0 (https://intercom.help/kognity)

Overview
(/study/ap
ai-
hl/sid-
132-
cid-
761618/ov

Teacher view

Table of
contents

Notebook

Glossary

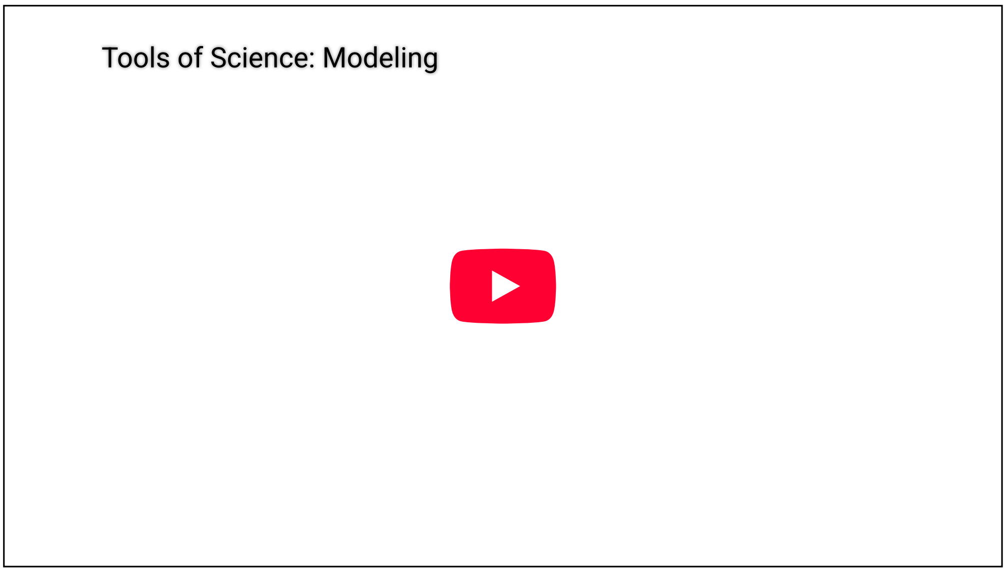
Reading
assistance

Index

4. Probability and statistics / 4.13 Non-linear regression

The big picture

Section Student... (0/0) Feedback Print (/study/app/math-ai-hl/sid-132-cid-761618/book/the-big-picture-id-27533/print/) Assign



Models are integral to many scientific investigations, as the video explains. However, models are not just important in science. They are also crucial to areas such as technological advancements, urban planning and the financial policies of corporations. As shown in the 2011 American movie *Moneyball* , models can be used in sports, too, in an attempt to give teams statistical advantages over their opponents.

However, once a model has been created, how can you determine how well the model fits the data being tested? Is there a way to quantify the effectiveness of the model?

Concept

Models can be used to explore the patterns and connections between quantities or properties in the world around you. Correlation and regression are powerful tools in identifying these patterns and in determining how well a model matches the real-world data.

Student
view

Overview (/study/app/math-ai-hl/sid-132-cid-761618/ov

4. Probability and statistics / 4.13 Non-linear regression

Sum of square residuals

Section Student... (0/0) Feedback Print (/study/app/math-ai-hl/sid-132-cid-761618/book/sum-of-square-residuals-id-27534/print/) Assign

Making connections

In section 4.4.3 (/study/app/math-ai-hl/sid-132-cid-761618/book/predicton-id-26081/) , you learned that regression lines are a mathematical way of working out a line of best fit between two variables plotted on a scatter graph. You also learned how to find the equation of the regression line using technology. In this section we will take a closer look at some of the algebraic theory behind the equation.

Ben is investigating whether the height of a catcher in baseball is related to their ‘pop time’. In baseball, a catcher squats behind home plate to catch missed balls. Pop time refers to how quickly a catcher can stand up and throw the ball to second base after catching it. Watch this video (https://www.youtube.com/watch?v=L_TXXQhyHjw) to see an example.

For his investigation, Ben has collected data for five different catchers and organised them in the table below.

Be aware

For the purpose of this discussion, you will only analyse a data set of five points. However, if you wanted to fully investigate whether there was a relationship between these two variables you would need a much larger data set.

Pop time (seconds)	1.81	1.95	1.91	2.25	2.10
Height (inches)	69	71	73	74	76

Example 1

★☆☆

Find the mean point of Ben’s data.

Student view

Overview
(/study/ap
ai-
hl/sid-
132-
cid-
761618/ov

Ben is trying to determine whether height affects pop time, or in other words, whether a person's height can **explain** their pop time. Therefore, height will be the explanatory variable and pop time will be the responsive variable.

Steps	Explanation
Use the formula for mean.	$\bar{x}_{\text{height}} = \frac{\sum x}{n} = \frac{69+71+73+74+76}{5} = 72.6$ $y_{\text{pop time}} = \frac{\sum y}{n} = \frac{1.81+1.95+1.91+2.25+2.10}{5} = 2.004$
Give the values as a coordinate point.	$\therefore$ the mean point is $(72.6, 2.004)$

Important

When carrying out regression line calculations, it is important to clearly identify which variable is the explanatory variable (the likely cause) and which one is the responsive variable (the likely effect) . The explanatory variable will be plotted on the x - axis and the responsive will be plotted on the y - axis.

Consider the importance of the mean point found in **Example 1**. As Ben investigates the data, he wants to know whether there is a trend that can be explained by a relationship between height and pop time. In other words, as height increases above 72.6 inches does the pop time increase or decrease from 2.004 s? What about when the height decreases below 72.6?

Making connections

Considering these questions should help you understand why the regression line *must* go through the mean point, as you learned in section 4.4.3 (/study/app/math-ai-hl/sid-132-cid-761618/book/predicton-id-26081/).

