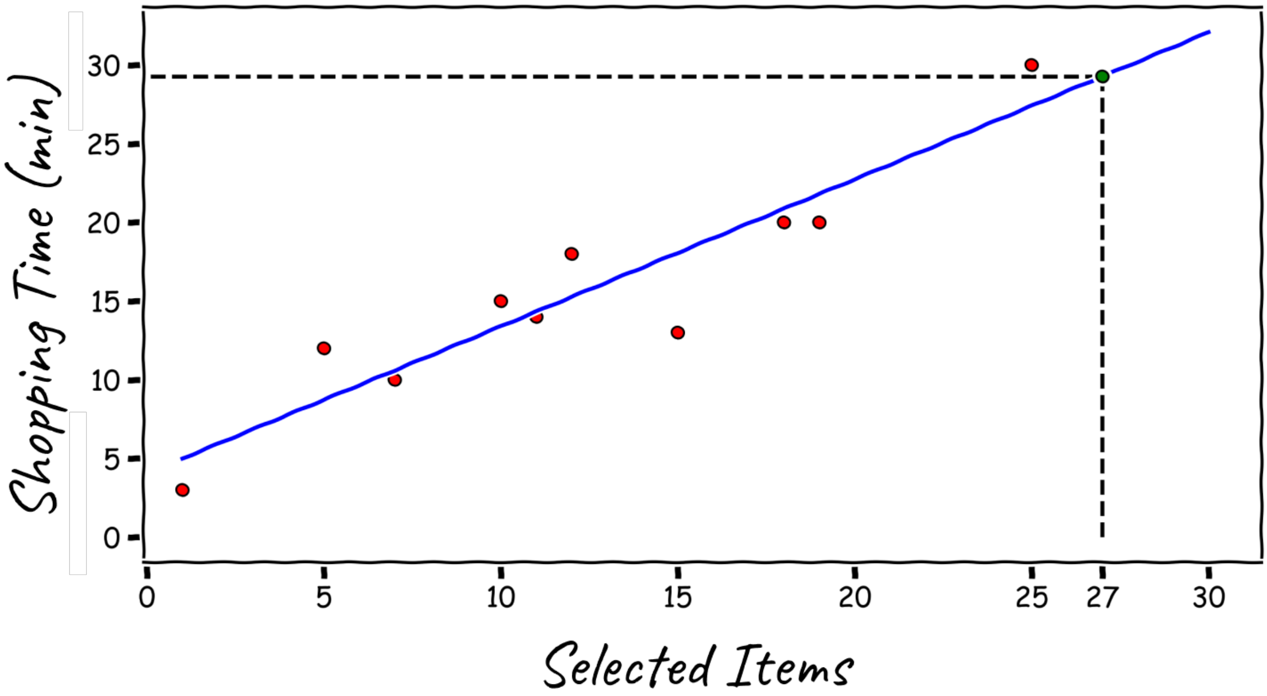

Figure 10: The relationship between shopping time and the number of purchased items as well as regression and prediction

1. Random variables (errors) $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ are independent and identically distributed.

2. Errors are not systematic: $\mathsf{E}\varepsilon_i = 0$, $i \in \{1, 2, ..., n\}$.

3. Error variances are the same: $\mathsf{D}\varepsilon_i = \sigma^2 > 0$, $i \in \{1, 2, ..., n\}$ (homoscedasticity).

4. $\varepsilon_i \sim \mathsf{N}_{0,\sigma^2}$.

Even based on these rigid assumptions about the distribution of random errors, the set of values $Y_1, Y_2, ..., Y_n$ of the random variable Y is not a sample in a common sense of statistics. Random variables Y_i are not identically distributed because, for example, their expected values differ since

$$\mathsf{E}Y_i = \mathsf{E}\left(\theta_0 + \theta_1 x_i + \varepsilon_i\right) = \theta_0 + \theta_1 x_i,$$

where the last values are not the same.

Now we are ready to study the main properties of estimators obtained using ordinary least squares. Recall their analytical expressions:

$$\widehat{\theta}_1 = \frac{\sum\limits_{i=1}^{n}(x_i - \overline{X_1})(Y_i - \overline{Y})}{\sum\limits_{i=1}^{n}(x_i - \overline{X_1})^2}, \quad \widehat{\theta}_0 = \overline{Y} - \widehat{\theta}_1\overline{X_1},$$

$$\overline{X_1} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

If the four discussed conditions are met, the estimators obtained using ordinary least squares match the estimators provided by the maximum likelihood estimation method.

Theorem 2.1.1 (About identical OLS and MLE estimators) *In the four assumptions, the estimators $\widehat{\theta}_0$ and $\widehat{\theta}_1$ are maximum likelihood estimators of the parameters θ_0 and θ_1.*

Proof. Let's use the maximum likelihood estimation. Although $Y_1, Y_2, ..., Y_n$ is not a sample in a common sense, the method and its implementation remain the same due to the independence of Y_i. Since

$$Y_i \sim \mathsf{N}_{\theta_0+\theta_1 x_i, \sigma^2},$$

the likelihood function takes the form:

$$f_\theta(\vec{Y}) = f_\theta(Y_1, Y_2, ..., Y_n) = \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\frac{\sum\limits_{i=1}^{n}(Y_i-\theta_0-\theta_1 x_i)^2}{2\sigma^2}} = (2\pi\sigma^2)^{-n/2} e^{-\frac{\varepsilon(\theta_0,\theta_1)}{2\sigma^2}},$$

where

$$\varepsilon(\theta_0, \theta_1) = \sum_{i=1}^{n}(Y_i - \theta_0 - \theta_1 x_i)^2.$$

The maximum of the likelihood function is attained when the minimum of the function $\varepsilon(\theta_0, \theta_1)$ is attained. It leads us to the following problem:

$$\arg\min_{\theta_0,\theta_1} \varepsilon(\theta_0, \theta_1) = \arg\min_{\theta_0,\theta_1} \sum_{i=1}^{n}(Y_i - \theta_0 - \theta_1 x_i)^2,$$

which echoes the problem solved using the ordinary least squares method. □

Remark 2.1.1 *Note that the parameter σ^2 is usually unknown. Since the sample $Y_1, Y_2, ..., Y_n$ is not a sample in a common sense, we cannot use the estimator $S^2(Y)$ or $S_0^2(Y)$ for the estimation.*

Corollary 2.1.2 *The maximum-likelihood estimator of the unknown parameter σ^2 is the following random variable:*

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \theta_0 - \theta_1 x_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2.$$

Proof. As obtained earlier,

$$f_{\theta,\sigma^2}(\vec{Y}) = f_{\theta,\sigma^2}(Y_1, Y_2, ..., Y_n) = \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\frac{\sum\limits_{i=1}^{n}(Y_i-\theta_0-\theta_1 x_i)^2}{2\sigma^2}}.$$

Thus, a log-likelihood function is rewritten as follows:

$$L_{\theta,\sigma^2}(\vec{Y}) = -\frac{n}{2}\left(\ln 2\pi + \ln \sigma^2\right) - \frac{\sum\limits_{i=1}^{n}(Y_i - \theta_0 - \theta_1 x_i)^2}{2\sigma^2}.$$

$$\frac{\partial L_{\theta,\sigma^2}(\vec{Y})}{\partial \sigma^2} = -\frac{n}{2} \cdot \frac{1}{\sigma^2} + \frac{\sum\limits_{i=1}^{n}(Y_i - \theta_0 - \theta_1 x_i)^2}{2\sigma^4}.$$

We equate the last expression to zero and solve the resultant equation to obtain that

$$\sigma^2 = \frac{\sum\limits_{i=1}^{n}(Y_i - \theta_0 - \theta_1 x_i)^2}{n}.$$

Based on the sufficient condition for an extremum, we can conclude that the obtained point is a maximum. Thus,

$$\widehat{\sigma^2} = \frac{\sum\limits_{i=1}^{n}(Y_i - \theta_0 - \theta_1 x_i)^2}{n}.$$

□

Remark 2.1.2 *When θ_0 and θ_1 are unknown, the estimator is rewritten as*

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{\theta_0} - \widehat{\theta_1} x_i)^2$$

Now let's consider the properties of estimators $\widehat{\theta_0}$ and $\widehat{\theta_1}$. The obtained facts will help us to construct confidence intervals and test hypotheses with respect to model parameters. Let's begin with $\widehat{\theta_0}$.

