The big picture

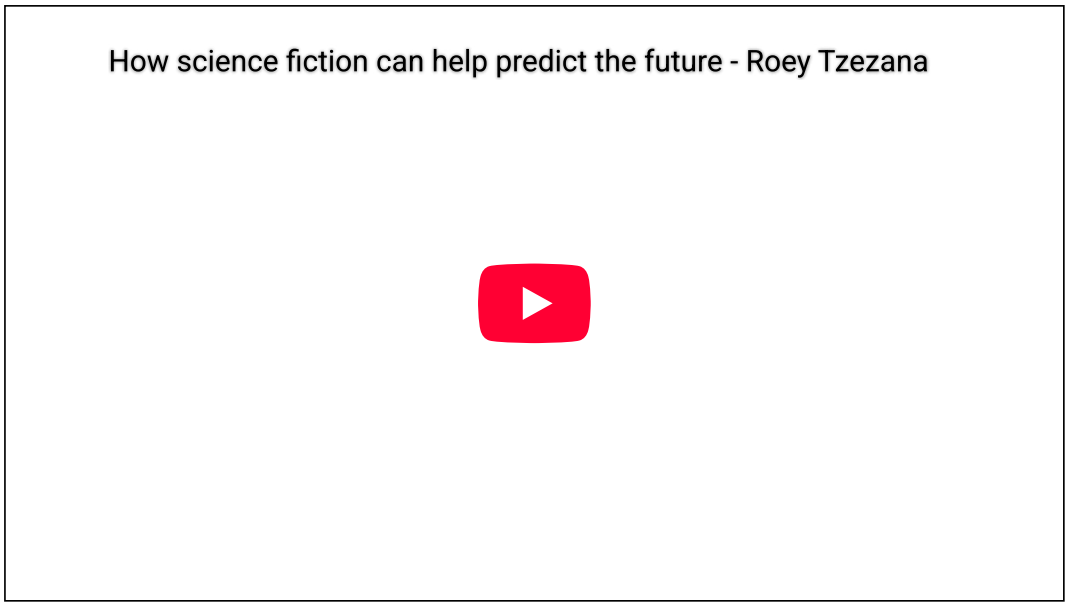
In our discussion of linear correlationin sections 4.4.1 (/study/app/math-aa-hl/sid-134-cid-761926/book/an-introduction-to-linear-correlation-id-25527/), 4.4.2 (/study/app/math-aa-hl/sid-134-cid-761926/book/pearsons-r-correlation-coefficient-id-25528/) and 4.4.3 (/study/app/math-aa-hl/sid-134-cid-761926/book/predicton-id-25529/), we discovered how to determine whether two variables are significantly related and how to find the least squares regression line $y = ax + b$ using technology. There you also learned how to use the equation to predict, or interpolate , the value of y from a given value of x . This is possible because the variable y is dependent on the independent variable x . You were warned, however, that attempting to predict the value of x from a value of y would produce a somewhat unreliable estimate. In this subtopic, we will explore why that prediction is unreliable and discover how to produce a more reliable prediction when possible.

Theory of Knowledge

Using linear regression to make predictions brings several significant questions to mind. For example:

- How reliable are the predictions we make?
- Can predictions made using statistics be considered truth?

Overview (/study/ap aa-hl/sid-134-cid-761926/o



As you think about the role that predictions play in shaping our world, another thing to consider is whether the prediction motivated someone to pursue that outcome. For example, did Arthur C. Clarke *predict* the use of tablet computers, or did modern-day product developers get the idea from his vision of the future?

Concept

It is important to remember that even though we are finding a *more reliable* method of predicting x from a value of y , it is still an approximation. Just as we found using theoretical and experimental probabilities, any time we are making a prediction, we are approximating what we believe is most likely to occur.

