ST101 Unit 6: Regression

Contents:

Regression

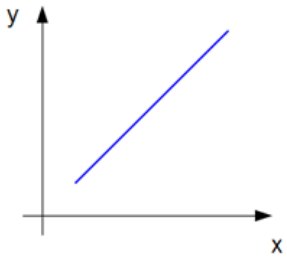
You should remember from the very beginning of this course that we can have data sets with more than one dimension. For example, the size of a house relative to its price:

In the first unit we saw different ways of presenting the data, like scatter-plots or bar charts, but we didn't look at what might be thought of as the 'Holy Grail' of statistics, fitting a line to the data points:

By the end of this unit you will be able to fit a line to data points like these, and you will even be able to state what the residual error is in that fit. This will allow you not only to understand the data, but also to make predictions about points that you have never seen before.

Lines

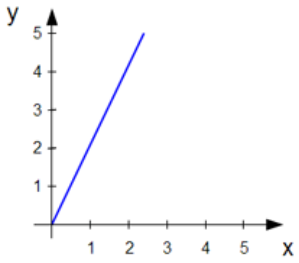
How do we specify a line? Well, suppose we have a horizontal axis, x, and a vertical axis, y. A straight line is commonly described by a functional relationship between x and y of the form:



$$y = bx + a$$

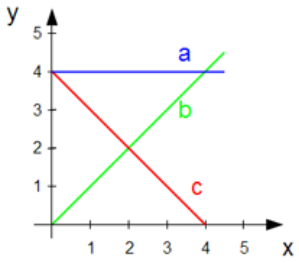
Consider the line described by the function y = 2x.

If x = 0 then y = 0. This means that the line must go through the origin. If x = 2 then y = 4. We now have two points and we can draw our line:



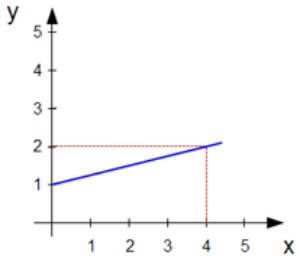
Pick the Line Quiz

What about the line, y = -x +4 i.e. where b = –1 and a = 4. Which of these lines best matches this function?



Find Coefficients Quiz

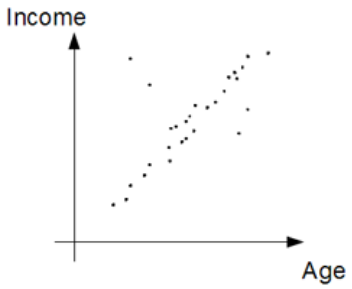
What are the coefficients a and b for the line shown below:

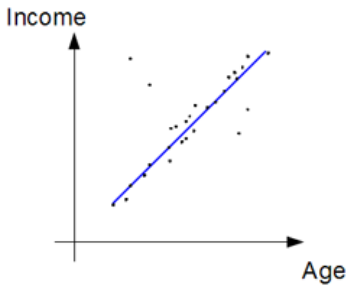


$$y = bx + a$$

Linear Regression

If we have 2-dimensional data, for example the age of a person and that person's income, linear regression is a technique for trying to fit a line that best describes that data:

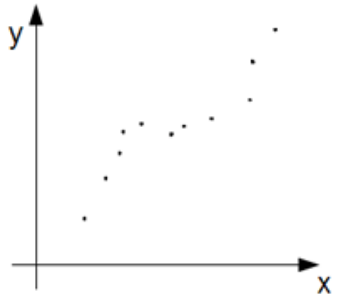




In linear regression we are given data (having more than 1-dimension), and we attempt to find the best line to fit the data. To put this differently, we are trying to identify the parameters a and b for the function:

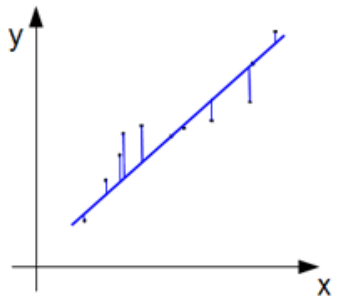
$$y = bx + a$$

We are trying to find the line that is the best-fit to our data, and the word 'best' is interesting in this context. Obviously it is impossible to draw a straight line that passes through every point in this data set:



This is what is known as **non-linear data** - the data points go up and down. Data often looks like this, even when the relationship between x and y is linear. This is because there is usually what is known as **noise** in the data. Noise is a random element in the data that we cannot explain.