Using technology, Ben finds that the equation of the regression line is

$$y = 0.04856x - 1.52158$$

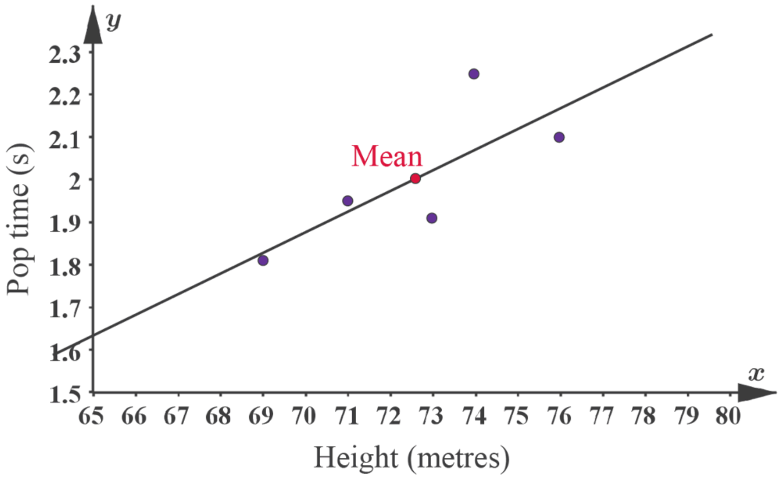
Student view

Overview
(/study/ap
ai-
hl/sid-
132-
cid-
761618/ov

Be aware

It is important to ensure you have chosen the correct variable for the explanatory variable. Ben has correctly chosen height as the explanatory variable for the above equation. What would the equation be if he incorrectly chose pop time as the explanatory variable?

He then plots the five data points, the mean point, and the regression line on the graph seen in the diagram below.



More information

The image is a graph with the X-axis labeled as "Height (metres)" ranging from 65 to 80 with intervals of one unit. The Y-axis is labeled "Pop time (s)" ranging from 1.5 to 2.3 with intervals of 0.1 seconds. Both axes have arrowheads indicating direction. On the graph are five purple data points scattered around a line that slopes upward. A red point labeled "Mean" is plotted on this line. The line through the data points is the regression line, representing the best fit through the data.

[Generated by AI]

Why is the regression line shown in the graph considered to be the line of best fit? What makes it the ‘best’? To answer these questions, let us explore the idea of a residual.

Important

A **residual** is the vertical distance between a given data point and the regression line.

Student
view

Overview (/study/app/math-ai-hl/sid-132-cid-761618/ov

In the table below, the actual pop times for each height are shown with the pop time that the regression line predicts for each height (see also the interactive activity below). For example, the estimated pop time for a height of 69 inches can be found as follows:

$$\begin{aligned} y &= 0.04856x - 1.52158 \\ &= 0.04856 \times 69 - 1.52158 \\ &= 1.82906 \approx 1.829 \text{ s} \end{aligned}$$

Height (inches)	69	71	73	74	76
Actual pop time (s)	1.81	1.95	1.91	2.25	2.10
Estimated pop time (s)	1.829	1.926	2.023	2.072	2.169

Example 2

★☆☆

Find the sum of all of the residuals of the data shown in the table above.

The residual for one point is the difference between the estimated pop time given by the regression line and the actual pop time from the data.

$$\begin{aligned} &\sum (\text{actual pop time} - \text{estimated pop time}) \\ &= (1.81 - 1.829) + (1.95 - 1.926) + (1.91 - 2.023) + (2.25 - 2.072) + (2.10 - \\ &= -0.019 + 0.024 - 0.113 + 0.178 - 0.069 = 0.001 \approx 0 \end{aligned}$$

Note that since there are data values both above and below the regression line, some of the differences are negative and some are positive. This causes the sum to be very close to 0. In fact, if the fully unrounded regression equation was used, the sum would be exactly zero. Why is it important for the sum to be zero?

The result from **Example 2** is always the case for the sum of the residuals for a regression line. So, if the sum is always zero, how can you decide how well the regression line fits the data?

The answer is that you square the differences to make all of the values positive and then take the sum.

Student view

Overview (/study/ap ai- hl/sid- 132- cid- 761618/ov

Example 3

★★☆

Find the sum of the square residuals for the data.

$$\begin{aligned} SS_{\text{res}} &= \sum (\text{actual pop time} - \text{estimated pop time})^2 \\ &= (1.81 - 1.829)^2 + (1.95 - 1.926)^2 + (1.91 - 2.023)^2 + (2.25 - 2.072)^2 + (2.10 - \\ &= 0.050151 \end{aligned}$$

Note that the value above is an approximate value, using the approximate estimated pop up times. If you use the more accurate values or you use your GDC to find the sum of the square residuals, then you get a slightly different 0.050259589.

Exam tip

Take note of the notation SS_{res} .

This notation will be used in exams for the sum of square residuals.

Making connections

The sum of square residuals can be used as a measure of fit for the regression line or a model. In section 4.13.2 (/study/app/math-ai-hl/sid-132-cid-761618/book/the-coefficient-of-determination-id-27535/) , you will learn how to quantify this fit.

Finally, consider the graph of the data in the interactive activity below. Note that the residuals are now shown along with the value of the sum of square residuals. Click and drag point S. What happens to the sum of square residuals as the new line moves away from the regression line?

Student view

Interactive 1. Sum of Square Residuals.

Credit: GeoGebra (https://www.geogebra.org/m/dksqbkmz) Nicholas Broom

More information for interactive 1

This interactive is a graph that allows the user to understand the graph of the sum of square residuals.

The graph displays a scatter plot showing the relationship between "Height" in meters on the x-axis ranging from 66 to 86 and "Pop time" in seconds on the y-axis ranging from 1.6 to 2.6. There are several data points plotted. A blue best-fit line is shown with a red point labeled "Mean," accompanied by vertical lines connecting the data points to the blue line. The Sum of Squared Residuals represented by SSres with its value is displayed at the top.

The black and purple dots represent individual data points, where each point corresponds to a specific height and its associated pop time.

The Line of Best Fit, also known as the regression line (represented in blue), illustrates the overall trend in the data. This line is positioned in a way that best "fits" the scattered data points based on a specific criterion. The red point represents the mean values of both height and pop time, serving as the centroid of the data. Vertical lines connect each observed data point, represented by the purple dots, to the corresponding black dots on the line of best fit at the same height.