Lemma 2.1.1 (About properties of the estimator $\widehat{\theta_1}$) *According to the Gauss–Markov assumptions, the estimator $\widehat{\theta_1}$ has the following properties:*

1. *It is an unbiased estimator of the parameter θ_1.*

2. *It is an effective linear unbiased estimator.*

3. *Its variance equals*

$$\mathsf{D}\widehat{\theta}_1 = \frac{\sigma^2}{\sum\limits_{i=1}^{n}(x_i - \overline{X_1})^2}.$$

4. *Given that the fourth condition is true (that is, $\varepsilon_i \sim \mathsf{N}_{0,\sigma^2}$), it has a normal distribution with the parameters θ_1 and $\mathsf{D}\widehat{\theta}_1$:*

$$\widehat{\theta}_1 \sim \mathsf{N}_{\theta_1, \mathsf{D}\widehat{\theta}_1}.$$

Let's explain the second property. It means that MSE in the class of linear unbiased estimators is minimal on the estimators obtained using ordinary least squares. Since

$$\mathsf{MSE} = \mathsf{E}||\theta - \widehat{\theta}||^2 = \mathsf{D}\widehat{\theta} + (\mathsf{E}\widehat{\theta} - \theta)^2,$$

and the last term for $\widehat{\theta}_1$ is zero due to unbiasedness, then minimality of MSE is minimality of the variance $\mathsf{D}\widehat{\theta}$ of the estimator $\widehat{\theta}$. The minimality of variance ensures minimality of spread of possible estimates from the true value of the parameter.

Proof. 1. Using the linearity property of the expected value, we obtain

$$\mathsf{E}\widehat{\theta}_1 = \frac{1}{\sum\limits_{i=1}^{n}(x_i - \overline{X_1})^2} \sum_{i=1}^{n}(x_i - \overline{X_1})(\mathsf{E}Y_i - \mathsf{E}\overline{Y}).$$

Since $\mathsf{E}\varepsilon_i = 0$,

$$\mathsf{E}Y_i = \mathsf{E}\left(\theta_0 + \theta_1 x_i + \varepsilon_i\right) = \theta_0 + \theta_1 x_i,$$

therefore,

$$\mathsf{E}\overline{Y} = \frac{1}{n}\sum_{i=1}^{n}\mathsf{E}Y_i = \frac{1}{n}\sum_{i=1}^{n}\left(\theta_0 + \theta_1 x_i\right) = \theta_0 + \theta_1\overline{X_1}.$$

Then,

$$\mathsf{E}Y_i - \mathsf{E}\overline{Y} = \theta_1(x_i - \overline{X_1})$$

and, after plugging it in the expression for $\mathsf{E}\widehat{\theta}_1$, we obtain

$$\mathsf{E}\widehat{\theta}_1 = \frac{1}{\sum\limits_{i=1}^{n}(x_i - \overline{X_1})^2}\sum_{i=1}^{n}\theta_1(x_i - \overline{X_1})^2 = \theta_1.$$

2. We will not prove this point in the general case. Its special case follows from the properties of estimators obtained using the maximum likelihood estimation

and the previous theorem.
3. Since the expression for $\widehat{\theta}_1$ can be rewritten as:

$$\widehat{\theta}_1 = \frac{1}{\sum\limits_{i=1}^{n}(x_i - \overline{X_1})^2}\left(\sum_{i=1}^{n}(x_i - \overline{X_1})Y_i - \sum_{i=1}^{n}(x_i - \overline{X_1})\overline{Y}\right),$$

and since

$$\sum_{i=1}^{n}(x_i - \overline{X_1})\overline{Y} = \overline{Y}\left(\sum_{i=1}^{n} x_i - n\overline{X}\right) = 0,$$

it's sufficient to study

$$\widehat{\theta}_1 = \frac{1}{\sum\limits_{i=1}^{n}(x_i - \overline{X_1})^2}\sum_{i=1}^{n}(x_i - \overline{X_1})Y_i.$$

Since errors are independent and by variance properties, we obtain

$$\mathsf{D}\widehat{\theta}_1 = \frac{1}{\left(\sum\limits_{i=1}^{n}(x_i - \overline{X_1})^2\right)^2}\sum_{i=1}^{n}(x_i - \overline{X_1})^2\mathsf{D}Y_i.$$

Since

$$\mathsf{D}Y_i = \mathsf{D}\left(\theta_0 + \theta_1 x_i + \varepsilon_i\right) = \mathsf{D}\varepsilon_i = \sigma^2,$$

then

$$\mathsf{D}\widehat{\theta}_1 = \frac{\sigma^2}{\left(\sum\limits_{i=1}^{n}(x_i - \overline{X_1})^2\right)^2}\sum_{i=1}^{n}(x_i - \overline{X_1})^2 = \frac{\sigma^2}{\sum\limits_{i=1}^{n}(x_i - \overline{X_1})^2}.$$

4. This property follows from the fact that the sum of independent random variables having the normal distribution has the normal distribution, which parameters were calculated in points 1 and 3.

□

Lemma 2.1.2 (About properties of the estimator $\widehat{\theta}_0$) *According to the Gauss–Markov assumptions, the estimator $\widehat{\theta}_0$ has the following properties:*

1. *It is an unbiased estimator of the parameter θ_0.*
2. *It is an effective linear unbiased estimator.*
3. *Its variance equals*

$$\mathsf{D}\widehat{\theta}_0 = \frac{\sigma^2}{n}.$$

4. *Given that the fourth condition is true (that is, $\varepsilon_i \sim \mathsf{N}_{0,\sigma^2}$), it has a normal distribution with the parameters θ_0 and $\mathsf{D}\widehat{\theta_0}$:*

$$\widehat{\theta_0} \sim \mathsf{N}_{\theta_0,\mathsf{D}\widehat{\theta_0}}.$$

5. *Assuming that the fourth condition is true (that is, $\varepsilon_i \sim \mathsf{N}_{0,\sigma^2}$), it is a maximum-likelihood estimator of the parameter θ_0.*

Proof. 1. Using the linearity property of the expected value, we obtain

$$\mathsf{E}\widehat{\theta_0} = \mathsf{E}\left(\overline{Y} - \theta_1\overline{X_1}\right) = \mathsf{E}\overline{Y} - \theta_1\overline{X_1} = \frac{1}{n}\sum_{i=1}^{n}(\theta_0 + \theta_1 x_i) - \theta_1\overline{X_1} =$$

$$= \theta_0 + \theta_1\overline{X_1} - \theta_1\overline{X_1} = \theta_0.$$

2. We will not prove this point in the general case. Its special case follows from the properties of estimators obtained using the maximum likelihood estimation and the theorem about related OLS and MLE estimators.
3. By variance properties, we obtain

$$\mathsf{D}\widehat{\theta_0} = \mathsf{D}(\overline{Y} - \theta_1\overline{X_1}) = \mathsf{D}\overline{Y} = \frac{1}{n^2}\sum_{i=1}^{n}\mathsf{D}Y_i.$$