4. Probability and statistics / 4.10 Further linear regression

Predicting x from y

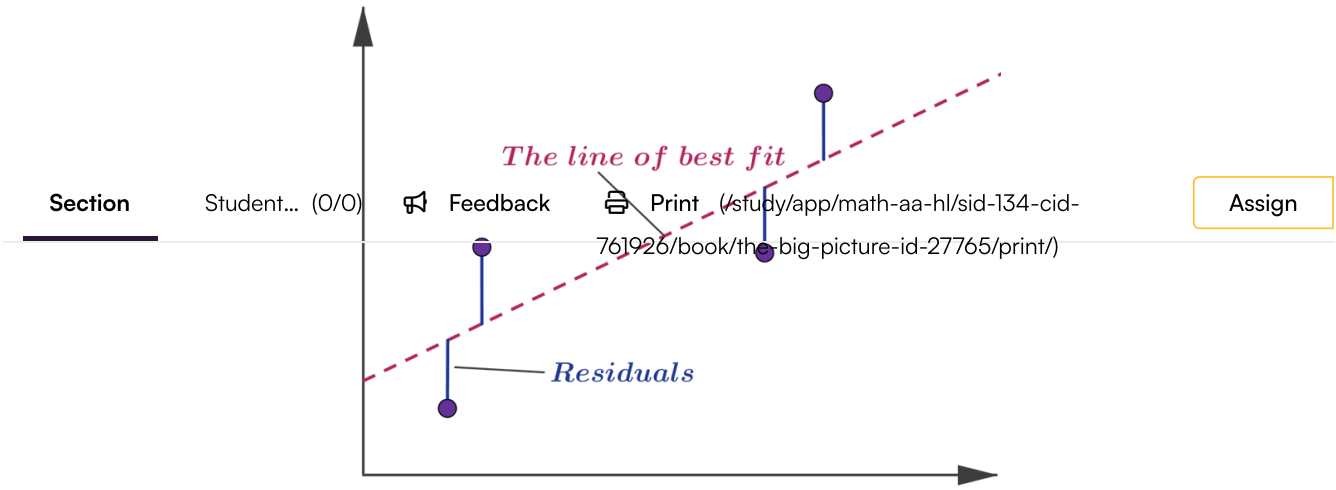
Student view

Overview (/study/ap aa-hl/sid-134-cid-761926/o

Independent and dependent variables

The process of linear regression

The process of finding the least-squares regression line utilises residuals, the difference between the value predicted by the equation describing the dependence of y on x , and the measured value of y . To find the regression line without the help of technology, you have to minimise the sum of the squares of these values, which is where the term ‘least-squares regression’ originates from. As shown in the diagram below, the residuals are measured vertically from each data point to the line. These values are squared to eliminate negatives.



More information

The diagram illustrates a graph of a least-squares regression line, also known as the line of best fit. This dashed pink line runs diagonally upwards, indicating a positive trend. The graph has axes, which are unlabelled, but they typically represent x and y coordinates.

The diagram highlights data points as filled purple circles placed above or below the line of best fit, illustrating their actual measured values compared to predicted values on the line. Vertical blue lines connect each data point to the regression line, representing the residuals, or the difference between actual and predicted values.

The line of best fit and residuals are labeled on the graph. The residuals are measured vertically from each data point to the line, demonstrating the concept of minimizing these differences to find the best fit for a given set of data points.

[Generated by AI]

