In few words, the correlation measures how well the association between two variables can be described by a straight line. Let us consider the variable $x$ and the four variables $y$ in Table 22. Figure 31 shows the corresponding scatter plots.

- The association between $x$ and $y_1$ (top–left panel) is clearly linear and very strong, in fact, we have a perfect linear association. This fact is adequately described by the correlation coefficient, we get $r_{xy_1} = -1$. Note that the sign indicates a negative association.
- There is a perfect association between $x$ and $y_2$ (top–right panel), in fact, we have that $y_2 = \exp(x)$. However, the association is not linear and the correlation coefficient equals $r_{xy_2} = 0.6914$ which, in loose words, can be interpreted as "the association between $x$ and $y_2$ can be adequately described by a straight line".
- There is a perfect association between $x$ and $y_3$ (bottom–left panel), in fact, we have that $y_3 = (x-6)^2$. However, the association is not linear and the correlation coefficient equals $r_{xy_3} = 0$, which can be interpreted as "the association between $x$ and $y_3$ cannot be described by a straight line at all".
- The points in the bottom–right panel show no clear association of any type, in particular, they show no linear association. This fact is reflected by the correlation coefficient which is $r_{xy_4} = -0.2545$, indicating that a straight line will be very poor at describing the association between both variables.

| $x$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ |
|---|---|---|---|---|
| 1 | 11 | 3 | 25 | 9 |
| 2 | 10 | 7 | 16 | 7 |
| 3 | 9 | 20 | 9 | 6 |
| 4 | 8 | 55 | 4 | 1 |
| 5 | 7 | 148 | 1 | 10 |
| 6 | 6 | 403 | 0 | 5 |
| 7 | 5 | 1097 | 1 | 3 |
| 8 | 4 | 2981 | 4 | 11 |
| 9 | 3 | 8103 | 9 | 8 |
| 10 | 2 | 22026 | 16 | 4 |
| 11 | 1 | 59874 | 25 | 2 |

Table 22: Different types of association

In R, the function `cor()` can be used for computing the correlation coefficient.

**Correlation matrix** Let us say that we have $J$ numerical variables, $x_1, x_2, \cdots, x_J$. It is possible to calculate $J(J-1)/2$ correlations between pairs of variables and present them in a matrix that is called the *correlation matrix*:

$$\begin{bmatrix} 1 & r_{x_1x_2,U} & \cdots & r_{x_1x_J,U} \\ r_{x_2x_1,U} & 1 & \cdots & r_{x_2x_J,U} \\ \vdots & \vdots & \ddots & \vdots \\ r_{x_Jx_1,U} & r_{x_Jx_2,U} & \cdots & 1 \end{bmatrix}$$

The ones in the diagonal are obtained by taking into account that $r_{x_jx_j,U} = 1$. Alternatively, taking into account that the correlation is symmetric, i.e. that $r_{x_jx_{j'},U} = r_{x_{j'}x_j,U}$, we can simply

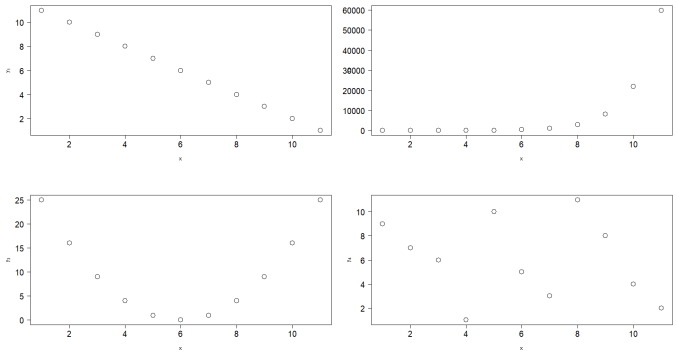

Figure 31: Scatter plots of the variables in Table 22.

write the correlation matrix as

$$\begin{bmatrix} 1 & — & \cdots & — \\ r_{x_2x_1,U} & 1 & \cdots & — \\ \vdots & \vdots & \ddots & \vdots \\ r_{x_Jx_1,U} & r_{x_Jx_2,U} & \cdots & 1 \end{bmatrix}$$

## 5.2 Spearman's correlation coefficient

In the previous subsection we discussed Pearson's correlation coefficient, a measure of the linear association between two variables. In this subsection we will discuss another measure that allows to measure a more general class of associations, namely, monotonic associations. In other words, this measure allows to determine if the association between the two variables can be adequately described by a monotonic function.

