Linear Regression

STAT 1F92 ❄ Component 02 ❄ September 20 – October 03, 2023[1]

Table of Content

❄ **September 21, 2023** ❄

2 Linear Regression

We will learn about scatter plots and correlation, the least east-squares regression method and how to perform diagnostics on the least-squares regression line.

Definition 2.1: Scatter plot

A *scatter* plot is a graph that shows the relationship between two quantitative variables measured on the same individual that is marked as a point on the graph. The points are *not* connected by a line.

Definition 2.2: Response and explanatory variables

The value of a *response* variable can be explained by the value of the *explanatory* (or predictor) variable. In other words, an explanatory (cause) will influence the response (effect) variable.

Here a some examples to further explain response and explanatory variables:

- Age of a person influences their height
  - Suppose that the older the person is, the taller they are
  - The explanatory variable is ‘age’ and response variable is ‘height’
- Time spent studying influences the grade earned
  - Suppose that the longer the person studies, the higher the grade earned
  - The explanatory variable is ‘study hours’ and response variable is ‘grade’
- Years of experience influences salary
  - Suppose that the more experience a person has, the higher their salary
  - The explanatory variable is ‘years of experience’ and response variable is ‘salary’

[1] Complied at: Thursday March 28th, 2024 12:55pm -04:00

On a scatter plot, the explanatory variable is plotted on the x-axis (horizontal) and the response variable on the y-axis (vertical)[2]. A generic scatter plot can be seen in Figure 2.1.

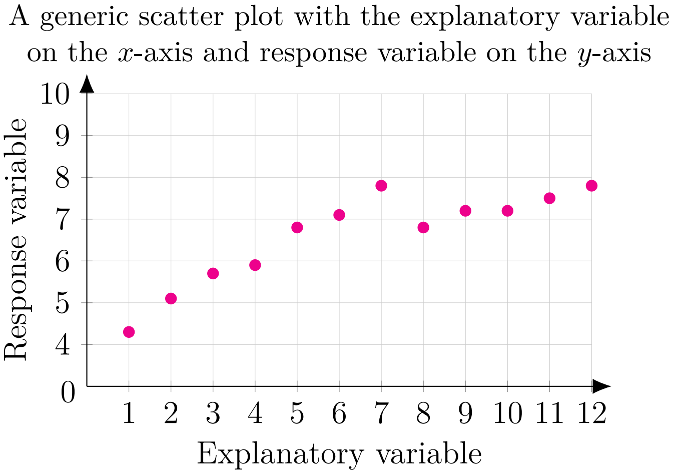


Figure 2.1: A generic scatter plot

Example 2.1: Generating a scatter plot

Suppose we are given the following data regarding the number of hours studied and the grade received:

Hours studied	Grade (%)
1	32
2	39
3	42
4	48
5	50
6	65
7	78
8	84
9	90
10	86
11	82
12	78

We see that the number of hours of studying influences the grade. Less hours means low grade and long study time doesn't necessarily yield optimal results. For instance, we are better off studying seven hours instead of twelve. Hence, let the explanatory variable be the number of hours studied and the response variable be the grade achieved. The scatter plot that represents our data is:

[2] To remember which one is which, the word explanatory has 'x' in it and is plotted on the x-axis.

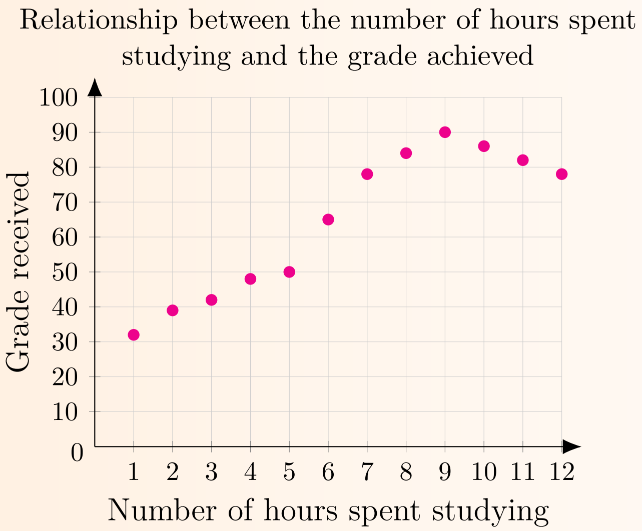


Figure 2.2: The scatter plot of the relationship between the number of hours spent studying and the grade achieved

There are several scenarios we could encounter when graphing the scatter plot. We are interested in two cases: positively associated and negatively associated variables.

2.1 Linear Correlation Coefficient (r)

Definition 2.3: positively associated and negatively associated variables

- **Positively associated:** when above-average values of one variable are associated with above-average values of the other variable, and below-average values of one variable are associated with below-average values (*i.e.*, when one variable goes up, the other variable goes up as well)
- **Negatively associated:** when below-average values of one variable are associated with below-average values of the other variable, and below-average values of one variable are associated with below-average values (*i.e.*, when one variable goes down, the other variable goes down as well)

The linear correlation coefficient is a measure of the strength and direction of the linear relation between two quantitative variables. Some scatter plots have been given graphed in Figure 2.3. We need to mathematically decide whether something has a positive correlation or not. To do that, we will use the Sample Linear Correlation Coefficient formula[3].