Since

$$\mathsf{D}Y_i = \mathsf{D}(\theta_0 + \theta_1 x_i + \varepsilon_i) = \mathsf{D}\varepsilon_i = \sigma^2,$$

then

$$\mathsf{D}\widehat{\theta_0} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

4. This property follows from the fact that the sum of independent normally distributed random variables has the normal distribution and from the calculations in points 1 and 3. □

Remark 2.1.3 *The variance of the estimator $\widehat{\theta_0}$ derived from the last theorem is based on the fact that θ_1 is a known variable. If θ_1 is estimated using $\widehat{\theta_1}$, the calculations become more complicated and lead us to the following variance expression:*

$$\mathsf{D}\widehat{\theta_0} = \mathsf{D}\overline{Y} + \mathsf{D}\widehat{\theta_1}\cdot\overline{X_1}^2,$$

hence,

$$\mathsf{D}\widehat{\theta_0} = \frac{\sigma^2}{n} + \frac{\sigma^2}{\sum\limits_{i=1}^{n}(x_i - \overline{X_1})^2}\overline{X_1}^2 = \sigma^2\left(\frac{1}{n} + \frac{\overline{X_1}^2}{\sum\limits_{i=1}^{n}(x_i - \overline{X_1})^2}\right)$$

2.2 Constructing Confidence Intervals for Regression Coefficients

Well, before we formulate the main theory of confidence intervals, let's return to the estimator of the unknown parameter σ^2. Recall that, according to MLE, it has the following form:

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{\theta}_0 - \widehat{\theta}_1 x_i)^2.$$

Remark 2.2.1 *We can show that, according to the assumptions 1-4:*

$$\frac{n}{\sigma^2}\widehat{\sigma^2} \sim \chi^2_{n-2}.$$

2 degrees of freedom are taken because we don't know the parameter θ_0 and the parameter θ_1 and only estimate them. Using this and the linearity property of the expected value, as well as considering that the expected value of a random variable having the chi-square distribution with $(n-2)$ degrees of freedom equals $(n-2)$, we obtain

$$\mathsf{E}\left(\frac{n}{\sigma^2}\widehat{\sigma^2}\right) = \frac{n}{\sigma^2}\mathsf{E}\widehat{\sigma^2} = (n-2).$$

Hence,

$$\mathsf{E}\widehat{\sigma^2} = \frac{n-2}{n}\sigma^2,$$

and the obtained estimator $\widehat{\sigma^2}$ is biased but asymptotically unbiased because:

$$\lim_{n\to+\infty}\frac{n-2}{n} = 1.$$

The unbiased estimator is obtained as follows:

$$\widehat{\sigma^2_0} = \frac{n}{n-2}\widehat{\sigma^2} = \frac{n}{n-2}\cdot\frac{\sum\limits_{i=1}^{n}(Y_i - \widehat{\theta}_0 - \widehat{\theta}_1 x_i)^2}{n} = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{\theta}_0 - \widehat{\theta}_1 x_i)^2.$$

Since the true value of the parameter σ^2 is unknown, we have to estimate it together with the variances of the estimators $\widehat{\theta}_0$ and $\widehat{\theta}_1$ obtained earlier. Well, we can evaluate the spread of possible estimators $\widehat{\theta}_0$ and $\widehat{\theta}_1$.

Definition 2.2.1 *The values*

$$\mathsf{SE}(\widehat{\theta}_0) = \sqrt{\frac{\sum\limits_{i=1}^{n}(Y_i - \widehat{\theta}_0 - \widehat{\theta}_1 x_i)^2}{n-2}} \cdot \sqrt{\frac{1}{n} + \frac{\overline{X_1}^2}{\sum\limits_{i=1}^{n}\left(x_i - \overline{X_1}\right)^2}},$$

$$\mathsf{SE}(\widehat{\theta_1}) = \sqrt{\frac{\sum\limits_{i=1}^{n}(Y_i - \widehat{\theta_0} - \widehat{\theta_1}x_i)^2}{n-2}} \cdot \sqrt{\frac{1}{\sum\limits_{i=1}^{n}\left(x_i - \overline{X_1}\right)^2}}$$

are standard errors of estimators $\widehat{\theta_0}$ and $\widehat{\theta_1}$ respectively.

Remark 2.2.2 *It's useful to know how the introduced standard errors are obtained and what they mean. Since*

$$\mathsf{D}\widehat{\theta_0} = \sigma^2 \left(\frac{1}{n} + \frac{\overline{X_1}^2}{\sum\limits_{i=1}^{n}(x_i - \overline{X_1})^2} \right),$$

and

$$\mathsf{D}\widehat{\theta_1} = \frac{\sigma^2}{\sum\limits_{i=1}^{n}(x_i - \overline{X_1})^2},$$

the introduced standard errors are nothing but the standard deviation estimators $\widehat{\theta_0}$ and $\widehat{\theta_1}$ while replacing σ^2 by its unbiased estimator obtained earlier, that is, by the estimator:

$$\widehat{\sigma_0^2} = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{\theta_0} - \widehat{\theta_1}x_i)^2.$$

We can construct a so-called confidence interval based on the standard errors. Recall the definition.

Definition 2.2.2 *Assume that $0 < \varepsilon < 1$. The confidence interval (θ^-, θ^+) of the confidence level $1 - \varepsilon$ is the interval within which the real value of the parameter would fall with the probability of $1 - \varepsilon$.*

In practice, we often consider the values $\varepsilon = 0.1$, $\varepsilon = 0.05$, or $\varepsilon = 0.01$.

Theorem 2.2.1 (Confidence intervals for θ_0 and θ_1) *According to the assumptions 1-4, the confidence interval at the confidence level of $(1 - \varepsilon)$ for the parameter θ_0 is the interval*

$$\left(\widehat{\theta_0} - t_{1-\varepsilon/2} \cdot \mathsf{SE}(\widehat{\theta_0}),\ \widehat{\theta_0} + t_{1-\varepsilon/2} \cdot \mathsf{SE}(\widehat{\theta_0})\right),$$

where $t_{1-\varepsilon/2}$ is $(1 - \varepsilon/2)$ quantile of the Student's t-distribution with $(n - 2)$ degrees of freedom.

The confidence interval at the confidence level of $(1-\varepsilon)$ *for the parameter* θ_1 *is the interval*

$$\left(\widehat{\theta_1} - t_{1-\varepsilon/2}\cdot \mathsf{SE}(\widehat{\theta_1}),\ \widehat{\theta_1} + t_{1-\varepsilon/2}\cdot \mathsf{SE}(\widehat{\theta_1})\right),$$

where $t_{1-\varepsilon/2}$ *is* $(1-\varepsilon/2)$ *quantile of the Student's t-distribution with* $(n-2)$ *degrees of freedom.*

Proof. Let's prove the second relation. The first one is obtained similarly. As follows from the lemma about properties of the estimator $\widehat{\theta_1}$,

$$\widehat{\theta_1} \sim \mathsf{N}_{\theta_1, \frac{\sigma^2}{\sum\limits_{i=1}^{n}(x_i-\overline{X_1})^2}}.$$

Thus, using the properties of linear transformations,

$$\frac{\theta_1 - \widehat{\theta_1}}{\sqrt{\frac{\sigma^2}{\sum\limits_{i=1}^{n}(x_i-\overline{X_1})^2}}} \sim \mathsf{N}_{0,1}.$$

According to the note made in the beginning of this section,

$$\frac{n\widehat{\sigma^2}}{\sigma^2} \sim \chi^2_{n-2}.$$

Thus, due to independence, the random variable

$$t_{n-2} = \left(\frac{\theta_1 - \widehat{\theta_1}}{\sqrt{\frac{\sigma^2}{\sum\limits_{i=1}^{n}(x_i-\overline{X_1})^2}}}\right) : \left(\sqrt{\frac{n\widehat{\sigma^2}}{\sigma^2(n-2)}}\right) \sim \mathsf{T}_{n-2}$$

has the Student's t-distribution with $(n-2)$ degrees of freedom. The last one is equivalently rewritten as follows:

$$t_{n-2} = \frac{\theta_1 - \widehat{\theta_1}}{\mathsf{SE}(\widehat{\theta_1})}.$$

The subsequent construction of the confidence interval is typical. □

2.3 Interpreting Confidence Intervals

Let's look at the examples of phone and plane sales that we discussed in the very beginning.