The vertical distance between each observed data point and the line of best fit is called the residual (or error).

On Clicking and dragging the point S on the blue line, the sum of square residuals increases as the new line moves away from the regression line and it decreases as it moves closer to the regression line.
This interactive will help the user to understand that the best-fit line occurs when the sum of square residuals is minimized.

Overview (/study/ap ai- hl/sid- 132- cid- 761618/ov

As you can see from the applet above, the best-fit line occurs when the sum of square residuals is minimised. For this reason, regression equations are often referred to as least-squares regression curves.

International Mindedness

Over centuries and around the globe, many different measuring systems have been developed, often originally based on the human body (for example length in handspans in China and feet and inches — thumbs — in Europe). Starting in the 19th century, countries began to agree on a single system in which all units are related to seven base units, and can be easily scaled up and down in steps of 10^3. Now almost every country has adopted this system, the SI (Système Internationale d'Unités, International System of Units) except the USA and Myanmar.

It is easy to measure objects in terms of your handspan, say. Can you think of reasons why this is not a good system for everybody to use?

What are the advantages of countries collaborating to develop and use SI units? Why do people sometimes still use other units — for example, buying food in pounds or measuring horses in hands?

3 section questions ⌄

4. Probability and statistics / 4.13 Non-linear regression

The coefficient of determination

Section Student... (0/0) Feedback Print (/study/app/math-ai-hl/sid-132-cid-761618/book/the-coefficient-of-determination-id-27535/print/) Assign

Sum of square residuals as a proportion

In the previous section, you used Ben's data on player height versus pop time to create a best-fit line. The line you created was the best possible line for the given five pieces of data. However, even though you found the best-fit line for the data, you still need to consider whether a straight line is the best option for the regression line.

Recall that in the previous section that you learned how to calculate the sum of square residuals, and for Ben's data you found that $SS_{\text{res}} = 0.050151$ (**Example 3** in section 4.13.1 (/study/app/math-ai-hl/sid-132-cid-761618/book/sum-of-square-residuals-id-27534/)). What does that value tell you about how well the regression line fits the data? To answer this question, let us revisit the significance of the

Student view

Overview (/study/ap ai- hl/sid- 132- cid- 761618/ov

mean point. Recall, from subtopic 4.4 (/study/app/math-ai-hl/sid-132-cid-761618/book/the-big-picture-id-26078/), that the x -coordinate of the mean point is the average value of the explanatory variable and the y -coordinate is the average value of the responsive variable.

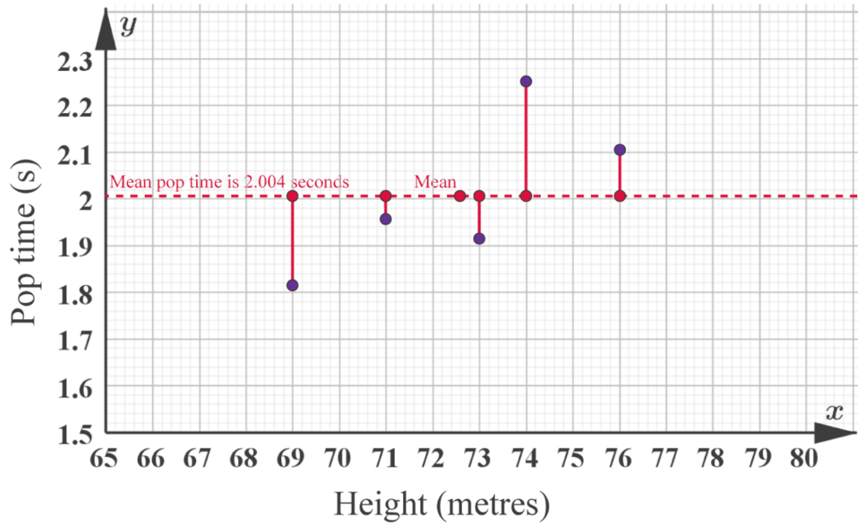
✓ **Important**

When investigating the relationship between two variables, you want to test whether moving away from the mean value of the explanatory variable results in a similar move away from the mean value of the responsive variable.

Let us look at Ben's data once more and consider a slightly different graph of the data.

						Mean
Pop time (seconds)	1.81	1.95	1.91	2.25	2.10	2.004
Height (inches)	69	71	73	74	76	72.6

In the graph in below, the distance from each data point to the mean pop time of 2.004 has been represented with a vertical red line. As you did with the residuals in the previous section, you can find the sum of the squares of each of these distances.



More information

The graph shows a relationship between pop time in seconds on the Y-axis and height in meters on the X-axis. The Y-axis ranges from 1.5 to 2.3 seconds, while the X-axis ranges from 65 to 80 meters. The graph contains several data points plotted on varying heights. A red horizontal dashed line at 2.004 seconds indicates the mean pop time. Each data point is connected to this mean

Student view

Overview (/study/app/math-ai-hl/sid-132-cid-761618/ov

line by a vertical red line, representing the distance from the mean. The text 'Mean pop time is 2.004 seconds' is slightly above the red dashed line, providing a reference point for comparison with the individual data points, and the term 'Mean' is written near the center below the dashed line.

[Generated by AI]

Example 1

★☆☆

Find the total sum of the squares of the distances between each data point and the mean pop time of 2.004.

$$\begin{aligned} SS_{\text{tot}} &= \sum (\text{pop time} - \text{mean pop time})^2 \\ &= (1.81 - 2.004)^2 + (1.95 - 2.004)^2 + (1.91 - 2.004)^2 + (2.25 - 2.004)^2 + (2.10 - \\ &= 0.11912 \end{aligned}$$

Exam tip

Take note of the notation SS_{tot} .

This notation will be used in exams for the total sum of the squares of the differences between each responsive data value and the mean responsive value.

Now that you have found SS_{tot} , you can see how large SS_{res} is in comparison by expressing the two values as a fraction:

$$\frac{SS_{\text{res}}}{SS_{\text{tot}}} = \frac{0.050151}{0.11912} \approx 0.421$$

To explore the significance of this value, consider the graph in the interactive activity below.