Definition 2.4: Sample Linear Correlation Coefficient formula

The linear correlation coefficient formula, denoted by r, measures the strength and

[3] Only the sample formula will be used in this course, the population formula is not a part of this course

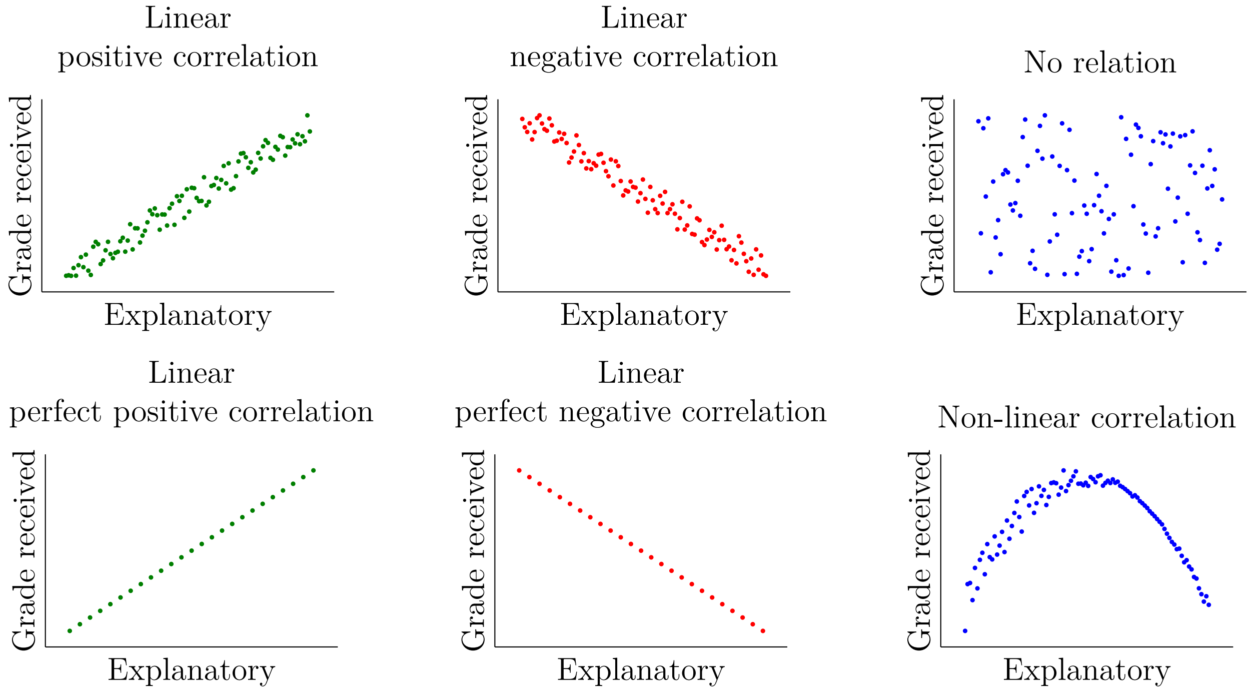
direction of the linear relation between two quantitative:

$$r = \frac{\displaystyle\sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x}\right) \cdot \left(\frac{y_i - \overline{y}}{s_y}\right)}{n-1}, \tag{1}$$

where x_i the ith data value of the explanatory variable (horizontal)
$\overline{x}$ sample mean of the explanatory variable (horizontal)
s_x sample standard deviation of the explanatory variable (horizontal)
y_i is the ith observation of the response variable (vertical)
$\overline{y}$ sample mean of the response variable (vertical)
s_y sample standard deviation of the response variable (vertical)
n the number of individuals in the sample

Some properties of the Linear Correlation Coefficient r:

- The value r is between -1 and 1 (both inclusive, *i.e.*, $[-1, 1]$) and is unitless
- $r = +1$: correlation between the two variables is perfectly *positive* linear (going up ↗)
- $r = -1$: correlation between the two variables is perfectly *negative* linear (going down ↘)
- When r is close to $+1$ (*e.g.*, 0.8), there is a strong evidence of positive association
- When r is close to -1 (*e.g.*, -0.9), there is a strong evidence of negative association
- r is close to 0, then there is little or no evidence exists of a linear relation between the two variables. There *might* be a non-linear relationship between them but not a linear one.