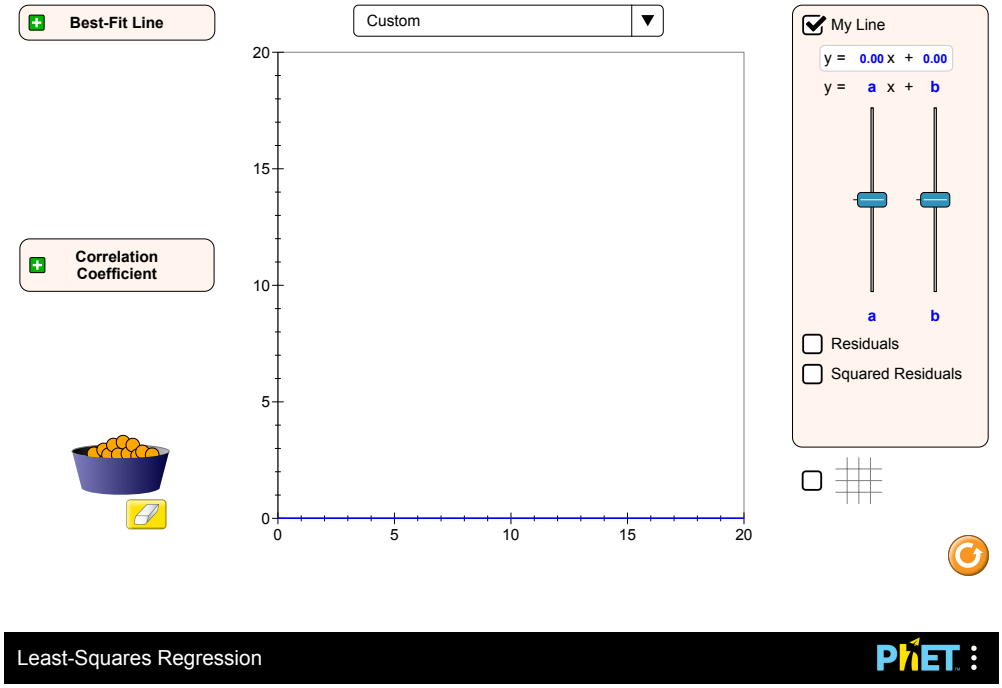
Student view

Overview (/study/app/aa-hl/sid-134-cid-761926/o

Making connections

The formulae for variance and standard deviation also deal with distances when they require you to subtract the mean from each value (see section 4.9.1 (/study/app/math-aa-hl/sid-134-cid-761926/book/the-normal-distribution-id-25667/)), and negatives are eliminated by squaring those distances like they are here. Do you think there might be a link between these concepts?

You can explore the concept of residuals and their squares using the applet below. Drag orange points onto the graph to create custom data or select pre-defined data from the drop-down list.



Interactive 1. Least-squares Regression.

More information for interactive 1

This interactive enables users to explore the concept of linear regression, residuals, and squared residuals through hands-on experimentation. By manipulating data points and adjusting the parameters of a line, users can visualize how the best-fit line minimizes the sum of squared residuals, providing an intuitive understanding of the least squares method. The applet also displays the correlation coefficient, offering insight into the strength and direction of the linear relationship between variables.

The interface allows users to drag orange points onto the graph to create custom datasets or select pre-defined data from a drop-down menu. Users can then adjust the slope (a) and intercept (b) of a line to fit the data manually, comparing their custom line to the calculated best-fit line. By toggling options like "Residuals" and "Squared Residuals,"

Student view

Overview (/study/ap aa-hl/sid-134-cid-761926/o

users can observe the vertical distances between data points and the line, as well as how squaring these distances affects the overall fit. The applet dynamically updates calculations, such as the sum of residuals and squared residuals, reinforcing the learning process.

By experimenting with different lines and observing how residuals change, they can internalize why the best-fit line is optimal. This hands-on approach demystifies abstract statistical concepts, making them more accessible and memorable. Ultimately, the applet fosters active learning, encouraging students to explore, hypothesize, and discover the principles of linear relationships on their own.

Finding the value of x given y

Icicles

Credit: fhm GettyImages

Example 1

★☆☆

Consider the data in the table, which shows the relationship between x , the length of an icicle in centimetres, and y , its weight in grams.

x (cm)	y (g)
16	89
31	239

Student view

Overview (/study/app/aa-hl/sid-134-cid-761926/o

x (cm)	y (g)
39	297
55	402
84	580
115	802

Find out whether the weight of the icicle depends on its length, and if so, find the equation linking the two variables.

Using technology, you can find the correlation coefficient $r = 0.997280458$, which indicates a very strong linear correlation. Since that is the case, it is also appropriate to use the least-squares regression line given by the calculator, that is, the equation for the line of x on y in the form $y = ax + b$.

Therefore , $y = 6.929864253x + 8.807692308$, or $y = 6.93x + 8.81$ rounded to 3 significant figures.

Now suppose you wanted to approximate the length of an icicle given that it weighs 350 g. If you were to use the equation you just found, you would substitute $y = 350$ and solve for x like this.

$$\begin{aligned} 350 &= 6.93x + 8.81 \\ 341.19 &= 6.93x \\ x &\approx 49.2 \text{ cm} \end{aligned}$$

Generalising this process by solving the original equation of the line for x in terms of y you would get $x = \frac{y}{a} - \frac{b}{a}$, giving this equation: $x = 0.144y - 1.271$.