In the phone example, the confidence interval $\left(\theta_0^-,\ \theta_0^+\right)$ for θ_0 (when $\varepsilon = 0.1$) takes the form

$$\left(\theta_0^-,\ \theta_0^+\right) = (3.88,\ 5.87)\,,$$

for θ_1 it takes the form

$$\left(\theta_1^-,\ \theta_1^+\right) = (0.28,\ 0.31)\,.$$

Later, we will review the detailed calculation, but now let's take some time to think things over. Recall that the regression equation looks as follows:

$$Y = \theta_0 + \theta_1 X_1.$$

Based on the calculations, we can conclude that, without advertising (when $X_1 = 0$), sales will drop to $3.88 - 5.87$ thousand units on average (since the predicted value is $Y = \theta_0$). Moreover, for each \$1,000 increase in advertising, there will be an average increase in sales by $0.28 - 0.31$ thousand units.

In the aircraft example, the confidence interval of confidence level 0.9 for θ_0 takes the form

$$\left(\theta_0^-,\ \theta_0^+\right) = (4.73,\ 5.52)$$

for θ_1 it takes the form

$$\left(\theta_1^-,\ \theta_1^+\right) = (1.12,\ 1.57).$$

Based on these results, we can conclude that, on average, sales will drop down to $4.73 - 5.52$ thousand units if there's no advertising. Moreover, for each hundred thousand increase in advertising, there will be an average increase in sales by $1.12 - 1.57$ units.

2.4 Confidence Intervals. Example

Let's return to the shopping time example and calculate the confidence intervals for the model parameters. The formulas for ten inputs can be written as follows:

$$\mathsf{SE}(\widehat{\theta_0}) = \sqrt{\frac{\sum\limits_{i=1}^{10}(Y_i - \widehat{\theta_0} - \widehat{\theta_1}x_i)^2}{10-2}} \cdot \sqrt{\frac{1}{10} + \frac{\overline{X_1}^2}{\sum\limits_{i=1}^{10}\left(x_i - \overline{X_1}\right)^2}},$$

$$\mathsf{SE}(\widehat{\theta}_1) = \sqrt{\frac{\sum\limits_{i=1}^{10}(Y_i - \widehat{\theta}_0 - \widehat{\theta}_1 x_i)^2}{10-2}} \cdot \sqrt{\frac{1}{\sum\limits_{i=1}^{10}\left(x_i - \overline{X_1}\right)^2}},$$

$$\overline{X_1} = \frac{1}{10}\sum_{i=1}^{10} x_i.$$

Let's see how to find $t_{1-\varepsilon/2}$ using the formula for the confidence interval. So, we have $n = 10$, that is, ten degrees of freedom. Assume that $\varepsilon = 0.1$, then $1-\varepsilon/2 = 0.95$, $n-2 = 8$, therefore, we open the table provided in the additional materials and find the value at the intersection of the eighth row and the column that corresponds to the probability 0.95. In our case, $t_{0.95} \approx 1.86$. Since $\mathsf{SE}(\widehat{\theta}_0) = 0.91$ and $\mathsf{SE}(\widehat{\theta}_1) = 0.13$,

$$\left(\theta_0^-,\ \theta_0^+\right) = (2.37,\ 5.75)$$

and

$$\left(\theta_1^-,\ \theta_1^+\right) = (0.69,\ 1.17)\,.$$

What conclusion can be made based on the calculations? Let's begin with considering the second interval. On average, an increase in the number of goods by one increases the shopping time from 0.69 to 1.17 minutes in 90% of cases (because $\varepsilon = 0.1$, therefore, the probability $1-\varepsilon = 0.9$). The first interval shows that one can buy $3-4$ products on average during a zero-minute trip. This anomaly is caused by a small amount of real-life data and the model error (because the relationship is not completely linear). If we assume that we spend at least one minute in the store, the anomaly disappears.

2.5 Hypothesis Testing

Standard errors are also used in the so-called problem of hypothesis testing. One of the most frequently tested hypotheses is a hypothesis about statistical significance of the parameter $\widehat{\theta}_1$. It's formulated as follows:

$$\mathsf{H}_0:\ \text{There is no relationship between } X_1 \text{ and } Y.$$

The alternative hypothesis is

$$\mathsf{H}_a:\ \text{There is a relationship between } X_1 \text{ and } Y.$$

From the mathematical point of view, null and alternative hypotheses state that

$$\mathsf{H}_0:\ \theta_1 = 0,$$

$$\mathsf{H}_a: \ \theta_1 \neq 0.$$

Indeed, when $\theta_1 = 0$, the model is rewritten in the form $Y = \theta_0$, and the values X are not taken into account at all.

To test the hypothesis, it is necessary to determine how far from null the value of our estimator $\widehat{\theta_1}$ of the true value of θ_1 is. Clearly, it depends on the standard error $\mathsf{SE}(\widehat{\theta_1})$. When the latter is low, even sufficiently low values of $\widehat{\theta_1}$ can prove that $\theta_1 \neq 0$. When the error is high, the value $|\widehat{\theta_1}|$ must be high too in order to reject the null hypothesis. In practice, the Student's t-test is usually used. To do this, we compute statistics

$$t = \frac{|\widehat{\theta_1} - 0|}{\mathsf{SE}(\widehat{\theta_1})} = \frac{|\widehat{\theta_1}|}{\mathsf{SE}(\widehat{\theta_1})}.$$

Comparing the real and table value of $t_{1-\varepsilon/2}$ at the confidence level $1-\varepsilon$ with the number of degrees of freedom $(n-2)$, a decision is made:

1. If $t_{1-\varepsilon/2} < t$, the hypothesis H_0 is rejected, and the estimator $\widehat{\theta_1}$ is considered statistically significant at significance level ε.

2. If $t_{1-\varepsilon/2} \geq t$, the hypothesis H_0 is accepted, and the estimator $\widehat{\theta_1}$ is considered statistically insignificant at significance level ε.

Of course, we are interested in the first one, otherwise, in terms of statistics, our model does not reflect the real relationship between the variables.

In the phone example, when $\varepsilon = 0.1$, we get the value $t = 28.49$ that is greater than the value $t_{1-\varepsilon/2} \approx 1.66$. Thus, the hypothesis H_0 is rejected, and the alternative hypothesis H_a is accepted. In the plane example, we obtain the value $t = 9.81$, that is also greater than the value $t_{1-\varepsilon/2}$ from the table. Thus, the hypothesis H_0 is rejected, and the alternative hypothesis H_a is accepted.

Remark 2.5.1 *We can also test the hypothesis about equality of θ_1 to a specific value. To do so, it makes sense to use the statistics*

$$t = \frac{|\widehat{\theta_1} - \theta_1|}{\mathsf{SE}(\widehat{\theta_1})}.$$

The remaining actions are the same as in the discussed algorithm.

Remark 2.5.2 *We can similarly test the hypotheses with respect to the values of the parameter θ_0. We will discuss it together with multivariate regression.*

2.6 Hypothesis Testing. Example

We have calculated all the necessary values. Let's assume $\varepsilon = 0.1$. In our case

$$t = \frac{|\widehat{\theta_1}|}{\mathsf{SE}(\widehat{\theta_1})} = \frac{0.93}{0.13} \approx 7.15.$$

which is larger than 1.86. It means that the hypothesis H_0 is rejected, and the alternative hypothesis H_a is accepted. Thus, a non-zero response to the predictor is set, and there is a relationship.