Linear positive correlation

Linear negative correlation

No relation

Linear perfect positive correlation

Linear perfect negative correlation

Non-linear correlation

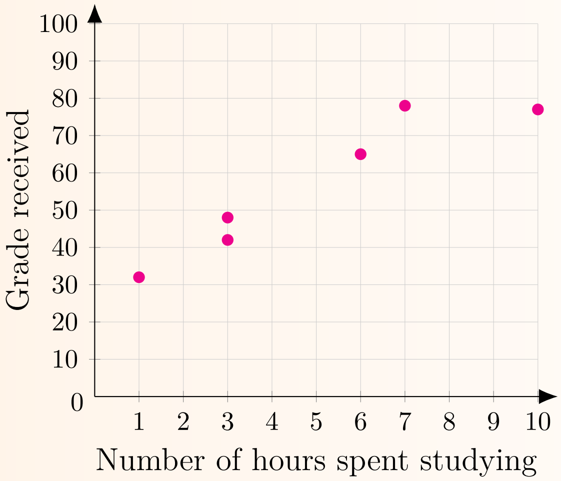
Figure 2.3: Different types of graphs with different correlations

Example 2.2: Linear Correlation Coefficient

Find the linear correlation coefficient of the following data:

x Hours studied	y Grade
1	32
3	42
3	48
6	65
7	78
10	77

Scatter plot of the table on the left showing the relationship between the number of hours spent studying and the grade achieved



First, let us find both means $\overline{x}$ and $\overline{y}$, and both standard deviations s_x and s_y.

$$\begin{aligned} \overline{x} &= \frac{1+3+3+6+7+10}{6} = 5 \\ \overline{y} &= \frac{32+42+48+65+78+77}{6} = 57 \end{aligned}$$

$$s_x = \sqrt{\frac{(1-5)^2+(3-5)^2+(3-5)^2+(6-5)^2+(7-5)^2+(10-5)^2}{6-1}}$$

$$s_x = 3.28634$$

$$s_y = \sqrt{\frac{(32-57)^2+(42-57)^2+(48-57)^2+(65-57)^2+(78-57)^2+(77-57)^2}{6-1}}$$

$$s_y = 19.16246 \tag{2}$$

Hence, $\overline{x} = 5$, $\overline{y} = 57$, $s_x = 3.28634$ and $s_y = 19.16246$. Now we have all of the numbers required to use the linear correlation coefficient formula and find r. It is best to places the data in a tabular form:

x	y	$\frac{x_i - \overline{x}}{s_x}$	$\frac{y_i - \overline{y}}{s_y}$	$\left(\frac{x_i - \overline{x}}{s_x}\right) \cdot \left(\frac{y_i - \overline{y}}{s_y}\right)$
1	32	$\frac{1 - 5}{3.28634} = -1.21716$	$\frac{32 - 57}{19.16246} = -1.30463$	$(-1.21716) \cdot (-1.30463) = 1.58795$
3	42	$\frac{3 - 5}{3.28634} = -0.60858$	$\frac{42 - 57}{19.16246} = -0.78278$	$(-0.60858) \cdot (-0.78278) = 0.47638$
3	48	$\frac{3 - 5}{3.28634} = -0.60858$	$\frac{48 - 57}{19.16246} = -0.46967$	$(-0.60858) \cdot (-0.46967) = 0.28583$
6	65	$\frac{6 - 5}{3.28634} = 0.30429$	$\frac{65 - 57}{19.16246} = 0.41748$	$(0.30429) \cdot (0.41748) = 0.12704$
7	78	$\frac{7 - 5}{3.28634} = 0.60858$	$\frac{78 - 57}{19.16246} = 1.09589$	$(0.60858) \cdot (1.09589) = 0.66694$
10	77	$\frac{10 - 5}{3.28634} = 1.52145$	$\frac{77 - 57}{19.16246} = 1.04371$	$(1.52145) \cdot (1.04371) = 1.58795$

The far-right column contains the numbers we are interested in. We need to sum them up and then divide by $n - 1$:

$$r = \frac{1.58795 + 0.47638 + 0.28583 + 0.12704 + 0.66694 + 1.58795}{6 - 1} = 0.94642$$

To test for a linear relation, we need to find the absolute value of the correlation coefficient and then the critical value in Table II from Appendix of the textbook, which is also included in Table 2.1 in this document. In case the absolute value is greater than the critical value, we conclude that there exists a linear relation between the two quantitative variables. Otherwise, we conclude that a linear relation doesn't exists. In our case, the critical value in the table is 0.811. Since our correlation coefficient value r is 0.94642 , we conclude that there is a linear relation between the number of hours spent studied and the grade received. So, in this example, any r value that is equal to 0.811 or larger means that there is a linear relationship.

n	Critical Value
3	0.997
4	0.950
5	0.878
6	0.811
7	0.754
8	0.707
9	0.666
10	0.632
11	0.602
12	0.576
13	0.553
14	0.532
15	0.514
16	0.497
17	0.482
18	0.468
19	0.456
20	0.444
21	0.433
22	0.423
23	0.413
24	0.404
25	0.396
26	0.388
27	0.381
28	0.374
29	0.367
30	0.361

Table 2.1: Critical Values for Correlation Coefficient

2.2 Least-Squares Regression

Suppose we have a scatter plot and want to find the line of best fit, how can we do that? The line of best fit is a line that 'cuts through' the data points and is the best approximation of the data values. This is referred to as the 'regression line'. Suppose we have the same data as Example 2.2. It 'should' predict where the future data points will lie on the scatter plot. One approach is to select two point from the data set and use them to find the equation of the line. For now, we can choose the points by eyeballing them. Let us choose the $(1, 32)$ and $(10, 77)$. Let us find the equation of the line.