In trying to find the 'best-fit' to the data, we are trying to find a line that minimises the differences between the data points and the line in the y-direction:



The reason for this is that we are assuming that our data results from some unknown linear function plus noise:

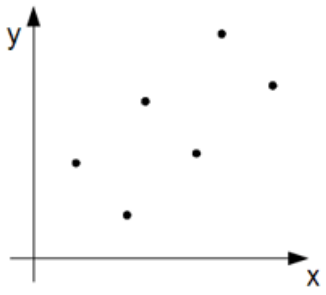
$$y = bx + a + noise$$

If the noise is assumed to be Gaussian, then minimising the quadratic deviation between the data points and the line provides the correct mathematical answer. In practice, what we are doing is adding, over all data points, the difference between our function and the y-value of the data point, squared:

$$\sum_{x_i} (bx_i + a - y_i)^2$$

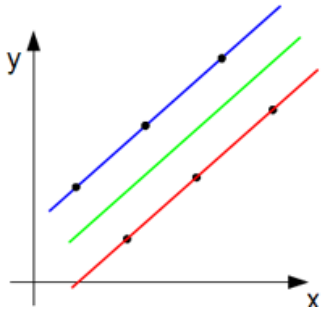
This is the distance that we are minimising.

Consider the following six data points:

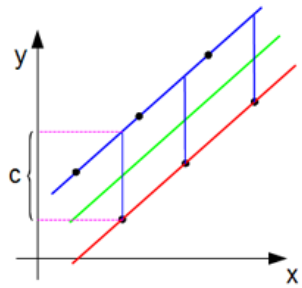


What is the best fit line for this data?

Well, consider the three possible lines shown:



In the case of the blue line, there is no loss for the three points that it passes through, but a fairly substantial loss for the other three points:



If the distance along the y-axis between the blue line and the data points is c, then the error, e, for the blue line is:

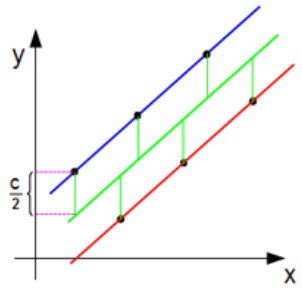
$$e = 3 \times c^2$$

since there are three points, and we are using the quadratic distance.

Similarly for the red line, it suffers no loss for the three points that it passes through, but a fairly substantial loss for the other three points, and the error, e , for the red line is also:

$$e = 3 \times c^2$$

In the case of the green line there are errors for all six points, but the error is only half as big in each case:



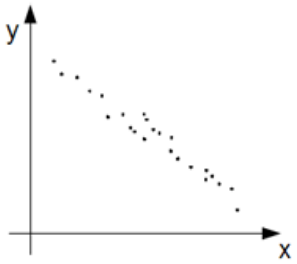
Now, the error is:

$$e = 6 \times \left(\frac{c}{2}\right)^2 = \frac{6}{4}c^2 = \frac{3}{2}c^2$$

This means that the total quadratic error for the green line is half as big as it is for either the blue or the red lines. This is because, when we use the quadratic, larger errors count much, much more than smaller errors. Clearly, in this case, the green line would be the best fit for these data points.

Negative or Positive Quiz

Will the parameters a and b be negative or positive for the data shown below?



$$y = bx + a$$

Regression Formula

The function $y = bx + a$ is the Holy Grail of linear regression, and much of statistics is concerned with how to use the data to determine the value of b, and the value of a. If we can do this with the data, then we have solved the problem of fitting the best line.

If the data comes in pairs:

x_1	x_2	x_3	…	x_N
y_1	y_2	y_3	…	y_N

Then we can calculate b using the formula:

$$b = \frac{\sum_i \left[\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)\right]}{\sum_i \left(x_i - \bar{x}\right)^2}$$

Note that $\bar{x} = mean(X_i)$ and $\bar{y} = mean(Y_i)$.

Previously, we've used μ for the mean, but now that we have more than one variable we are going to use the bar-notation.

This formula shouldn't be entirely unfamiliar. When we calculated the variance, we used: $(x_i - \bar{x})^2$, whereas now we have 2-dimensional data we are using $(x_i - \bar{x})(y_i - \bar{y})$.

Now that we know b, and we know that:

$$y = bx + a$$

It turns out that this function is also true for the average values, $\bar{x}$ and $\bar{y}$. We can calculate the value for a using:

$$a = \bar{y} - b\bar{x}$$

Let's try it out in an example.