Move the predicting line by clicking on the open circle to investigate how the quotient changes. Why is the denominator not changing? When is the numerator and the quotient smallest? Can this ratio be viewed as an indication of how well different lines can be used to predict pop time for a given height based on the data? Which is better, small or large ratio?

Student view

Interactive 1. Investigating Quotient Change.

More information for interactive 1

This interactive allows users to explore how well a regression line fits the given data by adjusting the predicting line manually.

A scatterplot with Height in meters on the x-axis, ranging from 65 to 80, and Pop Time in seconds on the y-axis, ranging from 1.5 to 2.5. A red dashed line is displayed along the x-axis, representing the mean pop time of 2.004 seconds with 5 red dots along it. The data points are marked in blue and joined to the vertical red dots. A blue dashed line, passing through the mean point, represents the fixed denominator of the ratio $\frac{\mathrm{SS}_{\mathrm{res}}}{\mathrm{SS}_{\mathrm{tot}}}$. A solid blue line, also passing through the mean point, represents the numerator of the ratio $\frac{\mathrm{SS}_{\mathrm{res}}}{\mathrm{SS}_{\mathrm{tot}}}$ and serves as the prediction line. Users can adjust this line by dragging the open circle on it.

By clicking and dragging the red circle, you can change the slope or position of the line and observe how the sum of squared residuals $\mathrm{SS}_{\mathrm{res}}$—the total error between predicted and actual values—varies accordingly. The denominator $\mathrm{SS}_{\mathrm{tot}}$, which represents the total variability of the data around the mean pop time, remains fixed because it is independent of the regression line. The ratio $\frac{\mathrm{SS}_{\mathrm{res}}}{\mathrm{SS}_{\mathrm{tot}}}$ indicates the proportion of unexplained variation; a smaller ratio means the line fits the data better, with the minimum value occurring when the line is the least squares regression line.

This interactive tool helps visualize the concept of model fit, demonstrating why minimizing residuals leads to the most accurate predictions and how deviations from the best-fit line increase prediction errors. By experimenting with different lines, you can better understand the trade-offs in regression analysis and the importance of selecting the optimal model.

The coefficient of determination

By finding that $\frac{SS_{\text{res}}}{SS_{\text{tot}}} = 0.421$ you have shown that the sum of the squares of the residuals from the graph above is about 42.1% of the total sum of the squares of the differences of the values and the mean value. The remaining percentage gives us a measure of how well the regression line accounts for the given data.

Overview (/study/app/math-ai-hl/sid-132-cid-761618/ov

Example 2

★☆☆

Find the percentage of the variation in pop times accounted for by Ben's regression line.

This can be found by subtracting the value for $\frac{SS_{\text{res}}}{SS_{\text{tot}}}$ from 1:

$$1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}} = 1 - 0.421 = 0.579$$

Therefore, the regression line accounts for 57.9% of the variation of pop times of the given heights.

The expression $1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$ is known as the coefficient of determination and is given the notation R^2 .

✓ **Important**

The coefficient of determination, R^2, gives the proportion of variability in the responsive variable accounted for by the chosen model (in this case the regression line).

Exam tip

The equation $R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$ is helpful for understanding the meaning of the coefficient of determination, but you will not be expected to use it in exams.

Let us take this one step further by considering an ideal situation with different data.

Example 3

★☆☆

Find the value of SS_{res} for the following data, using x as the explanatory variable.

Student view

Overview
(/study/ap
ai-
hl/sid-
132-
cid-
761618/ov

x	13	37	24	17	42
y	31.7	83.3	55.35	40.3	94.05

Use technology to find the regression line.

$$y = 2.15x + 3.75$$

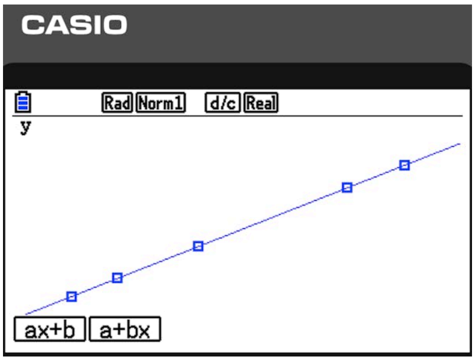
Use the regression line to find the estimated values for y

x	13	37	24	17	42
$y\,(\text{actual})$	31.7	83.3	55.35	40.3	94.05
$y\,(\text{estimated})$	31.7	83.3	55.35	40.3	94.05

Therefore, $SS_{\text{res}} = 0$ because all of the estimated values match the actual values.

In **Example 3**, you found that there were no residuals. What does that tell you about the line of regression and the five data points? Plot the five data points and the regression line on the same graph within your calculator.

The result screens from the calculators



Casio fx-CG50

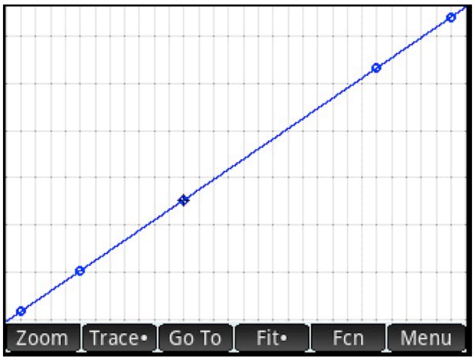
More information

The image displays a graph on what appears to be a calculator screen. The screen header shows options like "Rad," "Norm1," "d/c," and "Real." The graph indicates a linear trend with data points marked along the line. There are also two labeled buttons at the bottom of the screen: "ax+b" and "a+bx." The graph has an X-axis and a Y-axis, but only the Y-axis is labeled with a 'y'. This setup suggests a focus on linear equations and graphing functions on the calculator.