First, let us clarify what a monotonic function is. Let us take a look at the functions plotted in Figure 32:

- The top–left panel shows an *increasing* function: as $x$ increases, $y$ also increases;
- The top–central panel shows a *decreasing* function: as $x$ increases, $y$ decreases;
- The top–right panel shows a *constant* function: as $x$ increases, $y$ does not change;
- The bottom–left panel shows a *non–decreasing* function: a function that is partly increasing and partly constant;
- The bottom–central panel shows a *non–increasing* function: a function that is partly decreasing and partly constant.

Any of the five type of functions described above is called a monotonic function. In loose words a monotonic function is a function that is not increasing and decreasing. For instance, the bottom–right panel shows a function that is decreasing first and then becomes increasing, this is not a monotonic function.

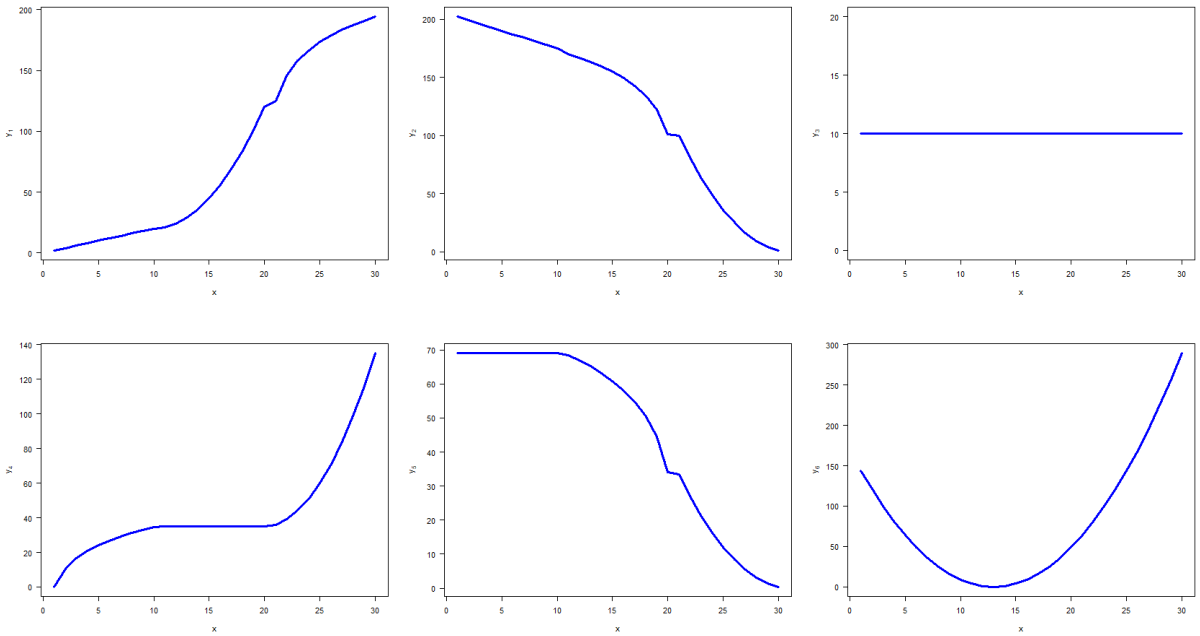

Figure 32: Six functions.

Before defining Spearman's correlation coefficient, we need to define the ranks of the observations of a variable. The rank of the observation $x_i$ in the population $U$, denoted by $R(x_i)$ or simply $R_i$, is the position occupied by the observation when the values are sorted from smallest to largest. This is one of the situations when things are simpler than they sound as illustrated by the following example.

**Example 45.** Consider the population of ten students. The ranks of the number of points in the assignment ($x$) and the exam ($y$) is shown in Table 23.