Suppose we have the following data sample:

x	y
6	7
2	3
1	2
-1	0

Now, with this data set we notice that the y-value is always exactly 1 larger than the x-value, so:

$$y = x + 1$$

This means that b = 1 and a = 1.

Let's see what we get when we use the formula. The first thing we do is to calculate the means:

$$\bar{x} = \frac{6+2+1-1}{4} = \frac{8}{4} = 2$$

$$\bar{y} = \frac{7+3+2+0}{4} = \frac{12}{4} = 3$$

Now we can evaluate b using:

$$b = \frac{(6-2)(7-3)+(2-2)(3-3)+(1-2)(2-3)+(-1-2)(0-3)}{(6-2)^2+(2-2)^2+(1-2)^2+(-1-2)^2}$$

$$b = \frac{16+0+1+9}{16+0+1+9} = 1$$

Now we can calculate a using:

$$a = \bar{y} - b\bar{x} = 3 - (1 \times 2) = 1$$

Which agrees with our initial observation.

Regression Quiz

Calculate the values for a and b for the following data set:

x	y
4	7
3	9
7	1
2	11

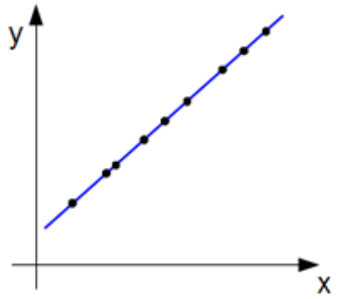
Correlation

This section is all about correlation. Correlation is a measure of how closely two variables are related. We can calculate something called the correlation coefficient which gives us a measure of how closely two variables are related.

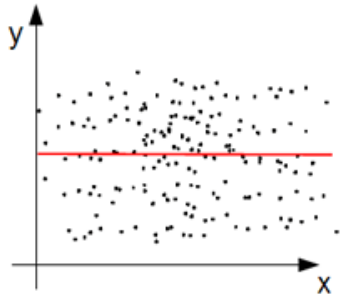
The correlation coefficient, r, will have a value between -1 and 1, so that:

$$-1 < r < 1$$

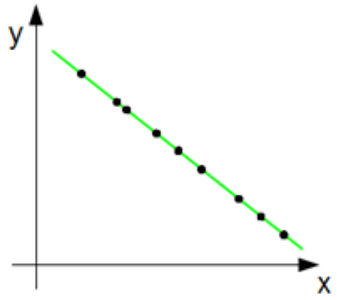
If the data is completely linear, without noise, as shown below, then we can fit a line exactly to the data and the correlation will be 1:



If the data is completely unrelated then the correlation will be 0:



The correlation coefficient can also be -1 in the case where the data is still perfectly aligned to the line, but where there is a negative relationship between the two variables as shown:



Correlation From Regression Quiz

Suppose that we ran linear regression on our data, and we found that:

b = 4

a = -3

Where these values describe this linear relationship between x and y:

y = 4x – 3

Which of the following statements about the correlation coefficient, r, is true?

- r is positive
- r is negative
- r = 0
- We can't tell

Correlation Formula

So, to summarise what we know so far about the correlation coefficient:

- It has a value between -1 and 1
- It tells us how "related" two variable are
- Both +1 and -1 describe perfectly linear data

So how do we compute a value for r?

Well, one way to compute the correlation coefficient is very similar to the way that we calculated the value of b when we looked at linear regression:

$$r = \frac{\sum_i \left[\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)\right]}{\sqrt{\sum_i \left(x_i - \bar{x}\right)^2 \cdot \sum_i \left(y_i - \bar{y}\right)^2}}$$

Where: $\bar{x} = \frac{1}{N}\sum x_i$ and $\bar{y} = \frac{1}{N}\sum y_i$

Now this is probably the most complex formula that you have encountered so far in this class, but you will see it is related to a lot of the stuff we have seen before, like variance and so on, and that using it is not that difficult in practice.

If we look at the denominator:

$$\sqrt{\sum_i \left(x_i - \bar{x}\right)^2 \cdot \sum_i \left(y_i - \bar{y}\right)^2}$$

We notice that the term within the square root is the product of the non-normalised variance of x and the non-normalised variance of y.

The numerator is a kind of "mixed-variance":

$$\sum_i \left[\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)\right]$$

This is sometimes called the **covariance**, because it calculates the variance over two co-occurring variables. However, you will have noticed that, once again, there is a missing normaliser. What has happened is that the normalisers on the numerator and denominator cancel each other out.