Student
view

Overview
(/study/ap
ai-
hl/sid-
132-
cid-
761618/ov

[Generated by AI]

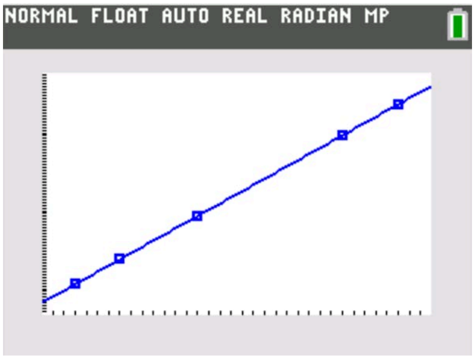


HP Prime

More information

The image shows a graph featuring a blue line with markers plotted across a grid. At the bottom, there are buttons labeled Zoom, Trace, Go To, Fit, Fcn, and Menu. The graph appears to represent a linear relationship with consistent spacing between the markers. The X-axis and Y-axis are marked with evenly spaced intervals. Although specific values are not visible, the graph suggests a direct correlation between the variables plotted on the axes.

[Generated by AI]



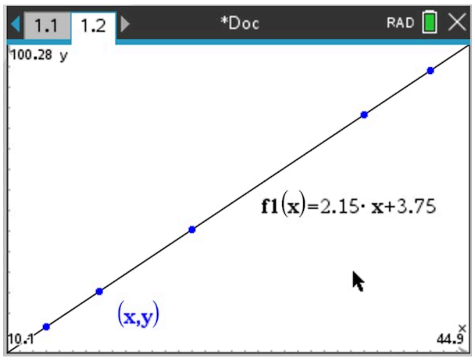
TI-84 plus CE

More information

The image shows a graph displayed on a calculator screen. The screen has a text header with the words "NORMAL FLOAT AUTO REAL RADIAN MP" and a battery indicator. The graph features a simple linear plot with a blue line that climbs from the bottom left to the top right, overlaid with square markers at several data points. The X-axis and Y-axis are marked by small dashes, but the specific labels or units are not visible. The graph suggests a consistent upward trend, characteristic of a linear equation or function.

[Generated by AI]

Student
view



TI-nspire CX

More information

The image shows a graph from a TI-nspire CX display with a linear function plotted. The graph includes a line representing the function f1(x) = 2.15x + 3.75. The x-axis ranges approximately from 10.1 to 44.9, and the y-axis ranges from 10.1 to 100.28. Several blue data points are placed along the line, indicating specific (x, y) values for the plotted function. The equation of the line is displayed prominently near the line in the graph's central area.

[Generated by AI]

In your graph you see that the regression line goes perfectly through each of the points. This makes sense as there were no residuals. Therefore, the regression line accounts for 100% of the variation of the y -values in the data.

Example 4

★☆☆

Let us consider Ben’s data on pop time one final time.

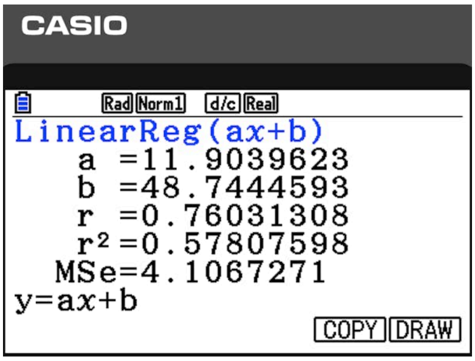
Pop time (seconds)	1.81	1.95	1.91	2.25	2.10
Height (inches)	69	71	73	74	76

Using your graphic display calculator, find the correlation coefficient, R, of Ben’s data, and thus the value of R^2.

From your calculator, you find that $R = 0.7603$ and $R^2 = 0.578$.

Student view

The result screens from the calculators



LinearReg(ax+b)
a =11.9039623
b =48.7444593
r =0.76031308
r^2 =0.57807598
MSe=4.1067271
y=ax+b

Casio fx-CG50

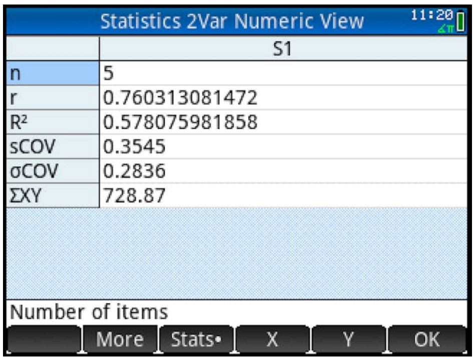
The image shows a calculator screen displaying results of a linear regression analysis. The top of the screen has the text 'LinearReg(ax+b)' highlighted in blue. Below, several calculated values are presented:

- a = 11.9039623
- b = 48.7444593
- r = 0.76031308
- r² = 0.57807598
- MSe = 4.1067271

At the bottom, the formula 'y=ax+b' is displayed.

The screen has function buttons at the top for 'Rad', 'Norm1', 'd/c', and 'Real'. At the bottom right corner, there are buttons labeled 'COPY' and 'DRAW'. The brand 'CASIO' is visible at the top of the screen.

[Generated by AI]



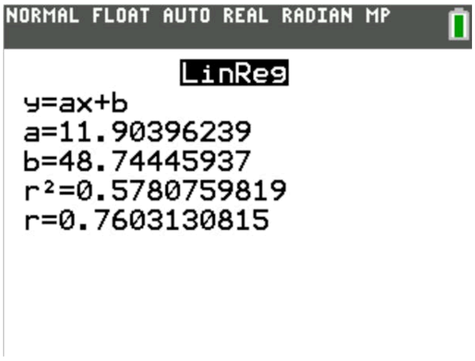
	S1
n	5
r	0.760313081472
R²	0.578075981858
sCOV	0.3545
σCOV	0.2836
ΣXY	728.87

Number of items

HP Prime

The image shows a screen from a statistical software titled "Statistics 2Var Numeric View." The screen is divided into a header and several rows displaying variable names and their associated values under the column labeled "S1." The labels on the left include: n, r, R², sCOV, σCOV, and ΣXY. Their corresponding values are 5, 0.760313081472, 0.578075981858, 0.3545, 0.2836, and 728.87 respectively. Below these entries, there is text reading "Number of items." At the bottom of the screen, there are buttons labeled "More," "Stats," "X," "Y," and "OK." The display suggests a numeric output from statistical analysis.