2.2.1 Finding the Equation of a Line

A *slope* is defined as rise over run: the rise (vertical) of a given line segment is the change in its y value, and its run (horizontal) is the change in its x-value. We can represent any line using the following formula given two points (x_1, y_1) and (x_2, y_2):

$$y = m \cdot x + b,$$

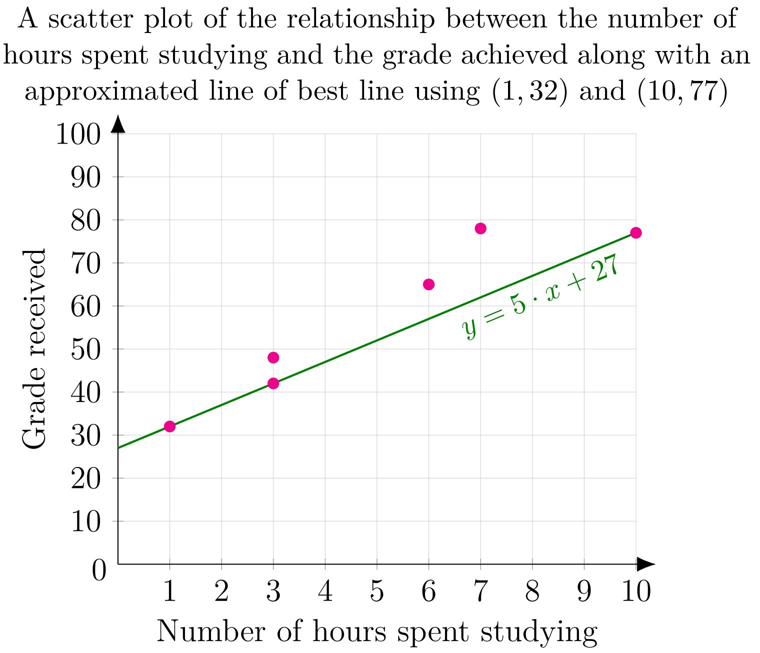
- The *rise* is $y_2 - y_1$
- The *run* is $x_2 - x_1$
- The *slope* of a line is constant across the entire line, and is $m = \frac{rise}{run}$
- b is the y-intercept (*i.e.*, when $x = 0$, the vertical line)

To understand the idea of slope, slope is defined as $\frac{\text{rise}}{\text{run}}$ or $\frac{\text{change in } y}{\text{change in } x}$. For instance, if we have the line equation as $y = \frac{3}{4} \cdot m + 5$, then this means every time we go up three units, we have to move four units to the right. Another example if we have $y = 2 \cdot m + 8$, which can be seen as $y = \frac{2}{1} \cdot m + 5$ then this means every time we go up two units, we move one unit to the right. Lastly, if we have $y = (-2) \cdot m + 3$, then this is the same as $y = \frac{-2}{1} \cdot m + 3$, which has a negative slope. Hence, if we go down twice, then we move right once, which is the same as $y = \frac{2}{-1} \cdot m + 3$, when we move two units up, we move one unit to the left (do the check yourself by drawing a line that has multiple points on it and). Let us find the equation of the line given the two points:

$$(x_1, y_1) = (1, 32),$$
$$(x_2, y_2) = (10, 77).$$

Our slope is:

$$m = \frac{y_2 - y_1}{x_2 - x_1} = \frac{77 - 32}{10 - 1} = \frac{45}{9} = 5 \tag{3}$$



A scatter plot of the relationship between the number of hours spent studying and the grade achieved along with an approximated line of best line using $(1, 32)$ and $(10, 77)$

Figure 2.4: The graph $y = 5 \cdot x + 27$ to scale. Finding the slope between any two points on the blue line will give the same m value, 5.

The y-intercept, denoted by b, is when the line intersects the y-axis, *i.e.*, when $x = 0$. We can choose any of the two points $(1, 32)$ or $(10, 77)$ (or other points in green) to find the b value. Note that the y-intercept will be denoted by the coordinate $(0, b)$. We can let $(x_1, y_1) = (1, 32)$ and $(x_2, y_2) = (0, b)$:

$$\begin{aligned}
m &= \frac{y_2 - y_1}{x_2 - x_1} \\
5 &= \frac{b - 32}{0 - 1} \\
5 &= \frac{b - 32}{-1} \\
5 &= (-1) \cdot (b - 32) \\
5 &= -b + 32 \\
5 - 32 &= -b \\
-27 &= -b \\
27 &= b \\
b &= 27
\end{aligned} \tag{4}$$

Hence, $m = 5$ and $b = 27$, the equation of the line is:

$$y = 5 \cdot x + 27.$$

We will use the notation $\hat{y} = 5 \cdot x + 27$ to denote that this line is the *predicted* line of best fit.