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $x_i$ | 0 | 9 | 2 | 24 | 25 | 23 | 4 | 20 | 28 | 5 |
| $R(x_i)$ | 1 | 5 | 2 | 8 | 9 | 7 | 3 | 6 | 10 | 4 |
| $y_i$ | 8 | 15 | 5 | 36 | 40 | 30 | 9 | 21 | 32 | 27 |
| $R(y_i)$ | 2 | 4 | 1 | 9 | 10 | 7 | 3 | 5 | 8 | 6 |

Table 23: Ranks of the points of ten students in an exam in Statistics

Spearman's correlation coefficient is simply the correlation coefficient (11) calculated over the ranks of $x$ and $y$, instead of the actual values. More formally:

**Definition 46.** Let $x_i$ and $y_i$ be the values of two variables associated to the $i$th element in $U$ ($i = 1, 2, \cdots, N$). Let also $R(x_i)$ and $R(y_i)$ be their corresponding ranks. *Spearman's correlation coefficient* between $x$ and $y$ is defined as

$$r^s_{xy,U} \equiv \frac{\sum_U (R(x_i) - \bar{R}_U)(R(y_i) - \bar{R}_U)}{\left(\sum_U (R(x_i) - \bar{R}_U)^2 \sum_U (R(y_i) - \bar{R}_U)^2\right)^{1/2}} \tag{12}$$

where $\bar{R}_U = (N+1)/2$. □

Let us illustrate the steps needed to calculate Spearman's coefficient.

**Example 47.** Let $U$ be the population of $N = 10$ students taking a Master course in statistics. Let $x_i$ and $y_i$ be, respectively, the scores in a home assignment and the final exam of the $i$th student ($i = 1, 2, \cdots, N$). The ranks of $x$ and $y$ were shown in Table 23. Using them let us find Spearman's correlation coefficient. We have $\bar{R}_U = (N+1)/2 = (10+1)/2 = 5.5$.

Let us calculate the numerator first:

$$\sum_U (R(x_i) - \bar{R}_U)(R(y_i) - \bar{R}_U) = (1-5.5)(2-5.5) + (5-5.5)(4-5.5) + \cdots + (4-5.5)(6-5.5) = 15.75 + 0.75 + \cdots + -0.75 = 75.5.$$

Now we calculate the first term in the denominator

$$\sum_U (R(x_i) - \bar{R}_U)^2 = (1-5.5)^2 + (5-5.5)^2 + \cdots + (4-5.5)^2 = 20.25 + 0.25 + \cdots + 2.25 = 82.5.$$

And the second term in the denominator is

$$\sum_U (R(y_i) - \bar{R}_U)^2 = (2-5.5)^2 + (4-5.5)^2 + \cdots + (6-5.5)^2 = 12.25 + 2.25 + \cdots + 0.25 = 82.5.$$

This gives

$$r^s_{xy,U} = \frac{\sum_U (R(x_i) - \bar{R}_U)(R(y_i) - \bar{R}_U)}{\left(\sum_U (R(x_i) - \bar{R}_U)^2 \sum_U (R(y_i) - \bar{R}_U)^2\right)^{1/2}} = \frac{75.5}{(82.5 \cdot 82.5)^{1/2}} = 0.9152,$$

which means that there is a high monotonic association between the number of points that students get in the assignment and the exam. □

We have insisted in that Spearman's correlation allows to measure monotonic associations between two variables. Let us illustrate this with the variables in Table 22 and Figure 31:

- We found earlier that for the perfect linear association shown in the top–left panel, Pearson's correlation coefficient is equal to -1. This is the same value obtained by Spearman's correlation coefficient.
- We found that Pearson's correlation coefficient for the increasing association shown in the top–right panel is 0.6914. Spearman's correlation coefficient is equal to 1. Indicating that there is a perfect monotonic association between $x$ and $y$: higher values of $x$ are associated to higher values of $y$.
- Spearman's correlation for the variables shown in the bottom–left panel equals 0, the same value as Pearson's correlation. This indicates that there is no monotonic association between $x$ and $y$.
- Due to the the way the numbers were generated, Pearson's and Spearman's coefficient correlation also coincide for the variables illustrated in the bottom–right panel: -0.2545. Indicating a small monotonic association between the two variables.

## 5.3 Kendall's correlation coefficient

Another parameter that allows for measuring the monotonic association between two variables is Kendall's correlation coefficient. In words, it works as follows:

It compares all pairs of observations and identifies them as either *concordant* or *discordant*. A pair $(x_i, y_i)$ and $(x_j, y_j)$ is said to be concordant if the straight line that connects both points

has a positive slope. A pair $(x_i, y_i)$ and $(x_j, y_j)$ is said to be discordant if the straight line that connects both points has a negative slope. This is illustrated in Figure 33 for a few pairs. The three green segments represent three pairs that are concordant. The three red segments represent three pairs that are discordant.