[Generated by AI]

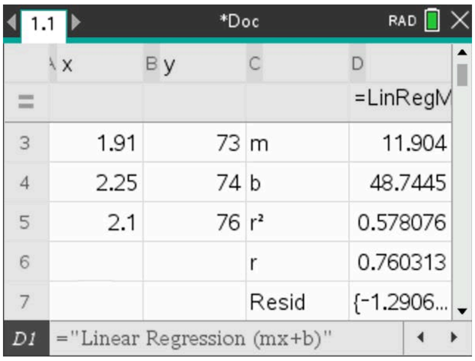


```
LinReg
y=ax+b
a=11.90396239
b=48.74445937
r²=0.5780759819
r=0.7603130815
```

TI-84 plus CE

The image shows a calculator screen with the linear regression formula y=ax+b at the top. It lists values for the coefficients and correlation: a=11.90396239, b=48.74445937, r²=0.5780759819, and r=0.7603130815. The calculator is in normal mode with floating point numbers displayed, in auto real and radian measurement mode, indicated by the top bar. The battery icon shows a full charge.

[Generated by AI]



	A x	B y	C	D
=				=LinRegM
3	1.91	73	m	11.904
4	2.25	74	b	48.7445
5	2.1	76	r²	0.578076
6			r	0.760313
7			Resid	{-1.2906...

D1 ="Linear Regression (mx+b)"

TI-nspire CX

The image displays a spreadsheet interface with several columns labeled: Ax, By, and an unlabeled column beside formulas. Rows 3 to 5 contain data entries. Row 3 shows Ax = 1.91, By = 73, with corresponding regression parameters: m = 11.904. Row 4 lists Ax = 2.25, By = 74, with b = 48.7445 indicated. Row 5 shows Ax = 2.1, By = 76, with parameters: r² = 0.578076 and r = 0.760313. A note associated with D1 states: "Linear Regression (mx+b)."

Overview
(/study/ap
ai-
hl/sid-
132-
cid-
761618/ov

[Generated by AI]

Consider the result for **Example 4**. How does it compare to the result of **Example 2** ?

Making connections

For linear models, the coefficient of determination is equal to the square of Pearson's product moment correlation coefficient, which you encountered in section 4.4.2 (/study/app/math-ai-hl/sid-132-cid-761618/book/pearsons-r-correlation-coefficient-id-26080/) .

In conclusion, consider again the coefficient of determination R^2 you found for Ben's data. The value tells us that, for the given data, 57.9% of the variation in the pop times can be accounted for by the regression line based on the heights of the catchers. Is it reasonable that height is not the only factor involved in quickly throwing the ball to second base? What other factors could be involved? If you collected more than five data points do you think the coefficient of determination would increase or decrease?

3 section questions ⌄

4. Probability and statistics / 4.13 Non-linear regression

Non-linear regression

Section Student... (0/0) Feedback Print (/study/app/math-ai-hl/sid-132-cid-761618/book/nonlinear-regression-id-27536/print/) Assign

Making connections

In topic 2 (/study/app/math-ai-hl/sid-132-cid-761618/book/the-big-picture-id-26012/) you learned about linear, quadratic, cubic, exponential, power and sine functions and how they can be used as models. Take a moment now to review the shape of the graph of each of these functions.

In section 4.13.2 (/study/app/math-ai-hl/sid-132-cid-761618/book/the-coefficient-of-determination-id-27535/) you learned that the coefficient of determination can tell you what proportion of the variability in the responsive variable is accounted for by a linear regression model. Let us now explore how it can also be useful for non-linear models.

Student
view

Overview (/study/ap ai-hl/sid-132-cid-761618/ov

Example 1

★☆☆

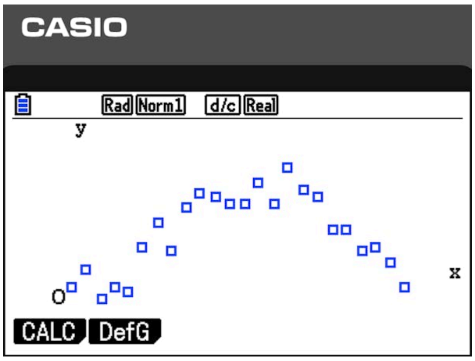
Yejun is doing a study on the temperature trends in Seoul, South Korea. He has collected the following temperature data from 2018.

Date	Jan 1	Jan 15	Feb 1	Feb 15	Mar 1	Mar 15	Apr 1	Apr 15	May 1	May 15	Jun 1	Jun 15
Day	1	15	32	46	60	74	91	105	121	135	152	16
Temp. (°C)	-1	4	-4	-1	-2	10	17	9	21	25	24	22

Date	Jul 1	Jul 15	Aug 1	Aug 15	Sept 1	Sept15	Oct 1	Oct 15	Nov 1	Nov 15	Dec 1
Day	182	196	213	227	244	258	274	288	305	319	335
Temp. (°C)	22	28	22	32	26	24	15	15	9	10	6

Plot Yejun's data in your graphic display calculator.

Using your calculator, a graph similar to one of those shown below should be seen. Remember that you can use the 'ZoomStat' function of the calculator to change the window to include all of your data.



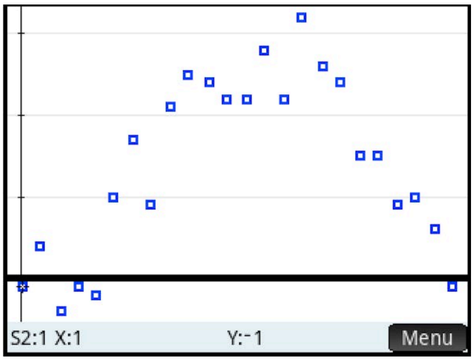
Casio fx-CG50

Student view

Overview (/study/app/math-ai-hl/sid-132-cid-761618/ov

The image shows a graph displayed on a Casio calculator screen. The graph consists of multiple blue square markers that form an upward curve, peaking and then curving downward. The X-axis is labeled with 'x' on the right side, and the Y-axis is labeled with 'y' at the top left. Above the graph, there are various options like 'Rad', 'Norm1', 'd/c', and 'Real'. At the bottom left, there are options labeled 'CALC' and 'DefG'. The graph seems to represent a data distribution or a statistical plot with scattered data points following a general trend over the X and Y axes.