Definition 2.5: Residual

The difference between the observed (data given in the table) and predicted values (found using the line equation) of $\hat{y}$ (*i.e.*, the error between the data points and our approximation)

Here are all of the residuals and the residuals squared of Example 2.2:

x	y	Approximated $\hat{y}$ $\hat{y} = 5 \cdot x + 27$	Residuals $(y - \hat{y})$	Residual squared $(y - \hat{y})^2$
1	32	$\hat{y} = 5 \cdot (1) + 27 = 32$	$32 - 32 = 0$	$0^2 = 0$
3	42	$\hat{y} = 5 \cdot (3) + 27 = 42$	$42 - 42 = 0$	$0^2 = 0$
3	48	$\hat{y} = 5 \cdot (3) + 27 = 42$	$48 - 42 = 6$	$6^2 = 36$
6	65	$\hat{y} = 5 \cdot (6) + 27 = 57$	$65 - 57 = 8$	$8^2 = 64$
7	78	$\hat{y} = 5 \cdot (7) + 27 = 62$	$78 - 62 = 16$	$16^2 = 256$
10	77	$\hat{y} = 5 \cdot (10) + 27 = 77$	$77 - 77 = 0$	$0^2 = 0$
				↑ Sum of residuals squared = 356

Note that negative residual means the data point lies *under* the line, which we don't have. The notation of adding a hat on top of a variable means it is predicted/approximated. Our regression line gave us a sum of residuals squared as 356. Can we come up with another line that gives us a sum of residuals squared that is less than 356? Remember that the smaller the sum of the residuals squared, the better the line of best fit (a sum of residuals squared of zero means all of our data points lie on the line perfectly). Recall that we chose the two points $(1, 32)$ and $(10, 77)$ to find the equation of the line, which was $\hat{y} = 5 \cdot x + 27$. Let us perform the same thing but now choosing different points to find different lines and test their approximation by calculating the residuals squared. We could also choose some arbitrary line, it is not a must that we choose from our data points, but preferred to choose from the data points. Let us use the same logic in Equation 3 and Equation 4.

Definition 2.6: Least-Squares Regression Criterion

The least-squares regression line is the line that minimizes the sum of the squared residuals. This line minimizes the sum of the squared vertical distance between the observed values of y and those predicted by the line, $\hat{y}$ (read 'y-hat'). It represents an approach to 'minimize Σ residuals^2'.

The objective is to ignore the concept of positive and negative errors as squaring a positive or a negative value yields a positive value, and is to minimize the square of the residuals. This is important because squaring values will change the data significantly. Think about having a residuals of, say, 5, 7 and 16. Their sum is 28 but their squared sum is $5^2 + 7^2 + 16^2 = 330$. Hence, the deviation or error is a lot larger when the residuals are squared. This approach minimizes the residuals/errors that will be encountered. The idea of 'minimizing Σ residuals^2'

is minimizing the same logic used in finding the numerator of variance but using the response variable (*i.e.*, $(y_1 - \hat{y_1})^2 + (y_2 - \hat{y_2})^2 + \cdots + (y_n - \hat{y_n})^2)$.

Definition 2.7: The Least-Squares Regression Line

The equation of the least-squares regression line is given by

$$\hat{y} = b_1 \cdot x + b_0,$$

where

$b_1 = r \cdot \dfrac{s_y}{s_x}$	*slope* of the least-squares regression line
$b_0 = \overline{y} - (b_1 \cdot \overline{x})$	*y-intercept* of the least-squares regression line
$\overline{x}$	mean of the explanatory variable
s_x	standard deviation of the explanatory variable
$\overline{y}$	mean of the response variable
s_y	standard deviation of the response variable

So, the line defined by $\hat{y} = b_1 \cdot x + b_0$ will always pass through the point $(\overline{x}, \overline{y})$ and is created to predict the value of y given x.

Example 2.3: Quick least-squares regression line

Use the data given to find the least-squares regression line:

$$r = -0.683, \qquad \overline{x} = 6, \qquad s_x = 4.7524, \qquad \overline{y} = 13, \qquad s_y = 7.8858$$

We need to find b_1 and b_0:

$$\begin{aligned} b_1 &= r \cdot \frac{s_y}{s_x} && = (-0.683) \cdot \frac{7.8858}{4.7524} && = (-0.683) \cdot (1.65933) && = -1.13332 \\ b_0 &= \overline{y} - b_1 \cdot \overline{x} && = (13) - (-1.13332) \cdot (6) && = 19.79993 \end{aligned}$$

Hence, the equation of the least-squares regression line is

$$\hat{y} = (-1.13332) \cdot x + (19.79993).$$

September 28, 2023

Example 2.4: Finding the least-squares regression line of Example 2.2

We need to find: r, $\overline{x}$, s_x, $\overline{y}$, s_y, $b_1 = r \cdot \dfrac{s_y}{s_x}$ and $b_0 = \overline{y} - b_1 \cdot \overline{x}$, to find $\hat{y} = b_1 \cdot x + b_0$.