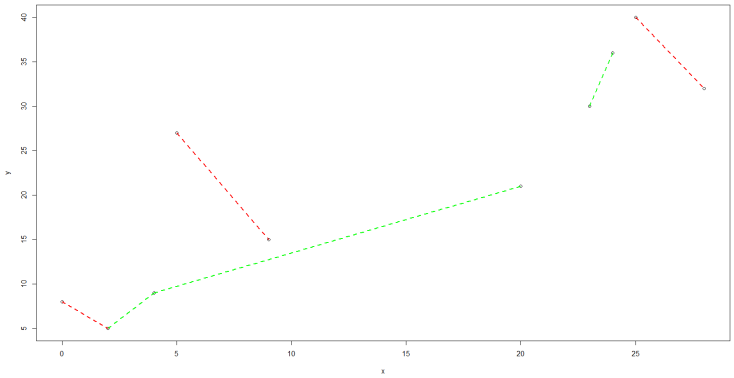

Figure 33: Three concordant (green) and three discordant (red) pairs in the scores of ten students.

Once we have identified concordant and discordant pairs, we take the total number of concordant pairs minus the total number of discordant pairs and divide this by the total number of pairs. More formally:

**Definition 48.** Let $x_i$ and $y_i$ be the values of two variables associated to the $i$th element in $U$ $(i = 1, 2, \cdots, N)$. *Kendall's correlation coefficient* between $x$ and $y$ is defined as

$$r^k_{xy,U} \equiv \frac{2}{n(n-1)} \sum_{i<j} \text{sgn}(x_j - x_i)\text{sgn}(y_j - y_i) \tag{13}$$

where

$$\text{sgn}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases} \qquad \square$$

Using R, we find that Kendall's correlation coefficient of the number of points of the students in the home assignment and the exam is 0.7778.

Up to this point we have introduced a set of *parameters* which are useful for describing different characteristics of one or two *variables* measured on the individuals of one *population* of interest. Note that we have highlighted in italics three words that are crucial to understand the theory that we have covered: parameters, variables and population. We measure or observe variables on the individuals of one population and these observations are used to compute the parameters of interest. The notation we are using takes these three concepts into account: we assign different symbols to the different parameters that have been defined, for instance, we use a bar ( $\bar{}$ ) to denote the mean, a breve ( $\breve{}$ ) to denote the median, an $r$ to denote the correlation, and so on. Furthermore, we explicitly indicate which variable or variables are involved in the calculation of the parameter and what population or set of individuals we are talking about. For instance, $\bar{x}_U$ means that we are calculating the mean of the variable $x$ for the individuals of the set $U$, $r_{xy,U}$ means that we are calculating the correlation between the variables $x$ and $y$ for the individuals of the population $U$. Mastering this notation is a very important and useful step for understanding the remaining part of the course.

# 6 Simple linear regression: the descriptive approach

Consider the situation where we have measurements of $K+1$ variables $x_1, x_2, \cdots, x_K$ and $y$ on a set $U$ of $N$ elements, i.e. the information available for the $i$th element is of the type $(x_{1,i}, x_{2,i}, \cdots, x_{K,i}, y_i)$ for all $i = 1, 2, \cdots, N$. This information is typically stored in a rectangular array as shown in Table 24.

| $\mathbf{x_1}$ | $\mathbf{x_2}$ | $\cdots$ | $\mathbf{x_K}$ | $\mathbf{y}$ |
|---|---|---|---|---|
| $x_{11}$ | $x_{21}$ | $\cdots$ | $x_{J1}$ | $y_1$ |
| $x_{12}$ | $x_{22}$ | $\cdots$ | $x_{J2}$ | $y_2$ |
| $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| $x_{1N}$ | $x_{2N}$ | $\cdots$ | $x_{KN}$ | $y_N$ |

Table 24: Array collecting the information in a regression problem

With this information we want to describe $y$ as a function of the $x$-variables, i.e. we want to express $y$ as $y = f(x_1, x_2, \cdots, x_K)$. To begin with, in this section we will consider the case with only one $x$-variable. Later in Section 7 we will return to the more general case with $K$ variables $x$. Thus the situation in this section is as follows.

We have measurements of two variables $x$ and $y$ on a set $U$ of $N$ elements, i.e. the information available is of the type $(x_1, y_1)$, $(x_2, y_2)$, $\cdots$, $(x_N, y_N)$. With this information we want to describe $y$ as a function of $x$, i.e. we want to express $y$ as $y = f(x)$. To begin with we will consider the case where $f(x) = b_0 + b_1 x$ . Therefore our task is to express $y$ as a linear function of $x$ *as best as possible*. Needless to say, if we want to express $y$ as a linear function of $x$ it is because we have reasons to believe that there is a linear relationship between the two variables.