Computing Correlation

Let's try the formula out on some data.

x:	3	4	5
y:	7	8	9

It is easy to see that the mean of x is:

$$\bar{x} = \frac{1}{3}\left(3+4+5\right) = \frac{12}{3} = 4$$

and the mean of y is:

$$\bar{y} = \frac{1}{3}\left(7+8+9\right) = \frac{24}{3} = 8$$

Armed with these mean values we can calculate the following:

$x - \bar{x}$	-1	0	1
$y - \bar{y}$	-1	0	1

We can plug these values into our equation, and we get:

$$r = \frac{\left(-1\times-1\right)+\left(0\times0\right)+\left(1\times1\right)}{\sqrt{\left[\left(-1\right)^2+\left(0\right)^2+\left(1\right)^2\right]\cdot\left[\left(-1\right)^2+\left(0\right)^2+\left(1\right)^2\right]}} = \frac{2}{\sqrt{2\times2}} = 1$$

Let's look at a different dataset,:

x:	3	4	5
y:	2	5	8

Again, we calculate the means:

$$\bar{x} = \frac{1}{3}(3+4+5) = \frac{12}{3} = 4$$

$$y = \frac{1}{3}(2+5+8) = \frac{15}{3} = 5$$

Guess r Quiz

Before we do the calculation, let's see what your intuition says about the value for r in this case. Which of the following do you think is correct?

- r = 1
- r = 3
- r = 2
- r = 0

Now let's tabulate the differences from the means:

$x - \bar{x}$	-1	0	1
$y - \bar{y}$	-3	0	3

Again, we can plug the values into our formula to get:

$$r = \frac{(-1\times -3)+(0\times 0)+(1\times 3)}{\sqrt{\left[(-1)^2+(0)^2+(1)^2\right]\cdot\left[(-3)^2+(0)^2+(3)^2\right]}} = \frac{6}{\sqrt{2\times 18}} = 1$$

Which is as we expected.

Reverse Order

Let's see what happens if we change the order of the y-values as shown:

x:	3	4	5
y:	8	5	2

Since only the order has been changed and the actual values are unchanged, the mean values will be the same as before:

$$\bar{x} = \frac{1}{3}(3+4+5) = \frac{12}{3} = 4$$

$$y = \frac{1}{3}(8+5+2) = \frac{15}{3} = 5$$

Now, since the y-values decrease as the x-values increase we would expect the correlation coefficient to be negative. Let's see what happens when we calculate the value of r.

Tabulating the differences from the means gives:

$x-\bar{x}$	-1	0	1
$y-\bar{y}$	3	0	-3

And once again, we just plug the values into our formula:

$$r = \frac{(-1\times 3)+(0\times 0)+(1\times -3)}{\sqrt{\left[(-1)^2+(0)^2+(1)^2\right]\cdot\left[(3)^2+(0)^2+(-3)^2\right]}} = \frac{-6}{\sqrt{2\times 18}} = -1$$

So the data is perfectly *negatively* correlated.

Uncorrelated Data

Let's try something a little bit tricky. Let's change the y-values to:

x:	3	4	5
y:	8	5	8

What seems to be happening is that the y-values start to decrease as the x-values increase, but then they start to increase again. While there may be a relationship between x and y, it doesn't appear to be a linear one, and we would expect the correlation to be zero.

Let's do the calculation. The means are now:

$$\bar{x} = \frac{1}{3}(3+4+5) = \frac{12}{3} = 4$$

$$y = \frac{1}{3}(8+5+8) = \frac{21}{3} = 7$$

Now when we tabulate the differences from the means we get:

$x - \bar{x}$	-1	0	1
$y - \bar{y}$	1	-2	1

And if we plug the values into our formula we get the correlation coefficient:

$$r = \frac{(-1\times 1)+(0\times -2)+(1\times 1)}{\sqrt{\left[(-1)^2+(0)^2+(1)^2\right]\cdot\left[(1)^2+(-2)^2+(1)^2\right]}} = \frac{0}{\sqrt{2\times 6}} = 0$$

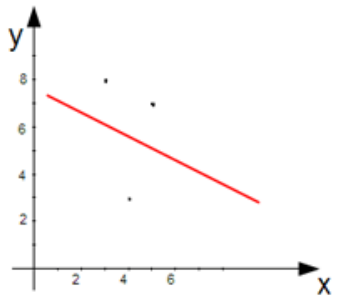
Of course, we didn't actually need to calculate the denominator. Once we knew the numerator was zero we also knew that r = 0.