We have:

$$r = 0.94642, \quad \overline{x} = 5, \quad s_x = 3.28634, \quad \overline{y} = 57, \quad s_y = 19.16246$$

Let us find

$$\begin{aligned} b_1 &= r \cdot \frac{s_y}{s_x} &&= (0.94642) \cdot \frac{19}{3.28634} &&= 5.47173 \\ b_0 &= \overline{y} - b_1 \cdot \overline{x} &&= 57 - (5.47173) \cdot 5 &&= 29.64133 \end{aligned}$$

Hence, the least-squares regression line is given by the equation:

$$\hat{y} = 5.47173 \cdot x + 29.64133$$

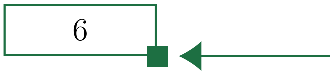
2.2.2 Excel: Absolute Referencing Versus Relative Referencing

We could use Excel to find both the r value and least square regression. Since we are using the online version, we have to manually[4] calculate the value of r. Then, we are able to easily find the least-squares regression line formula. We have seen how to find the the mean and standard deviation via excel. Now, given a set of values, how can we apply a single cell to all of the other values. For instance, to find the z-score of the x values, we need get the current x value, subtract it from $\overline{x}$ and then divide by the s_x. We have to do this for all of the values. Excel is designed to 'follow the pattern' of the formulas we write. For example, if we write `=A2+B2` in `C2`, then it will add the values of `A2` and `B2` and places the result in `C2`.

	A	B	C	D	E
1	**X**	**Y**	**Sum**		
2	5	1	6		
3	4	7			
4	9	6			
5	4	1			
6	3	3			
7	7	8			
8					

Table 2.2: Cell `C2` is storing the sum of cells `A2` and `B2` using the formula `=A2+B2`

When using Excel on a laptop/computer (*i.e.*, not a tablet[5]), you are able to drag down from the bottom-right corner. Ensure that cell `C2` is clicked and drag down on this:



We should get something like this:

[4] This will be the case for assignment questions that might ask for finding the r value or least-squares regression line formula. Everything needs to be found manually.

[5] When using a tablet, select the cells that need to be filled and the option 'fill' should appear, click on it

	A	B	C	D	E
1	**X**	**Y**	**Sum**		
2	5	1	6		
3	4	7	11		
4	9	6	15		
5	4	1	5		
6	3	3	6		
7	7	8	15		
8					

Excel is smart enough to realize we added the sum of the first row and we want to perform the same logic for the other rows. This is referred to as *relative referencing*. Suppose that we want to add the cell `B2` with all of the X values (*i.e.*, increment the X values by one and store the result in column `C`). We have to refer to the cell B2 each time. We could manually write the formula but we shouldn't. Instead, we can tell Excel that even if we drag down, refer to `B2` at all times. This is referred to as *absolute referencing*. To do that, we will place the following formula in cell `C2`: `=A2+$B$2`. We added a dollar sign `$` twice. Now, if we drag down once, then `A2` becomes `A3` and `$B$2` stays the same. Dragging down yet another row will make `A3` become `A4` but `$B$2` stays the same. The pattern follows through. The idea is that we will always refer to that one cell all the time.

2.2.3 The Significance of Least-Squares Regression Line

Let us understand the significance of the least-squares regression line by using the data from Example 2.2. Recall that we chose the two points $(1, 32)$ and $(10, 77)$ to find the equation of the line, which was $\hat{y} = 5 \cdot x + 27$, that did fit the data points somewhat well. Now, let us use different points to generate another line to approximate the data points. Suppose that we choose $(3, 42)$ and $(7, 78)$. Using the logic found in Equation 3 and Equation 4, it happens that the line will be $\hat{y} = 9 \cdot x + 15$. Now, we will choose an arbitrary line to fit our data. We can choose *any* line to fit the data, it is not a must to choose two points from the data points, but choosing from the data points is preferred when doing calculations manually. Say we arbitrarily choose the line $\hat{y} = 18 \cdot x - 8$, then we have the following three lines and a fourth one from Example 2.4:

$$\begin{aligned} \text{Line 1:} &\quad \hat{y} = 5 \cdot x + 27 \\ \text{Line 2:} &\quad \hat{y} = 9 \cdot x + 15 \\ \text{Line 3:} &\quad \hat{y} = 18 \cdot x - 8 \\ \text{Line 4:} &\quad \hat{y} = 5.47173 \cdot x + 29.64133 \end{aligned} . \tag{5}$$

They are graphed in Figure 2.5.

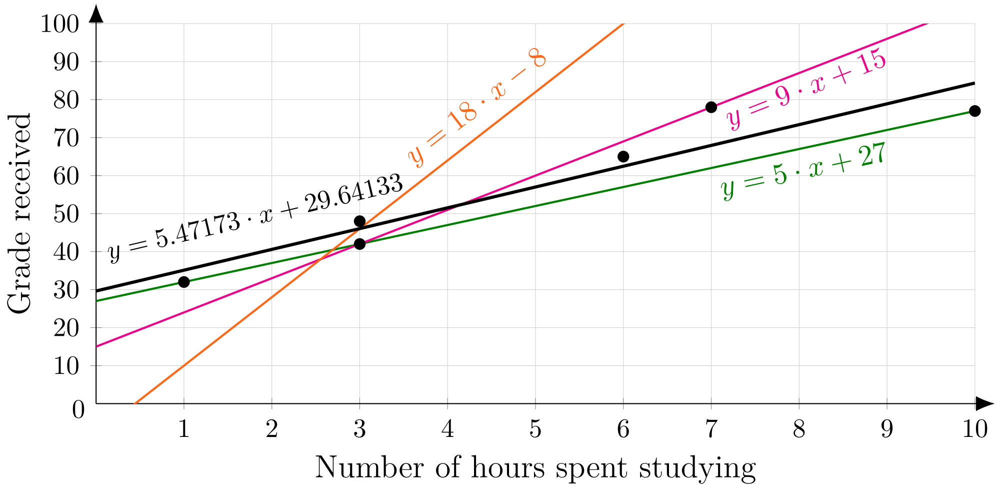

Scatter plot of the table on the left showing the relationship between the number of hours spent studying and the grade achieved