We illustrate the idea with an example.

**Example 49.** Let us consider a set of $N = 10$ companies, $U$, producing tables, let $x_i =$ number of workers in the $i$th company and $y_i =$ number of tables produced during one particular day by the $i$th company. Table 25 shows the values of $x$ and $y$ and Figure 34 shows a scatter plot of both variables. The scatter plot reveals some association between both variables, a higher number of workers produce a higher number of tables. Furthermore, this association can be adequately described by a straight line. □

| $i$ | $x_i$ | $y_i$ |
|---|---|---|
| 1 | 12 | 20 |
| 2 | 14 | 21 |
| 3 | 15 | 27 |
| 4 | 18 | 30 |
| 5 | 19 | 32 |
| 6 | 24 | 50 |
| 7 | 26 | 54 |
| 8 | 27 | 57 |
| 9 | 28 | 61 |
| 10 | 30 | 60 |

Table 25: Number of tables $y$ produced by $x$ workers

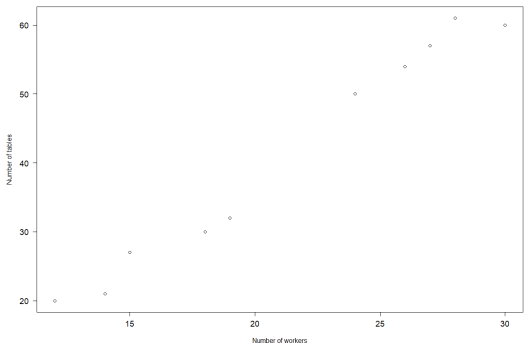


Figure 34: Scatter plot of number of workers $x$ and number of tables $y$.

We have said that we want to express $y$ as a function of $x$. In the previous example we want to express the number of tables as a function of the number of workers. It makes sense to believe that the number of tables *depends* on the number of workers, not the other way around. For this reason, we call $y$ the dependent variable and $x$ the independent variable. In some contexts $x$ and $y$ are given different names, for instance, $y$ is also called the *output* or the *response*; and $x$ is also called the *input* or the *explanatory* variable.

Once the problem has been identified, i.e. once we have decided that we have a pair of variables $x$ and $y$ one of which we would like to express as a linear function of the other one *as best as possible*, the next natural step is to fit this line. However, there are infinitely many lines, how do we choose one that is appropriate for our purposes? In other words, how do we choose the *best possible* straight line that relates $x$ and $y$? Some lines are, evidently, not adequate. For instance, in the top-left panel of Figure 35 we have fitted the line $\hat{y}_1 = 80 - 2\,x$ to the workers dataset. This line clearly does not describe in an adequate way the relation between both variables. In the top-right panel, we have fitted the line $\hat{y}_2 = 20 + 0.25\,x$, which does not look too bad for some values but looks quite bad for some others. On the other hand, in the bottom-left panel, we have fitted the line $\hat{y}_3 = -20 + 3\,x$ and in the bottom-right panel, we have fitted the line $\hat{y}_4 = -13 + 2.5\,x$. Both these lines seem to adequately describe the relation between $x$ and $y$. Which one should we prefer? Can we find any other line that can be considered to be better?

First, let us formalize what do we mean by a line $\hat{y}$ adequately describing $y$ as a function of $x$. Intuitively, we consider a line $\hat{y}$ to be adequate if the distances from it to the points are small. The lines in the top panels of Figure 35 have large distances to the points, that is why we immediately consider them to be inadequate. But the lines in the bottom panels have small distances to the points. Which one should we prefer?

In order to answer this question we need to define a criterion for measuring the distance from a line $\hat{y} = b_0 + b_1 x$ to the observed points. We will consider the *sum of squares error* —SSE—:

$$SSE = \sum_U e_i^2 \qquad \text{where} \qquad e_i = y_i - \hat{y}_i \qquad \text{and} \qquad \hat{y}_i = b_0 + b_1\, x_i.$$

The reason for the name *sum of squares error* can be interpreted as follows. The value $\hat{y}_i$ is the approximation to $y_i$ made by the straight line. Therefore the difference $y_i - \hat{y}_i$ can be interpreted as the *error* $e_i$ made by the straight line when approximating the $i$th observation. Our criterion is simply the sum of the squares of these errors.