Final Example

Let's look at a final example. This is our data:

x:	3	4	5
y:	8	3	7

Clearly, this doesn't look very correlated, although it looks *more* correlated than the data in the last example since the final y-value hasn't increased by as much. We might therefore expect that the correlation coefficient will be less than 1, and if we plot the data on a graph, we see that it should also be negative:



As usual, the first thing we do is to calculate the mean values for x and y:

$$\bar{x} = \frac{1}{3}(3+4+5) = \frac{12}{3} = 4$$

$$y = \frac{1}{3}(8+3+7) = \frac{18}{3} = 6$$

Now when we tabulate the differences from the means we get:

$x - \bar{x}$	-1	0	1
$y - \bar{y}$	2	-3	1

And if we plug the values into our formula we get the correlation coefficient:

$$r = \frac{(-1\times 2)+(0\times -3)+(1\times 1)}{\sqrt{\left[(-1)^2+(0)^2+(1)^2\right]\cdot\left[(2)^2+(-3)^2+(1)^2\right]}} = \frac{-1}{\sqrt{2\times 14}} = -0.189$$

So there is a negative correlation, as we suspected, but it is weak. This data really isn't well described by a linear function.

Summary

You now understand the basics of correlation coefficients. As we said earlier, the correlation coefficient:

- has a value between -1 and 1
- tells us how "related" two variable are

We also now know that:

- If $r > 0$ there is a positive relationship between x and y.
- If $r < 0$ there is a negative relationship between x and y.
- If $r = 0$ there is no relationship between x and y.

Furthermore, we have seen that $|r| \to 1$ as the relationship between x and y becomes increasingly linear. Both $r = +1$ and $r = -1$ describe perfectly linear data without any noise or deviation from the line.

Correlation is a very powerful tool. For any data set with multiple variables, such as salary versus age, you can now tell how closely the variables relate to each other using the relatively simple formula:

$$r = \frac{\sum_i \left[\left(x_i - \bar{x} \right) \left(y_i - \bar{y} \right) \right]}{\sqrt{\sum_i \left(x_i - \bar{x} \right)^2 \cdot \sum_i \left(y_i - \bar{y} \right)^2}}$$

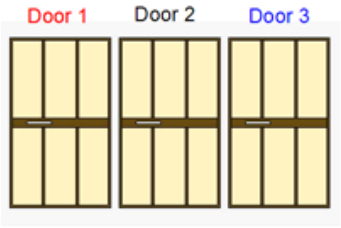
Monty Hall Problem (Optional)

This section should be fun. It is, however, entirely optional.

Some of you may be familiar with Monty Hall. He ran a US game show called **Let's Make a Deal** from 1963. The show ran for many years.

At the heart of the game show was a puzzle. This puzzle puzzles statisticians to the present day. In this unit, you get the chance to solve the Monty Hall program, and you also get the chance to program it, and verify that the assertion – which in not entirely obvious – is actually correct.

In the game there are three doors. Behind one door is a car. Monty knows which door, but he won't tell you where the car is.



The game is then played as follows:

You get to choose a door. Say, for the sake of argument that you pick door number 2. If the car is behind door 2, then you win the car, otherwise you win nothing. So far, so good.

Now comes the interesting bit. Obviously, at least one of the two remaining doors doesn't have a car behind it. Now, Monty knows which door has the car, and he reveals one of the doors that you didn't choose that doesn't have the car.

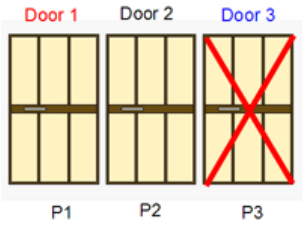
Monty then asks you if you want to switch from the door you have chosen to the other closed door. Do you want to change your choice in the hope that you will increase your chance of winning the car?

What makes this problem interesting is that when Monty opened the door, he really didn't tell you anything you didn't already know. You already knew in advance that one of the two doors didn't contain the car. The fact that Monty has just opened a door should give you zero information about which of the remaining doors the car is behind.

If you chose door 2, and Monty then opened door 3, all you now know is that door 3 doesn't contain the car. Why should door 1 now be more likely than door 2?

Door Chance Quiz

Given that you chose door 2, and Monty then revealed that door 3 didn't contain the car, what are the probabilities, P1, P2 and P3 that the car is behind each door?



Of course it is easy to see that P3 = 0. We already knew that the car wasn't behind door 3., But don't worry if you weren't able to figure out the other two probabilities, the answer to this problem is entirely non-intuitive.