2.7 Estimating Model Accuracy

If the null hypothesis is rejected in favor of an alternative hypothesis, we can determine the degree to which the model fits the data. Usually this 'estimate' of linear regression is given by two values: residual standard error (RSE) and R^2 statistics.

The model clearly shows that each trial contains the error ε. Because of this error, even if we know the real values of the coefficients θ_0 and θ_1, we cannot accurately predict the value Y, when the value X_1 is known. RSE is an estimator of the residual standard error σ^2 of ε. Roughly speaking, it shows how the constructed model differs from the true one and estimates the averaged square root of the sum of the squared errors of the model. RSE, as we know, can be computed using the formula

$$\widehat{\sigma_0} = \mathsf{RSE} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{\theta_0} - \widehat{\theta_1}x_i)^2}.$$

Since RSE is measured in the same units as Y, it is not always clear if the obtained parameter is good. R^2 statistics is dimensionless unlike RSE, and it lies between zero and one. To compute R^2, we will use the following formula:

$$\mathsf{R}^2 = 1 - \frac{\sum\limits_{i=1}^{n}(Y_i - \widehat{\theta_0} - \widehat{\theta_1}x_i)^2}{\sum\limits_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}.$$

If the constructed model perfectly matches the input data, the terms in the last fraction of the numerator are equal to zero, and if and only if $\mathsf{R}^2 = 1$, the model is ideal. On the contrary, if $\widehat{\theta_1} = 0$, that is, the model does not depend on X_1, then $\mathsf{R}^2 = 0$ (since, when $\widehat{\theta_1} = 0$, we have $\widehat{\theta_0} = \overline{Y}$), and the model is unsound because it doesn't reflect the relationship between the response and predictor.

Notwithstanding that R^2 statistics values lie between zero and one, we still cannot say what value R^2 is good. For example, in some physics problems, we know for sure that the relationship is linear and has an insignificant error. That's why we expect the coefficient to be very close to one. A small coefficient will indicate a serious problem with the experiment from which the data was taken. In many other fields, such as biology and marketing, the linear model is a rough approximation of the data, and errors are often high.

Recall that the sample correlation between X_1 and Y is defined as follows:

$$r\left(X_1, Y\right) = \frac{\sum\limits_{i=1}^{n}\left(x_i - \overline{X_1}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum\limits_{i=1}^{n}\left(x_i - \overline{X_1}\right)^2}\sqrt{\sum\limits_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}}$$

It turns out that $\mathsf{R}^2 = r^2(Y, \widehat{\theta_0} + \widehat{\theta_1}X_1)$.

In the phone example, we get $\mathsf{RSE} = 3.04$ and $\mathsf{R}^2 = 0.89$. According to the aforementioned, we can conclude that our model works well from the statistical point of view and is able to describe the considered relationship.

In the plane example, we get $\mathsf{RSE} = 1.73$ and $\mathsf{R}^2 = 0.49$. The characteristics of this model are much worse compared to the previous one.

2.8 Estimating Model Accuracy. Example

For our example, we rewrite the formulas as follows:

$$\mathsf{RSE} = \sqrt{\frac{1}{10-2}\sum_{i=1}^{10}\left(Y_i - 4.06 - 0.93 \cdot x_i\right)^2},$$

$$\mathsf{R}^2 = 1 - \frac{\sum\limits_{i=1}^{10}\left(Y_i - 4.06 - 0.93 \cdot x_i\right)^2}{\sum\limits_{i=1}^{10}\left(Y_i - \overline{Y}\right)^2}.$$

The calculations show that

$$\mathsf{RSE} = 2.77, \quad \mathsf{R}^2 = 0.87,$$

which indicates that the model is good from the statistical point of view.

3 Multivariate Linear Regression

The simple linear regression helps to predict the value of one variable when another variable is known. In practice, the desired value often depends on more than one variable. Say, sales depend on the amount spent on phone advertising as well as on the plane advertising budget. How can we extend the analysis to more variables?

3.1 Basic Definitions and Matrix Notations

Having thoroughly studied the univariate regression, we can write the multivariate linear regression model as:

$$Y = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + ... + \theta_p X_p + \varepsilon,$$

where $X_1, X_2, ..., X_p$ are (non-random) input data used to determine the (random) variable Y (response), and ε is a random error. Well, the relationship between Y and X_1, X_2, ..., X_p is assumed to be linear with the accuracy of some error ε.

Definition 3.1.1 *The model described by the relationship*

$$Y = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + ... + \theta_p X_p + \varepsilon,$$

where $\theta_0, \theta_1, ..., \theta_p$ are numeric parameters, $X_1, X_2, ..., X_p$ are non-random parameters, which values are either set or observed (to put it differently, known), ε is a random error, called a simple multivariate linear regression model.

The following two definitions are often used.

Definition 3.1.2 *A function*

$$f(X_1, X_2, ..., X_p) = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + ... + \theta_p X_p$$

in a multivariate linear regression model is called the line of regression of Y on X_1, X_2, ..., X_p.

Definition 3.1.3 *An equation*

$$Y = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + ... + \theta_p X_p$$

in a multivariate linear regression model is called the regression equation of Y on X_1, X_2, ..., X_p.

How to define the model coefficients on concrete observed values of X_1, X_2, ..., X_p and Y? Let's describe the scheme in detail. To begin with, we can carefully write what we have here.

On the ith observation, we get the value Y_i with respect to the values $(x_{i1}, x_{i2}, ..., x_{ip})$ of the predictors $X_1, X_2, ..., X_p$ respectively. On each observation, the equality is satisfied:

$$Y_i = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + ... + \theta_p x_{ip} + \varepsilon_i, \quad i \in \{1, 2, ..., n\}$$

Further on, we will use matrix notations for convenience. Let's introduce designations $x_{10} = x_{20} = ... = x_{n0} = 1$ and

$$X = \begin{pmatrix} x_{10} & x_{11} & \dots & x_{1p} \\ x_{20} & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0} & x_{n1} & \dots & x_{np} \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \quad \Theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_p \end{pmatrix}.$$

Then our model can be rewritten in a matrix form as

$$Y = X \cdot \Theta + \Sigma.$$

Remark 3.1.1 *The introduction of additional notations adds one more predictor X_0 that takes one and the same value equal to 1 on each observation. Thus, the considered model*

$$Y = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + ... + \theta_p X_p + \varepsilon$$

is equivalently rewritten as follows:

$$Y = \theta_0 X_0 + \theta_1 X_1 + \theta_2 X_2 + ... + \theta_p X_p + \varepsilon.$$

Let's also consider another common definition.

Definition 3.1.4 *The introduced matrix X is often called a regressor.*

Now let's move on to finding the coefficients of the multivariate regression.