**Example 50.** Let us calculate the sum of squares error —SSE— for each of the four lines $\hat{y}_1$, $\hat{y}_2$, $\hat{y}_3$ and $\hat{y}_4$ fitted to the workers dataset in Figure 35.

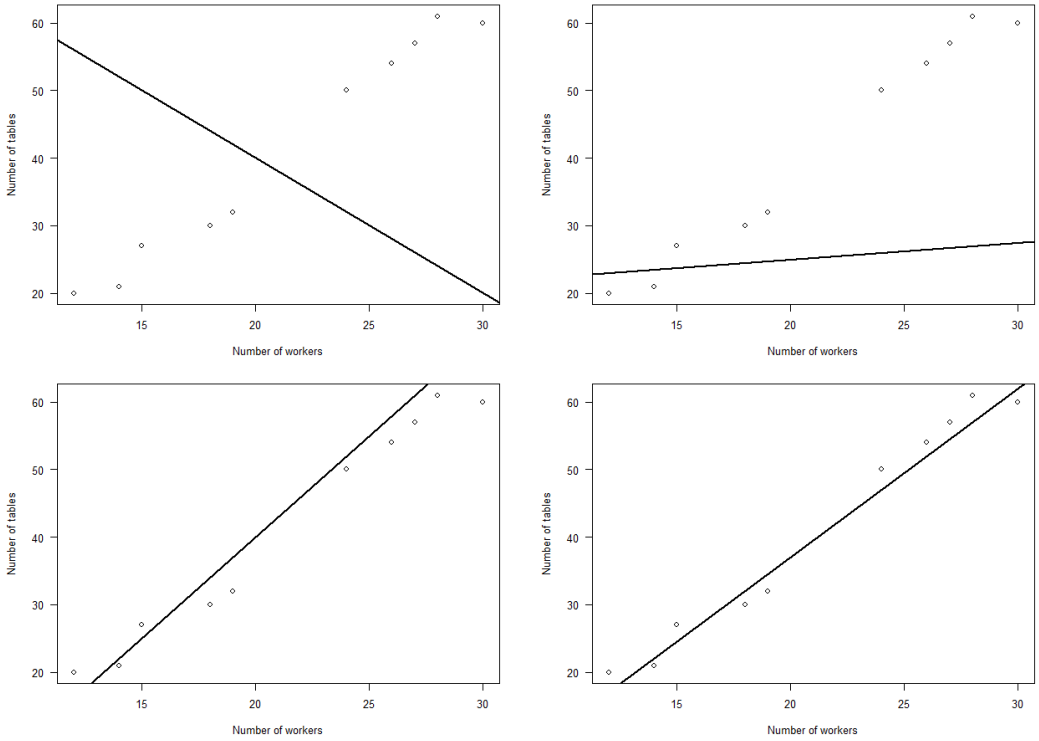


Figure 35: Four lines fitted to the workers dataset.

For $\hat{y}_1$ we have the intercept $b_0 = 80$ and the slope $b_1 = -2$, which yields fitted values $\hat{y}_1$ and errors $e_i = y_i - \hat{y}_{1,i}$ as shown in the fourth and fifth columns of Table 26, respectively. Therefore we have

$$SSE_1 = \sum_U e_i^2 = (-36)^2 + (-31)^2 + \cdots + 40^2 = 8012.$$

| $i$ | $x_i$ | $y_i$ | $\hat{y}_{1,i}$ | $e_i$ |
|---|---|---|---|---|
| 1 | 12 | 20 | 56 | -36 |
| 2 | 14 | 21 | 52 | -31 |
| 3 | 15 | 27 | 50 | -23 |
| 4 | 18 | 30 | 44 | -14 |
| 5 | 19 | 32 | 42 | -10 |
| 6 | 24 | 50 | 32 | 18 |
| 7 | 26 | 54 | 28 | 26 |
| 8 | 27 | 57 | 26 | 31 |
| 9 | 28 | 61 | 24 | 37 |
| 10 | 30 | 60 | 20 | 40 |

Table 26: Number of tables $y$ produced by $x$ workers

The sum of squares error for the remaining lines $\hat{y}_2$, $\hat{y}_3$ and $\hat{y}_4$ are found in a analogous way, we get $SSE_2 = 10046$, $SSE_3 = 207$ and $SSE_4 = 66$. Therefore, according to the SSE criterion, among these four lines, $\hat{y}_4 = -13 + 2.5\,x$ is the one that better fits the observations. □