We can perhaps best understand what is happening by building a truth table.

There are three possible true locations for the car, and we know that each of these possible locations has a probability of 1/3.

When we choose a door, Monty will open one of the other doors. We can construct the truth table as shown:

	True location	Monty Hall	Probability	Normalised probability
P(1) = 1/3	1	1	0	
	1	2	0	
	1	3	1/3	
P(2) = 1/3	2	1	1/6	
	2	2	0	
	2	3	1/6	
P(3) = 1/3	3	1	1/3	
	3	2	0	
	3	3	0	

Let's say, for the sake of argument that we have chosen door number 2. Now, we know that on the first row of the truth table, if the car is behind door number 1 then the probability that Monty will open door number 1 is zero. On the second line, the probability that Monty will open door 2 is also zero because we have chosen door number 2.

So the only option remaining is for Monty to open door 3. This has a probability of 1, but the posterior probability must take account that the total probability for P(1) = 3, so the probability becomes $1 \times 1/3 = 1/3$.

Similarly for the case where the car is behind door number 3, the probability that Monty will open door number 3 is zero, as is the probability that he will open the door that we have chosen (door number 2). The only remaining choice is door number 1.

If the car is behind door number 2, and we have chosen door number 2 then there is a fifty chance for door number 1 or door number 3. The posterior probability for both doors is therefore $1/2 \times 1/3 = 1/6$.

Now we have our truth table based on the fact that we chose door number 2.

Let's say that Monty opens door number 3. Now every other case, in which he might have picked door number 1 or door number 2 has zero probability. The only events that remain are highlighted here:

	True location	Monty Hall	Probability	Normalised probability
P(1) = 1/3	~~1~~	~~1~~	~~0~~	
	~~1~~	~~2~~	~~0~~	
	1	**3**	**1/3**	**2/3**
P(2) = 1/3	~~2~~	~~1~~	~~1/6~~	
	~~2~~	~~2~~	~~0~~	
	2	**3**	**1/6**	**1/3**
P(3) = 1/3	~~3~~	~~1~~	~~1/3~~	
	~~3~~	~~2~~	~~0~~	
	3	**3**	**0**	**0**

Now we just need to normalise these values so that the total probability adds up to 1. Currently the sum of the probabilities is $1/3 + 1/6 = 1/2$. If we divide each individual probability by this sum we get the normalised probabilities as shown. These are the true posterior probabilities given that we chose door number 2, and Monty opened door number 3.

Clearly we should switch our choice every time, as this will double our chances of winning the car!

Simulation

Write a function simulate() that runs 1000 iterations of a simulation of the Monty Hall problem and so empirically verify the probabilities that we have just calculated.

The function should count how many times you win the car in a variable, K, and then return K divided by the number of iterations, N.

Python includes the built-in function randint(). This generates random integers in the range specified by the function arguments, so:

```
randint(1,3)
```

with return a random integer in the range 1 to 3 (which is exactly what you will need in order to pick a random door).

You will also have to simulate the actions of Monty Hall. Sometimes these actions will be deterministic, at other times they will be stochastic (random). Recall that we saw both types of action when we constructed the truth-table.

Once the "Monty-simulator" has picked a door, flip your choice to the remaining door. If that matches the true location then you should increment K, otherwise not.

When you run this 1000 times, the output from your function should be approximately equal to 2/3.

The assignment will require some real knowledge of Python and is therefore considered to be a challenging problem. Good luck.

Answers

Pick the Line Quiz

c

Find Coefficients Quiz

a = 1
b = 0.25

Negative or Positive Quiz

a is positive. As x increases, y decreases, so b is negative. This is an example of ***negative correlation***.

Regression Quiz

$\bar{x} = 4$

$\bar{y} = 7$

$b = \frac{28}{14} = 2$

a = 15

Correlation From Regression Quiz

- **r is positive**
- r is negative
- r = 0
- We can't tell

Guess r Quiz

r = 1

Door Chance Quiz

P1 = 0.667
P2 = 0.333
P3 = 0

Simulation

```
from random import randint

N = 1000

def simulate(N):
    K = 0
    for i in range(N):
        TrueLoc = randint(1,3)
        guess = randint(1,3)
        if TrueLoc == guess:
            monty = randint(1,3)
            while monty == TrueLoc:
                monty = randint(1,3)
        else:
            monty = 6 - TrueLoc - guess
            switch = 6 - guess - monty
            if switch == TrueLoc:
                K = K + 1
    return float(K) / float(N)

print simulate(N)
```