3.2 OLS for Multivariate Regression

We will apply OLS to find the estimators of the unknown parameters in the same way as in the univariate regression case. Thus, we can find the estimators $\widehat{\theta_0}, \widehat{\theta_1}, ..., \widehat{\theta_p}$ of the coefficients $\theta_0, \theta_1, ..., \theta_p$ by minimizing the function $\varepsilon(\theta_0, \theta_1, ..., \theta_p)$ dependent on $(p + 1)$ variable.

$$\varepsilon(\theta_0, \theta_1, ..., \theta_p) = \sum_{i=1}^{n} \left(Y_i - \theta_0 - \theta_1 x_{i1} - \theta_2 x_{i2} - ... - \theta_p x_{ip}\right)^2.$$

Since we need the coefficients minimizing the function instead of the smallest value of the function, the following problem is solved:

$$\arg\min_{\theta_0,\ldots,\theta_p} \sum_{i=1}^{n} \left(Y_i - \theta_0 - \theta_1 x_{i1} - \theta_2 x_{i2} - \ldots - \theta_p x_{ip}\right)^2 .$$

Definition 3.2.1 *An estimator of the ordinary least squares method for unknown parameters θ_0, θ_1, ...,θ_p of the multivariate regression model is a set of parameter values minimizing the expression*

$$\varepsilon(\theta_0, \theta_1, \ldots, \theta_p) = \sum_{i=1}^{n} \left(Y_i - \theta_0 - \theta_1 x_{i1} - \theta_2 x_{i2} - \ldots - \theta_p x_{ip}\right)^2 .$$

We can solve the problem as earlier but the calculations will be more difficult and cumbersome. Let's formulate the final theorem and use geometrical considerations to explain it.

Theorem 3.2.1 *Let the columns of the regressor X be linearly independent, and $n > (p+1)$ (the number of observations exceeds the number of unknown model parameters). The minimum of the function*

$$\varepsilon(\theta_0, \theta_1, \ldots, \theta_p) = \sum_{i=1}^{n} \left(Y_i - \theta_0 - \theta_1 x_{i1} - \theta_2 x_{i2} - \ldots - \theta_p x_{ip}\right)^2$$

is unique and attained when

$$\Theta = \left(X^T X\right)^{-1} X^T Y.$$

Proof. Let's justify the case when $p = 1$, that is, given two unknown parameters θ_0 and θ_1. The justification in the general case is the same. However, it's less geometric.

Let X_0, X_1 be the columns of regressor X. They are linearly independent, which means they derive the two-dimensional space $L = \mathbb{R}^2$. Moreover,

$$X\Theta = X_0\theta_0 + X_1\theta_1 \in L.$$

The geometrical considerations (fig. 11) make it clear that, according to the ordinary least squares method, we minimize the square of the length of $Y - X\Theta$ (the last vector is shown by a dashed line). The length (and the squared length) is minimal when the vector $Y - X\Theta$ is orthogonal to L, that is, orthogonal to each vector X_0 and X_1:

$$(Y - X\Theta) \perp X_i,\ i = 0, 1.$$

To put it differently,

$$X^T(Y - X\Theta) = 0 \Leftrightarrow X^TY - X^TX\Theta = 0.$$

Since $\det(XX^T) \neq 0$,

$$\Theta = (X^TX)^{-1}X^TY.$$

□

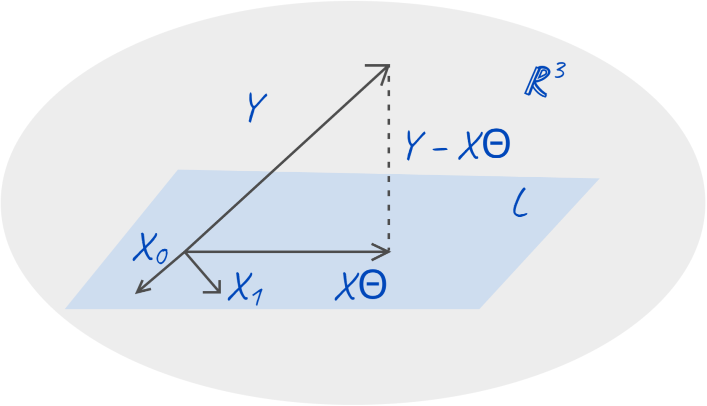

Figure 11: The proof.

Therefore,

$$\widehat{\Theta} = (X^TX)^{-1}X^TY.$$

Remark 3.2.1 *In practice, the obtained formulas are rarely used in their original form. Anyway, the respective functions are built in most data analysis packages.*

When the estimators of the model coefficients are known, and we can make a prediction using the following formula:

$$Y = \widehat{\theta}_0 + \widehat{\theta}_1X_1 + \widehat{\theta}_2X_2 + ... + \widehat{\theta}_pX_p.$$

Let's look at the example of creating a multivariate linear regression model based on the considered cases. We pose a question: how does the total sales volume of planes and phones depend on the advertising budget of each product. The input data distribution is shown in figure 12. In our case, the number of predictors equals two. Thus, the model takes the form:

$$Y = \theta_0 + \theta_1X_1 + \theta_2X_2 + \varepsilon.$$

Let's find the estimators $\widehat{\theta}_0$, $\widehat{\theta}_1$, and $\widehat{\theta}_2$ of the unknown coefficients θ_0, θ_1, and θ_2. The above formulas give

$$\widehat{\theta}_0 = 27.20, \quad \widehat{\theta}_1 = 1.08, \quad \widehat{\theta}_2 = 0.86,$$

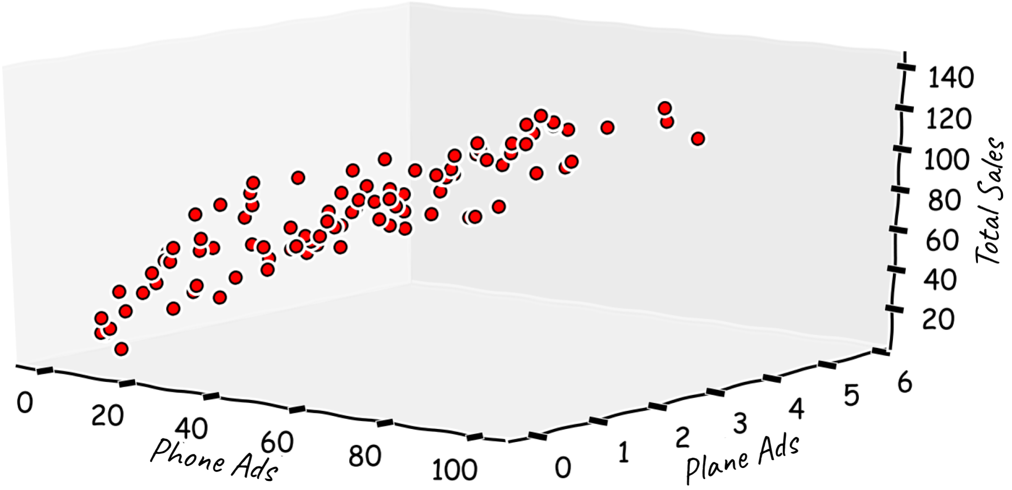

Figure 12: The relationship between total sales volume and advertising budget for planes and phones

and the prediction is made based on the relation

$$Y = 27.20 + 1.08 \cdot X_1 + 0.86 \cdot X_2.$$

The figure 13 show the attempts at illustrating the regression line (plane) and data spread from different angles.

3.3 Statistical Estimation of Parameters of Multivariate Linear Regression

As in the simple linear regression case, we can discuss in detail all possible statistical characteristics of parameters of multivariate regression model. However, we won't do this. Since, on the one side, the entire scheme is very similar, and, on the other side, it is technically challenging. So here we will limit ourselves to a brief overview of the important results.

Well, we will make the following important assumptions (the first three are often called Gauss–Markov assumptions):

1. Random variables (errors) $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ are independent and identically distributed.

2. Errors are not systematic: $\mathsf{E}\varepsilon_i = 0$, $i \in \{1, 2, ..., n\}$.

3. Error variances are the same: $\mathsf{D}\varepsilon_i = \sigma^2 > 0$, $i \in \{1, 2, ..., n\}$ (homoscedasticity).

4. $\varepsilon_i \sim \mathsf{N}_{0,\sigma^2}$.