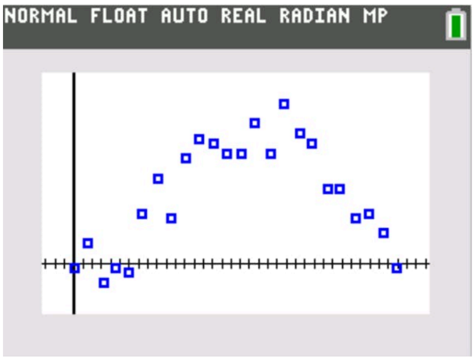
[Generated by AI]



HP Prime

The image shows a scatter plot graph, featuring a series of blue square data points distributed across a grid. The X-axis is labeled in increments, with visible ticks marking units. The Y-axis has a similar increment and tick structure. Data points are scattered in a pattern, showing a slight upward trend towards the middle and tapering off towards the end. There's a menu option visible at the bottom with the text 'S2:1 X:1 Y:-1' and 'Menu', providing additional features for interaction. The plot appears to be part of a graphical calculator display.

[Generated by AI]



TI-84 plus CE

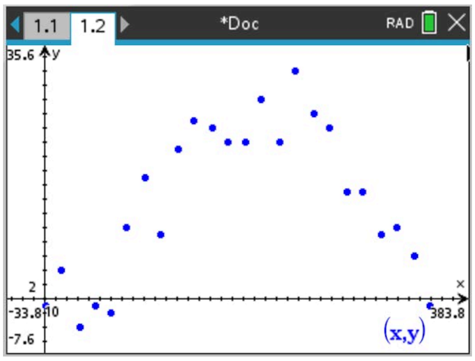
Student view

The image shows a graph displayed on a TI-84 calculator screen. The top of the screen shows text reading "NORMAL FLOAT AUTO REAL RADIAN MP." The graph includes a set of blue square data points plotted on a grid. The X-axis is a horizontal line marked with several tick marks, dividing the graph into increments. The Y-axis is a vertical line on the left

Overview
(/study/ap
ai-
hl/sid-
132-
cid-
761618/ov

side, intersecting the X-axis at the bottom edge of the image. The data points form a curve that rises and falls, indicating a trend. The right side of the top bar shows a battery indicator icon.

[Generated by AI]



TI-nspire CX

The image is a scatter plot displayed on a TI-nspire CX screen. The scatter plot features blue data points distributed across the graph. The X-axis is labeled with values ranging from -33.8 to 383.8, and the Y-axis displays values from -7.6 to 35.6. The data points form a roughly parabolic pattern, starting low on the left, rising and peaking towards the middle, and then descending towards the right. The axes intersect at the point where Y is approximately -7.6 and X is 2. The graph appears to be set in a mathematical software interface, with menu options visible at the top, including navigation arrows and a file name labeled as '*Doc'.

[Generated by AI]

Example 2

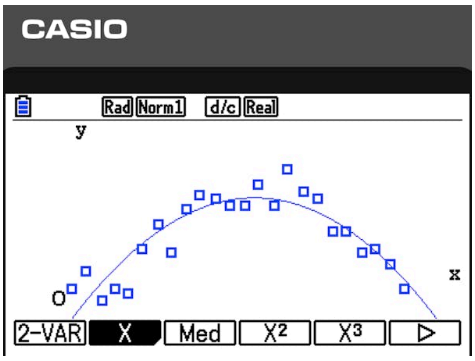
★☆☆

Look at **Example 1** again. Yejun notices that the graph seems to show a parabolic curve and decides to create a quadratic regression model for the data.

Using your graphic display calculator, create a quadratic regression curve for the data in the table above, graph the curve on the same axes as the plotted data, and find the coefficient of determination for the model.

Student view

Your calculator should produce a graph similar to one of those shown below.

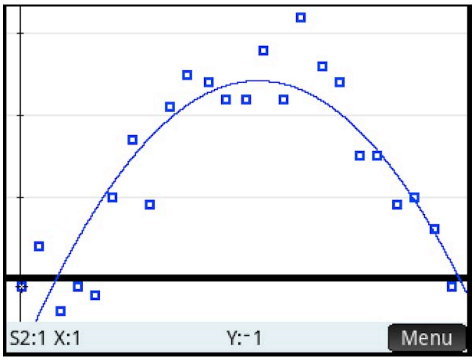


Casio fx-CG50

The image is a screenshot from a Casio calculator displaying a graph. The screen shows a scatter plot with several blue square data points arranged in an arching pattern, likely representing a parabolic curve. There is a curve fitted through these points. The X-axis is labeled 'x' at the bottom right, and the Y-axis is labeled 'y,' which is displayed vertically on the left.

Additional interface icons at the top display settings like 'Rad Norm1 d/c Real.' There are also buttons at the bottom for different settings labeled '2-VAR,' 'X,' 'Med,' 'X2,' 'X3,' followed by a right arrow symbol. The graphical representation suggests a statistical or mathematical analysis displaying a set of two-variable data.

[Generated by AI]

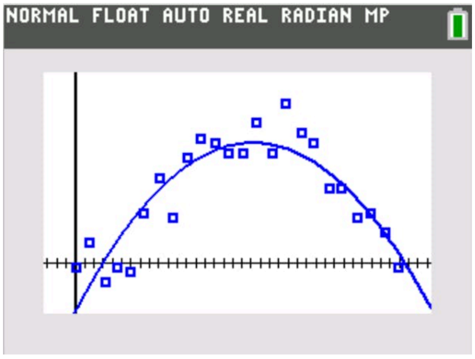


HP Prime

The image is a graph displaying a curve with blue data points plotted along it. The x-axis is labeled as 'S2:1 X:1', and the y-axis as 'Y:-1', suggesting some form of coordinate system or specific values related to the graph. The graph appears to have a parabolic curve, with data points scattered both on the curve and around it, indicating a potential fit or distribution. There is a black horizontal line that could be a part of the grid or an axis. There is also a button labeled 'Menu' in the lower right corner, which may indicate this graph is part of a software or calculator interface.