Figure 2.5: The graph that shows the lines we chose to fit the data. The thick line in black is the line $y = 5.47173 \cdot x + 29.64133$, which is the least-squares regression line (*i.e.*, as optimal as it can be)

We can have a look at Figure 2.5 to see which one has the least amount of the sum of the residuals squared.

x	y	Line 1 $\hat{y} = 5 \cdot x + 27$	Line 1 $\hat{y} = 5 \cdot x + 27$ Residuals Squared $(y - \hat{y})^2$	Line 2 $\hat{y} = 9 \cdot x + 15$	Line 2 $\hat{y} = 9 \cdot x + 15$ Residuals Squared $(y - \hat{y})^2$	Line 3 $\hat{y} = 18 \cdot x - 8$ Residuals Squared $(y - \hat{y})^2$
1	32	$5 \cdot (1) + 27 = 32$	$(32 - 32)^2 = 0$	$9 \cdot (1) + 15 = 24$	$(32 - 24)^2 = 64$	$(32 - 10)^2 = 484$
3	42	$5 \cdot (3) + 27 = 42$	$(42 - 42)^2 = 0$	$9 \cdot (3) + 15 = 42$	$(42 - 42)^2 = 0$	$(42 - 46)^2 = 16$
3	48	$5 \cdot (3) + 27 = 42$	$(48 - 42)^2 = 36$	$9 \cdot (3) + 15 = 42$	$(48 - 42)^2 = 36$	$(48 - 46)^2 = 4$
6	65	$5 \cdot (6) + 27 = 57$	$(65 - 57)^2 = 64$	$9 \cdot (6) + 15 = 69$	$(65 - 69)^2 = 16$	$(65 - 100)^2 = 1225$
7	78	$5 \cdot (7) + 27 = 62$	$(78 - 62)^2 = 256$	$9 \cdot (7) + 15 = 78$	$(78 - 78)^2 = 0$	$(78 - 118)^2 = 1600$
10	77	$5 \cdot (10) + 27 = 77$	$(77 - 77)^2 = 0$	$9 \cdot (10) + 15 = 105$	$(77 - 105)^2 = 784$	$(77 - 172)^2 = 9025$
			↑ Sum = 356		↑ Sum = 900	↑ Sum = 12354

Table 2.3: The least-squares residuals of the three lines: line 1, line 2 and line 3.

The summary is that:

- line 1 has a sum of residuals squared of 356
- line 2 has a sum of residuals squared of 900
- line 3 has a sum of residuals squared of 12354
- The black line ($y = 5.47173 \cdot x + 29.64133$), is the least-squares regression line, which has the sum of the residuals squared of 191.5997.

Out of the coloured lines, we see that line 1 has the least of deviation (the less the deviation, the more accurate the line of best fit). The question now is what is the best line that will minimize the sum of the residuals squared? The answer, simply, is the least-squares regression line! That is $y = 5.47173 \cdot x + 29.64133$, the line in black in Figure 2.5.

❋ October 03, 2023 ❋

2.2.4 The Coefficient of Determination R^2

From the previous section, we have found a line equation that minimizes the sum of the residuals squared. How good is that line? Is there a way for us to conclude whether the line fits the data perfectly or 'almost perfectly' or 'somewhat good'? How can we mathematically formulate our conclusion?

Definition 2.8: Coefficient of Determination R^2

The coefficient of determination, R^2, measures the proportion of total variation in the response variable that is explained by the least-squares regression line. It measures how good the approximation of the least-squares regression line in terms of the relation between the explanatory and response variables.

The value of R^2 will be between 0 and 1 (*i.e.*, $[0, 1]$). The closer the value of R^2 to 1 the better the approximation of how the explanatory variable affects the response variable. Having R^2 to be closer to 0 means that the line equation doesn't describe how the explanatory variable affects the response variable.