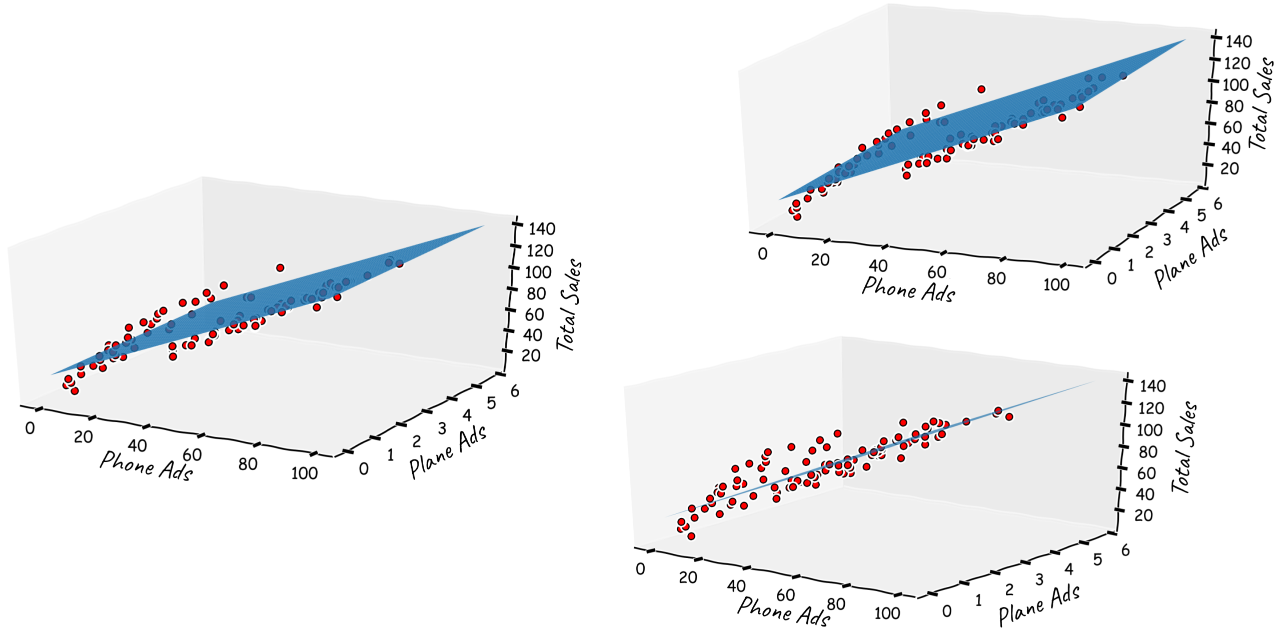

Figure 13: The relationship between total sales volume and advertising budget for planes and phones as well as regression

Without explaining the properties of estimators $\widehat{\Theta}$ (we invite you to formulate them yourself), we will jump right into the practically significant problems — formulas for confidence intervals. First, we are going to obtain the unbiased estimator of the variances σ^2 of the errors ε_i.

Lemma 3.3.1 *According to the assumptions 1-4, the estimator*

$$\widehat{\sigma_0^2} = \frac{1}{n-p-1}\sum_{i=1}^{n}(Y_i - \widehat{\theta_0} - \widehat{\theta_1}x_{i1} - \widehat{\theta_2}x_{i2} - \ldots - \widehat{\theta_p}x_{ip})^2 = \frac{1}{n-p-1}|Y - X\widehat{\Theta}|^2$$

is an unbiased estimator of the parameter σ^2.

Now we can provide the expression for the confidence interval.

Theorem 3.3.1 (A confidence interval for θ_i) *According to the assumptions 1-4, the confidence interval at the confidence level of $(1-\varepsilon)$ for the parameter θ_i is the interval*

$$\left(\widehat{\theta_i} - t_{1-\varepsilon/2}\cdot\widehat{\sigma_0}\sqrt{(X^TX)^{-1}_{(i+1)(i+1)}},\ \widehat{\theta_i} + t_{1-\varepsilon/2}\cdot\widehat{\sigma_0}\sqrt{(X^TX)^{-1}_{(i+1)(i+1)}}\right),$$

where $t_{1-\varepsilon/2}$ is $(1-\varepsilon/2)$ quantile of the Student's t-distribution with $(n-p-1)$ degrees of freedom, and $(X^TX)^{-1}_{(i+1)(i+1)}$ is an element of the matrix $(X^TX)^{-1}$ at the intersection of the $(i+1)$th row and $(i+1)$th column.

If you can construct confidence intervals, you can test hypotheses with respect to the values of regression coefficients, in particular, test the hypothesis about statistical significance of the obtained estimators $\widehat{\theta}_i$, $i \in \{0, 1, ..., p\}$.

Corollary 3.3.2 (About hypothesis testing with respect to θ_i) *Let's test the hypothesis $\theta_i = \theta_i^0 \in \mathbb{R}$ against the alternative hypothesis $\theta_i \neq \theta_i^0$, that is:*

$$\mathsf{H}_0: \ \theta_i = \theta_i^0$$

$$\mathsf{H}_a: \ \theta_i \neq \theta_i^0.$$

Assume that

$$t = \frac{|\widehat{\theta}_i - \theta_i^0|}{\widehat{\sigma}_0\sqrt{(X^TX)^{-1}_{(i+1)(i+1)}}},$$

where $(X^TX)^{-1}_{(i+1)(i+1)}$ is an element of the matrix $(X^TX)^{-1}$ at the intersection of the $(i+1)$th row and $(i+1)$th column and $t_{1-\varepsilon/2}$ is the quantile at level $(1-\varepsilon/2)$ of the Student's t-distribution with $(n-p-1)$ degrees of freedom. Thus,

1. *If $t > t_{1-\varepsilon/2}$, the hypothesis H_0 is rejected at the significance level ε.*
2. *If $t \leq t_{1-\varepsilon/2}$, the hypothesis H_0 is accepted at the significance level ε.*

We've reviewed the examples of application and interpretation of the discussed methods in the simple case, so we will not describe them again in this section.

3.4 Estimating Model Accuracy for Multivariate Regression

As in the simple regression case, we will also consider the estimators of RSE and R^2 for our model. As earlier, the error RSE is defined as follows:

$$\mathsf{RSE} = \widehat{\sigma}_0 = \sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n}\left(Y_i - \widehat{\theta}_0 - \widehat{\theta}_1 x_{i1} - ... - \widehat{\theta}_p x_{ip}\right)^2} =$$

$$= \sqrt{\frac{|Y - X\widehat{\Theta}|^2}{n-p-1}},$$

and shows the averaged sum of the squared errors of the model after the estimation of the parameters θ_0, θ_1, ..., θ_p. As in the simple case, R^2 statistics is introduced:

$$\mathsf{R}^2 = 1 - \frac{\sum\limits_{i=1}^{n}\left(Y_i - \widehat{\theta}_0 - \widehat{\theta}_1 x_{i1} - ... - \widehat{\theta}_p x_{ip}\right)^2}{\sum\limits_{i=1}^{n}(Y_i - \overline{Y})^2}$$

Again, as in the simple case, if the constructed model perfectly matches the input data, the terms in the numerator of the last fraction are equal to zero, and if and only if $\mathsf{R}^2 = 1$, the model is ideal. On the contrary, if $\widehat{\theta_1} = \widehat{\theta_2} = ... = \widehat{\theta_p} = 0$, that is, the model does not depend on X_1, X_2, ..., X_p, then $\mathsf{R}^2 = 0$, and the model is unsound because it doesn't reflect the relationship between the response and predictor. It turns out that the value R^2 only increases when we add new predictors to the model, even if they slightly affect the response.

We can show that the relationship between R^2 and the correlation r^2 in the multivariate regression case is as follows:

$$\mathsf{R}^2 = r^2\left(Y, \widehat{\theta_0} + \widehat{\theta_1}X_1 + .. + \widehat{\theta_p}X_p\right).$$

The multivariate regression case about the sales of phones and planes with respect to the advertising budget has the following coefficients: $\mathsf{R}^2 = 0.89$, and $\mathsf{RSE} = 11.04$. These values indicate a high-quality model.