[Generated by AI]

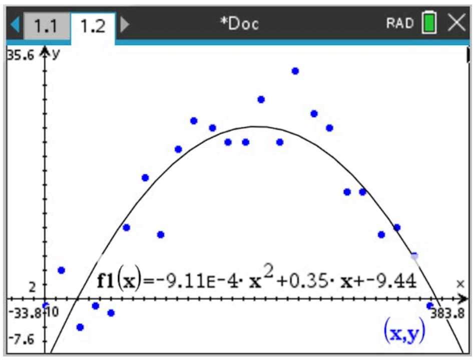
Student
view



TI-84 plus CE

The image shows a calculator screen with the top bar displaying the settings: "NORMAL FLOAT AUTO REAL RADIAN MP" with a battery icon on the right. Below the settings is a graph containing a series of blue square data points overlaid on a blue curved line forming a parabolic shape. The graph includes two axes: a vertical (Y-axis) and a horizontal (X-axis). The Y-axis and X-axis intersect with the curve indicating various plotted data values, representing a trend that rises to a peak and falls symmetrically.

[Generated by AI]



$f1(x)=-9.11E-4\cdot x^2+0.35\cdot x+-9.44$

TI-nspire CX

The image shows a scatter plot with a quadratic curve fitting the data points. The X-axis, representing an unspecified variable, ranges from 2 to 383.8. The Y-axis, representing another variable, ranges from -33.8 to 35.6. Data points are plotted in blue dots. The quadratic model is f(x) = -9.11E-4 * x^2 + 0.35 * x + 9.44, which is displayed over the curve. This curve follows a parabolic trajectory, peaking near the center of the graph. Key features include the consistent spacing of data points around the curve, indicating a moderate fit, as suggested by the R^2 value of 0.797570.

[Generated by AI]

For this quadratic model, $R^2 = 0.795915\ldots$

Overview (/study/app/math-ai-hl/sid-132-cid-761618/ov

As you learned in the previous section, the coefficient of determination is telling you that in this case 79.6% of the variation in the temperature data can be accounted for by this quadratic model. How confident should you be in this model?

⚠ Be aware

There are many factors that affect the validity of a model. The coefficient of determination should not be the only tool used to determine which model should be used for a particular data set.

As multiple years of temperature data would oscillate in a similar way to a sinusoidal function, Yejun reconsiders his decision and decides to instead create a sine regression for his data.

ⓘ Exam tip

For examination questions, you will be expected to know how to create linear, quadratic, cubic, exponential, power and sine regressions.

Using the GeoGebra app below, explore the effectiveness of different regressions for certain data sets. What factors should you consider in order to determine which regression model best fits the data?

Interactive 1. Exploring the Effectiveness of Different Regressions.

More information for interactive 1

Student view

This interactive tool enables users to explore regression analysis concepts by fitting various models to datasets and evaluating their effectiveness.

The screen is divided into two halves. On the left, a scatterplot with the x-axis ranging from 0 to 12 and the y-axis ranging from 0 to 14. Random points are generated on the graph, and the user can generate new random points by clicking on the "New Data" button on the top. On the right side of the screen, users can test linear, quadratic, cubic, exponential, power, and sine regressions by clicking on their respective checkboxes, with the tool displaying both the R ^{2} value (indicating goodness-of-fit) and the corresponding function equation for each model. A yellow line projects on the graph representing the checkbox the user has checked.

The visualization helps compare how different curves align with the data points, with higher R ^{2} values (closer to 1) suggesting better fits. When selecting the optimal model, users should consider multiple factors: the mathematical fit quality through R ^{2} values, the data's inherent pattern (linear trends, polynomial curves, exponential growth/decay, or periodic behavior), model complexity to avoid overfitting, residual analysis for validation, and real-world context to ensure practical relevance. Through hands-on experimentation, learners can discover how different regression types capture data relationships and develop skills in choosing the most appropriate model for various analytical scenarios, making this an effective tool for both introductory and advanced statistical exploration.

3 section questions ⌄

4. Probability and statistics / 4.13 Non-linear regression

Checklist

Section Student... (0/0) Feedback Print (/study/app/math-ai-hl/sid-132-cid-761618/book/checklist-id-27537/print/) Assign

What you should know

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

- create linear, quadratic, cubic, exponential, power, and sine regression models using technology
- calculate the sum of square residuals for a linear model
- use the sum of square residuals as a measure of fit for a model
- evaluate the coefficient of determination (R^2) using technology
- interpret the meaning of the coefficient of determination as the proportion of variability in the responsive variable accounted for by the chosen model

Overview
(/study/ap
ai-
hl/sid-
132-
cid-
761618/ov

- analyse various factors, including the coefficient of determination, to determine whether a certain model is the best fit for a given data set.

4. Probability and statistics / 4.13 Non-linear regression

Investigation

Section Student... (0/0) Feedback Print (/study/app/math-ai-hl/sid-132-cid-761618/book/investigation-id-27538/print/) Assign

Baseball players

Credit: Getty Images

In this subtopic, you did a lot of work with data on a baseball catcher’s pop time and their height.

However, all of the calculations were carried out with only five data points, which is not enough data to gain a full understanding of the possible connection between the two variables. What would the coefficient of determination be if you collected more data? Would the interpretations of the data be any different?

1. Visit this website (https://baseballsavant.mlb.com/poptime) to c ollect pop time data for the catchers with the best 50 pop times.
2. Visit this website (https://www.baseball-reference.com) to find the heights of the 50 players you chose in Step 2.
3. Use your data to re-calculate the coefficient of determination.

Student view

Compare your findings with the result you calculated earlier within the subtopic.

Overview (/study/app/math-ai-hl/sid-132-cid-761618/ov

Rate subtopic 4.13 Non-linear regression

Help us improve the content and user experience.

☆ ☆ ☆ ☆ ☆

Student view