In Example 50 we used SSE for deciding which line, among four different options, fits the observations best. The natural next step would be to find the best line according to the SSE

criterion, in other words, finding the values for the intercept $b_0$ and the slope $b_1$ that minimize the SSE. The solution is known as the *least squares regression.*

**Definition 51.** Let $(x_1\,, y_1)$, $(x_2\,, y_2)$, $\cdots$, $(x_N\,, y_N)$ be the observations of two variables $x$ and $y$ on a set $U$ of $n$ elements. The line that minimizes the SSE is given by the least squares regression,

$$\hat{y} = b_0 + b_1\, x \qquad \text{with} \qquad b_1 = r_{xy,U}\frac{S_{y,U}}{S_{x,U}} \qquad \text{and} \qquad b_0 = \bar{y}_U - b_1\bar{x}_U. \tag{14}$$

where $r_{xy,s}$ is the correlation coefficient between $x$ and $y$ in $s$, $S_{x,s}$ and $S_{y,s}$ are the standard deviations of $x$ and $y$, respectively. $\square$

**Remark:** The slope $b_1$ in (14) can be written alternatively as

$$b_1 = \frac{\sum_U (x_i - \bar{x}_U)(y_i - \bar{y}_U)}{\sum_U (x_i - \bar{x}_U)^2} \qquad \text{or} \qquad b_1 = \frac{\sum_U x_i y_i - N\bar{x}_U\bar{y}_U}{\sum_U x_i^2 - N\bar{x}_U^2}.$$

The intercept $b_0$ and, in particular, the slope $b_1$ in a least squares regression have interesting interpretations. The slope $b_1$ indicates that, on average, a unit change in the independent variable $x$ is associated to a change of $b_1$ in the dependent variable $y$. It is worth to emphasize the need for the expression "on average" in the previous sentence: $x$ and $y$ do not follow exactly a straight line, therefore the change is not deterministic, that is why we need to emphasize that the slope measures the "average change". The intercept, on the other hand, indicates the value of the dependent variable $y$ that is expected when the independent variable $x$ is equal to zero.

**Example 52.** Let us find the least squares regression for the workers dataset from Examples 49 and 50. We have $\bar{x}_U = 21.3$, $\bar{y}_U = 41.2$, $S_{x,U} = 6.482$ and $S_{y,U} = 16.69$ and $r_{xy,U} = 0.9888$. Therefore

$$b_1 = r_{xy,U}\frac{S_{y,U}}{S_{x,U}} = 0.9888\frac{16.69}{6.482} = 2.545 \qquad \text{and} \qquad b_0 = \bar{y}_U - b_1\bar{x}_U = 41.2 - 2.545\cdot 21.3 = -13.02.$$

The fitted regression line is

$$\hat{y} = -13.02 + 2.545\, x,$$

which is shown in Figure 36.

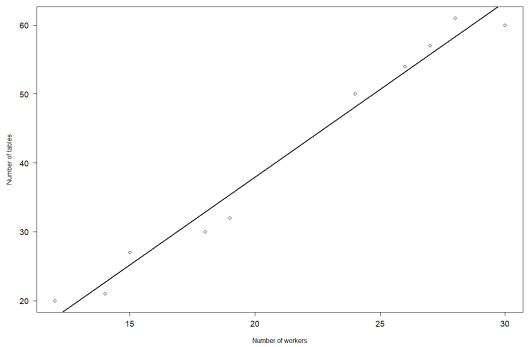


Figure 36: Least squares regression fitted to the workers dataset.

It can be verified that the sum of squares error for the least squares regression is $SSE = 56$, which is, in fact, smaller than the SSE of any of the four lines fitted in Example 50.

The intercept $b_0$ is interpreted as: a company with no workers is expected to produce around -13 tables. Admittedly, in this case, this interpretation makes no sense, it is simply not possible to produce a negative number of tables. This undesired result is due to the fact that the value $x = 0$ is not one of the observed values and it is, in fact, quite far from any of the observed values. One must be careful when extrapolating the results of a regression to observations that do not belong to the dataset, especially if they are quite different to the observed values.

The slope $b_1$ is interpreted as: in our set of 10 companies, on average, one extra worker is associated to 2.5 more tables produced. $\square$