There are two different types of deviations (*i.e.*, errors) we might encounter: *explained deviation* and *unexplained deviation*. In order to specify them, we need to use our least-square regression line with the means of the response variable $\overline{y}$ and the explanatory variable $\overline{x}$ and the value of data point to create two separate regions:

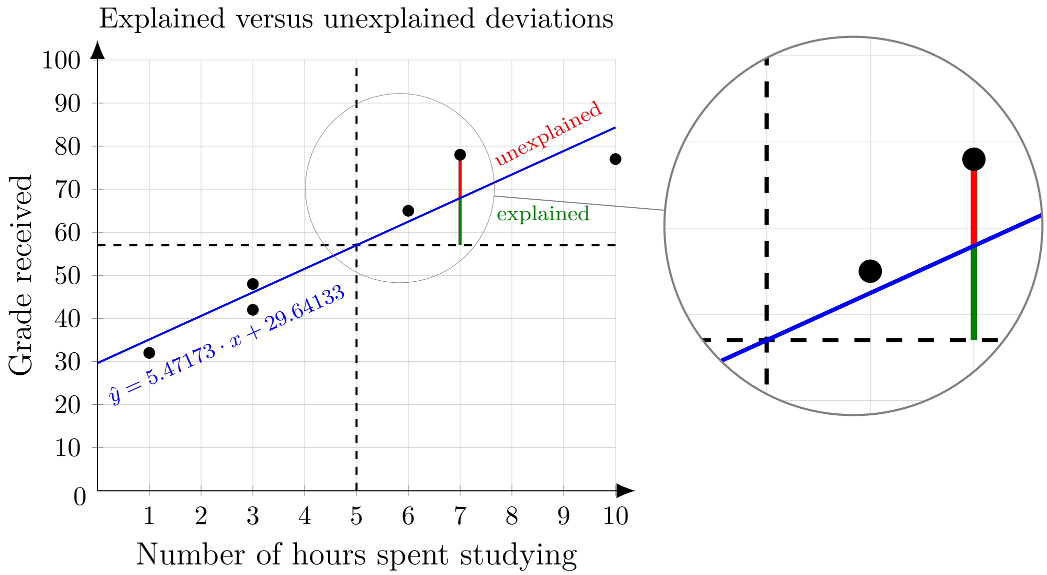


Figure 2.6: The blue line is $\hat{y} = 5.47173 \cdot x + 29.64133$ and black is the $\overline{y} = 57$ (horizontal) and $\overline{x} = 5$ (vertical) with the explained and unexplained deviations of a single data point shown

We can find the variation, which is how much we are away from the mean, using the following formula:

$$\begin{aligned} \text{deviation} &= \text{unexplained deviation} + \text{explained deviation} \\ y - \overline{y} &= (y - \hat{y}) + (\hat{y} - \overline{y}) \end{aligned} \tag{6}$$

The deviation squared is given as:

$$\begin{aligned} (\text{deviation})^2 &= (\text{unexplained deviation})^2 + (\text{explained deviation})^2 \\ (y - \overline{y})^2 &= (y - \hat{y})^2 + (\hat{y} - \overline{y})^2 \end{aligned} \tag{7}$$

We can find the *total* variation using the following formula:

$$\begin{aligned} \text{Total deviation} &= \text{unexplained deviation} + \text{explained deviation} \\ \sum_{i=1}^{n}(y_i - \overline{y})^2 &= \sum_{i=1}^{n}(y_i - \hat{y})^2 + \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 \end{aligned} \tag{8}$$

The sum of residuals squared is the unexplained variation. To find the value R^2, just square the value of r, which is found in Definition 2.4. Another way to look at it is by the follow formula:

$$\begin{aligned} R^2 &= \frac{\text{explained deviation}}{\text{total variance}}, \quad \text{which is the same as} \\ R^2 &= 1 - \frac{\text{unexplained deviation}}{\text{total variance}} \end{aligned} \tag{9}$$

which is some value between 0 and 1.

2.2.5 Linear Correlation Coefficient (r) VS Coefficient of Determination (R^2)

We use r to find the linear correlation coefficient (between ±1), which measures how much two variables are related. That is, if the x value is above $\overline{x}$, then the y value should be above $\overline{y}$, and if the x value is below $\overline{x}$ then the y value should be below $\overline{y}$. It will focuses on *that* relationship along with the trend of the points. In case all points lie on a single line, then there is a perfect association between the explanatory and response variables. In case the points are close to each other and can be appropriated by a line nicely, then this has a strong association (Figure 2.3 shows this graphically).

We use $R^2 = r^2$ to find the coefficient of determination (between 0 and 1), which is how much the line can explain the variation in y. Informally, r gives a overview of the correlation (or relationship) between the variables and R^2 gives how much the line explain the deviations/errors in the model. Suppose there is a single outlier in the data point that lies horizontally in the middle of the data points. By definition, it will be significantly above/-below the mean $\overline{y}$ (horizontal line). In other words, it could be way up or way down on the graph. The approaches that are used to fit the data using a line will not take it into consideration, because if it did, then all of the other data points will not fit the line well. It is better the focus on the majority of the points and leave one or two than vice-versa. Hence, our line will not cover it. When finding the deviation from the line of regression, we realize that the deviation is *very* large. Does the line of regression isn't a good line? We have to differentiate between two things: the distance from the mean to the line and from the line to the point. The line is responsible for the deviation from the mean to itself (the explained deviation). The unexplained deviation, which is from the line to the point, is not the line's responsibly. The point is, by default, an outlier, and away from the mean. It is unfair for us to put the blame on the line of regression, instead, we only penalize the line of regression based on the deviation from the mean to itself, the rest of the deviation isn't its fault. The diagram in Figure 2.6 shows this graphically.