3.5 Hypothesis about Testing Statistical Significance of Linear Regression Model

One of the most important applications of the estimator R^2 is testing the statistical significance of the obtained model in general. That is, testing the following assumption about the relationship between the response Y and predictors X_1, X_2, ..., X_p?

The null hypothesis is

$$\mathsf{H}_0: \text{ All model parameters } \theta_i \text{ are equal to zero for } i \in \{1, 2, ..., p\},$$

The alternative hypothesis is

$$\mathsf{H}_a: \text{ At least one model parameter } \theta_i \text{ is not equal to zero for } i \in \{1, 2, ..., p\},$$

We can write it shorter as:

$$\mathsf{H}_0: \ \theta_1 = \theta_2 = ... = \theta_p = 0,$$

$$\mathsf{H}_a: \ \theta_1^2 + \theta_2^2 + ... + \theta_p^2 \neq 0.$$

We will use F-statistics to test these hypotheses

$$F = \frac{\mathsf{R}^2}{1 - \mathsf{R}^2}\frac{n - p - 1}{p}.$$

It turns out that, in the assumptions 1-4, the introduced statistics has the F-Distribution with the parameters $(p, n - p - 1)$ (with p and $n - p - 1$ degrees of freedom). The hypothesis testing is carried out as follows.

Let $\varepsilon > 0$ and $f_{1-\varepsilon}$ be a quantile of the F-Distribution with the parameters $(p, n - p - 1)$ at level $(1 - \varepsilon)$.

1. If $F > f_{1-\varepsilon}$, the hypothesis H_0 is rejected at the significance level α. Thus, the constructed model is statistically significant.
2. If $F \leq f_{1-\varepsilon}$, the hypothesis H_0 is accepted at the significance level α. Thus, the constructed model is statistically insignificant.

4 Polynomial Regression

It has been well noted that linear regression assumes that there is a linear relationship between the response and predictors. The relationship may not be linear in real-life problems. It turns out that the linear regression model can be extended to the so-called polynomial regression model without significant complications. The simple polynomial regression model takes the form

$$Y = \theta_0 + \theta_1 X + \theta_2 X^2 + ... + \theta_p X^p + \varepsilon.$$

It is important to note that the model coefficients can be found using the OLS described earlier, because this is purely multivariate linear regression, but, instead of the predictors, we took the powers of X:

$$X_1 = X,\ X_2 = X^2, ..., X_p = X^p.$$

Thus, we can use the linear regression apparatus to create the model (more precisely, to find the estimators $\widehat{\theta_0}$, $\widehat{\theta_1}$, ..., $\widehat{\theta_p}$). X powers higher than the fourth are rare, since polynomial curves can take unpredictable shapes.

The values of response Y_i from the values of predictor x_i can be obtained by the rule that

$$Y_i = \widehat{\theta_0} + \widehat{\theta_1} x_i + \widehat{\theta_2} x_i^2 + ... + \widehat{\theta_p} x_i^p, \quad i \in \{1, 2, ..., n\}.$$

Let's assume that we have data shown in figure 14. The results of the polynomial regression with respect to the different powers p are shown in figure 15. The blue curve is classical linear regression, green is polynomial regression with the second-degree polynomial, violet is polynomial regression with the third-degree polynomial. You can see that the third-degree polynomial approximates the input data better than all the others do.

Polynomial regression is a special case of the model generalizing the linear regression model. We can take some functions φ_1, φ_2, ..., φ_p and construct a more general model

$$Y = \theta_0 + \theta_1 \varphi_1(X) + \theta_2 \varphi_2(X) + ... + \theta_p \varphi_p(X) + \varepsilon,$$

which calculations are similar to those in the polynomial case. Moreover, regression coefficients are estimated as we have already discussed.

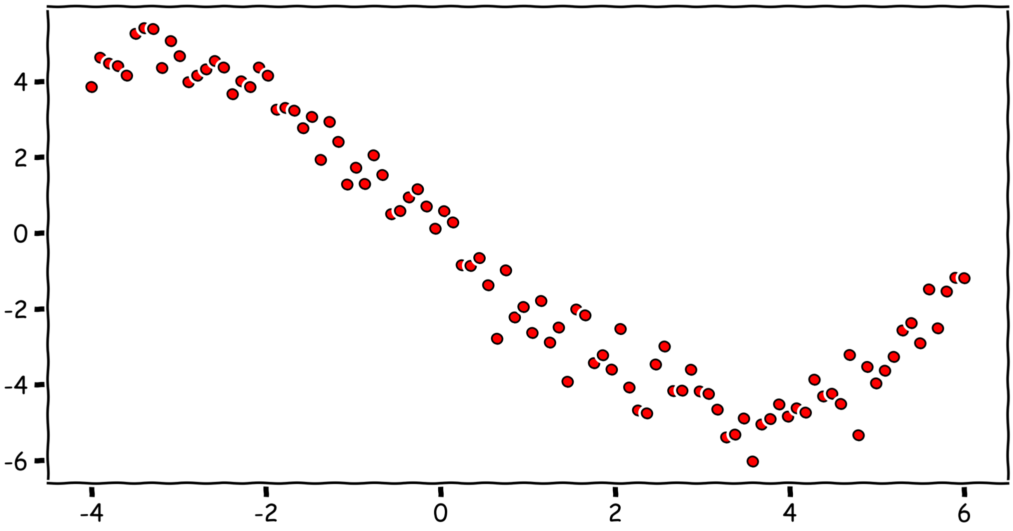

Figure 14: Polynomial regression dataset

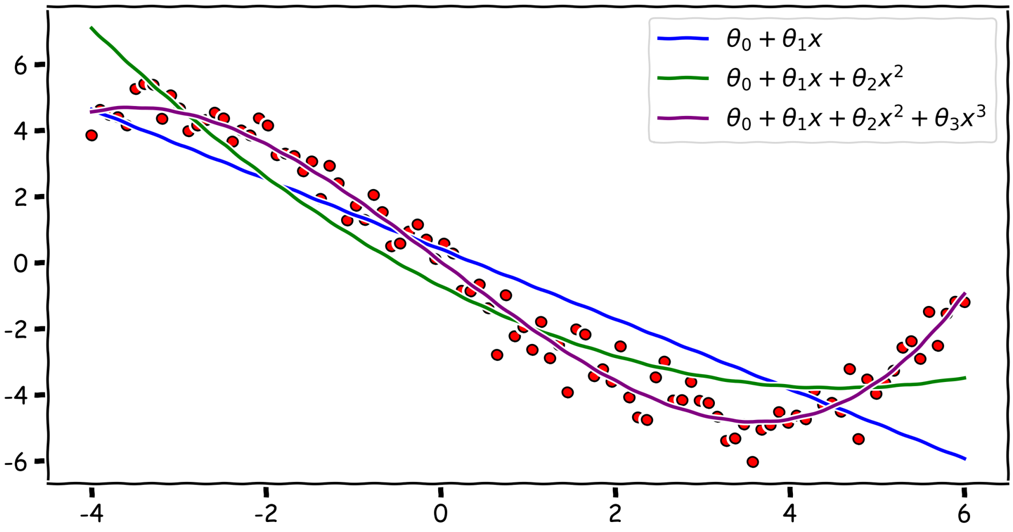

Figure 15: Polynomial regression

5 Conclusion

This time, we have considered the approaches to the regression problem. The apparatus used to solve this problem is well studied and supported by various statistical estimators. However, the choice of a set of functions φ_i used to construct generalized regression is put in charge of a researcher. Good luck!