It might seem reasonable to find values of x in this way, but it is not the best method. That is because in order to predict a value of x from a given value of y , you are treating x as the dependent variable and y as the independent variable.

In this way, you can find the equation $x = c + dy$.

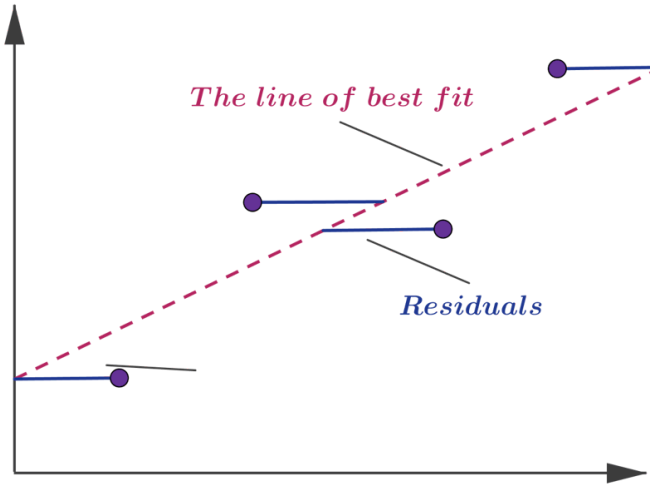
Student view

But what difference does it make if we treat y as the independent variable?

Overview (/study/ap aa-hl/sid-134-cid-761926/o

The difference is in the residuals – if y is the independent variable, then the residuals are measured horizontally from each point to the line, as illustrated in the diagram below.

This creates a slight difference in the values found through the process of linear regression.



More information

The image is a scatter plot illustrating the concept of linear regression. The X-axis is horizontal, and the Y-axis is vertical, although no specific labels or scale markings are visible. There are several data points marked as purple dots distributed around a dashed pink line, which represents the line of best fit. Lines extend vertically from each data point to the line of best fit, representing the residuals. The pink dashed line signifies the trend direction of the data points. Text labels are present: 'The line of best fit' in pink and 'Residuals' in blue, referring to the corresponding elements of the graph. This setup visually demonstrates the differences between actual data points and their predicted values on the line of best fit.

[Generated by AI]

Another question to ask is whether it makes sense to think of x as being dependent on the value of y . While you already know that correlation does not imply causation, the dependent variable should logically depend on the independent variable.

In our example, it makes sense to think of the weight of an icicle being dependent on the length or to think of the length being dependent on the weight, so attempting to predict one from the other makes sense either way.

Student view

Overview (/study/app/aa-hl/sid-134-cid-761926/o

International Mindedness

Correlations are often used when analysing relationships between variables on an international scale, but the broader the context, the more variables affect any given situation. Avoiding any implication of causation is particularly important in these instances. Can you think of a pair of related variables that might have a cause-and-effect relationship in your culture but might not even be correlated in another? How could societal structure impact this?

Finding the true line $x = c + dy$

Now that you understand the need to approach solving for x differently, let's look at how to find the form of the least-squares regression equation $x = c + dy$.

Your calculator likely does not have a built-in function that makes y the independent variable, but you can accomplish this simply by using the usual Linear Regression function on your calculator and reversing how you input the values of x and y. Since the calculator doesn't know what these variables stand for, you need to interpret the variables the calculator gives you in reverse.

For example, if you used the data from the table in **Example 1** and entered the weights of the icicles as the values of x and the lengths of the icicles as the values of y, the calculator would give you the equation $y = -0.956 + 0.144x$ rounded to 3 significant figures, but you would need to interpret it as the equation $x = -0.956 + 0.144y$.

How does the strength of the correlation, the value of r, compare with what you found before?

Exam tip

When doing problems that depend on a calculator function, identify which function you are using to show your working and earn full marks.

Be aware

While you may write down the equation using rounded values, use the exact values on your calculator to find your answer to ensure accuracy.

Student view

Overview (/study/app/aa-hl/sid-134-cid-761926/o

This new equation enables you to evaluate values of x more reliably. This means that you can approximate the length of an icicle that weights 350 g as

$$x = -0.956 + 0.144 \times 350 \approx 49.4 \text{ cm}.$$

How does this compare with the value we found with the original equation? This may seem like an insignificant difference, but the larger the values in the data set are and the weaker the correlation, the larger the difference will be.

Important

There are three key things to remember when finding the equation $x = c + dy$:

1. Confirm that the value of x could logically be dependent on the value of y.
2. Use the same calculator function, but reverse the data you enter for x and y.
3. Reversing the variables does not change r.

Activity

Find three or more sets of data from your notes that had strong, moderate or weak correlations and use them to predict the value of x given a value of y using both the equations $y = ax + b$ and $x = c + dy$.

What do you notice about your predictions for the three sets?

3 section questions

Question 1

Difficulty: ★☆☆

Write down the word that completes the statement: In order to properly predict the value of x from a given value of y, you need to find the least-squares regression line while treating y as the _____ variable.

independent ✓

Student view

Accepted answers

independent, independant

Overview (/study/ap aa-hl/sid-134-cid-761926/o

Explanation

To predict the value of x from a value of y, it needs to be possible to interpret x as dependent on y. Therefore, you need to treat y as the independent variable when using linear regression.

Question 2

Difficulty: ★☆☆

Alejandro is exploring the connection between the heights (x) in centimetres of his berry bushes and the kilograms of berries (y) he harvested from each. The data he collected is shown in the table.

x (cm)	y (kg)
90	1.1
100	2.8
110	4.3
120	5.9
130	8.1
140	9.9

Find the equation used to predict the height of a bush given the mass (in kilograms) of berries harvested from it.

1 $x = 5.67y + 84.7$ ✓

2 $y = 0.176x - 14.9$

3 $x = 0.176y - 14.9$

4 $y = 5.67x + 84.7$

Explanation

Using the linear regression function on your calculator, enter the values of y as the values of the independent variable and the values of x as the values of the dependent variable.

Once the calculator returns the equation, switch the x and y variables to get the equation $x = 5.67y + 84.7$.

Student view

Overview (/study/ap aa-hl/sid-134-cid-761926/o

Question 3

Difficulty: ★★☆

Chloe is exploring the relationship between the median monthly income of European countries (x) in thousands of euros and the annual value of its exports (y) in billions of euros. The data she collected is shown in the table below.

	x (thousand €)	y (billion €)
Germany	3.88	1557
Netherlands	2.855	723
France	2.957	568
Italy	2.595	547
Belgium	3.401	467
Spain	2.189	345
Poland	1.192	261
Czech Republic	1.26	202
Austria	2.688	185

Find the approximate median monthly income of a country that exports €625 billion in goods.

1 2.69 ✓

2 406 188

3 2.81

4 213 591

Explanation

First, use the linear regression function on the calculator and reverse the variables in the equation it gives to get $x = 0.00153y + 1.730$.

Student view

Finally, since y is in billions, substitute 625 into the equation to find $x = 0.00153\,(625) + 1.730 = 2.69$.

Therefore, the median monthly income of a European country with €$625\,000\,000\,000$ in annual exports would be approximately €2690.

4. Probability and statistics / 4.10 Further linear regression

Checklist

What you should know

By the end of this subtopic you should be able to:

- recognise when it is appropriate to try to predict x from a given value of y
- use technology to find the equation $x = c + dy$ for a set of data.

4. Probability and statistics / 4.10 Further linear regression

Investigation

While you are only required to calculate linear regression equations using technology for the IB Diploma Programme, finding the equation of the least-squares regression equation by hand will deepen your understanding of the process in general and of residuals in particular. Here are two key facts that will enable you to find the equation $y = ax + b$.

1. $a = \dfrac{\sum (x - \bar{x})\,(y - \bar{y})}{\sum (x - \bar{x})^2}$
2. $(\bar{x},\bar{y})$ is a point on the line .

(Remember that $\bar{x}$ and $\bar{y}$ are the means of all the x and y values, respectively.)

Overview (/study/ap aa-hl/sid-134-cid-761926/o

Find or collect a set of bivariate data in which either variable could be the independent variable.

Calculate the least-squares regression equation $y = ax + b$.

Can you identify the portions of the formula for a that involve residuals?

How can you adapt the formula for a to help you find $x = c + dy$?

Find it by hand, and then use your calculator to check the accuracy of both the equations.

Student view